ABSTRACT

Title of dissertation: RECOGNIZING HUMAN FACES: PHYSICAL MODELING AND PATTERN CLASSIFICATION

Gaurav Aggarwal, Doctor of Philosophy, 2008

Dissertation directed by: Professor Rama Chellappa
Dept. of Electrical & Computer Engineering
and Dept. of Computer Science

Although significant work has been done in the field of face recognition, the performance of the state-of-the-art face recognition algorithms is not good enough to be effective in operational systems. Most algorithms work well for controlled images but are quite susceptible to changes in illumination, pose, etc. In this dissertation, we propose methods which address these issues, to recognize faces in more realistic scenarios. The developed approaches show the importance of physical modeling, contextual constraints and pattern classification for this task.

For still image-based face recognition, we develop an algorithm to recognize faces illuminated by arbitrarily placed, multiple light sources, given just a single image. Though the problem is ill-posed in its generality, linear approximations to the subspace of Lambertian images in combination with rank constraints on unknown facial shape and albedo are used to make it tractable. In addition, we develop a purely geometric illumination-invariant matching algorithm that makes use of the bilateral symmetry of human faces. In particular, we prove that the

set of images of bilaterally symmetric objects can be partitioned into equivalence classes such that it is always possible to distinguish between two objects belonging to different equivalence classes using just one image per object.

For recognizing faces in videos, the challenge lies in suitable characterization of faces using the information available in the video. We propose a method that models a face as a linear dynamical system whose appearance changes with pose. Though the proposed method performs very well on the available datasets, it does not explicitly take the 3D structure or illumination conditions into account. To address these issues, we propose an algorithm to perform 3D facial pose tracking in videos. The approach combines the structural advantages of geometric modeling with the statistical advantages of a particle filter based inference to recover the 3D configuration of facial features in each frame of the video. The recovered 3D configuration parameters are further used to recognize faces in videos.

From a pattern classification point of view, automatic face recognition presents a very unique challenge due to the presence of just one (or a few) sample(s) per identity. To address this, we develop a cohort-based framework that makes use of the large number of non-match samples present in the database to improve verification and identification performance.

RECOGNIZING HUMAN FACES: PHYSICAL MODELING AND PATTERN CLASSIFICATION

by

Gaurav Aggarwal

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Professor Larry S. Davis
Professor David W. Jacobs
Professor David Mount
Professor K. J. Ray Liu

Dedication

To my parents.

Acknowledgments

I would like to thank a number of people without whose support and encouragement, this dissertation would not have been possible. First of all, I express my sincere gratitude to my advisor Prof. Rama Chellappa whose invaluable guidance and encouragement helped me throughout the course of my PhD. Not only did he work hard so that I never had to worry about financial support but also gave me complete freedom to work on problems of my choice. I am also thankful to Prof. David Jacobs, Prof. Larry Davis, Prof. David Mount and Prof. K. J. R. Liu for serving in my dissertation committee. I would also like to thank Dr. Nalini Ratha and Dr. Ruud Bolle whole guided me during my long internships at IBM Research. Like my PhD advisor, they gave me full liberty to pursue my interests during my internships. I am also grateful to Object Video, GE Research and IBM Research for finding me worthy enough to offer me full time research positions.

I would not have been able to overcome the hiccups of graduate studies, had there not been care and support of my friends and colleagues. The mere presence of friends like Ashok, Aswin, Indrajit, Kaushik, Mahesh, Narayanan, Naresh, Soma, Sameer, Saurabh and Shiv made my stay extremely cherishable. A special thanks to Soma, Amit (ARC) and Ashok who were courageous enough to let me work with them for several projects.

Finally, I am extremely indebted to my parents, bhaiya-bhabhi and didi-jijjo for their unstinting support and affection in all my endeavors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Humans make use of face as an important cue for identifying people. In fact, a photograph showing subject's face is an integral part of most state-issued identifying documents. Though humans have unmatched abilities to recognize familiar faces under arbitrary external (illumination, pose, etc.) or internal (expression, deformation, etc.) transformations, we are not so efficient when given the task of memorizing and matching a large number of unfamiliar faces. This makes automatic face recognition very crucial from the point of view of wide range of commercial and law enforcement applications. Moreover, automatic face recognition is probably one of the most well-defined problems in the field of computer vision and image analysis. These reasons justify the kind of attention, it has received from academic researchers and corporate vendors in the past decade. In spite of the large amount of work that has been done in this area, there are quite a few important issues which prevent the current algorithms to be effective in real conditions. Though current algorithms are able to recognize faces from images/videos taken under controlled conditions, they struggle to generalize across variations in illumination condition, pose, expression, aging, etc. Such variations, though difficult to model algorithmically, occur commonly in real life. In this dissertation, we address some of these issues to bring face

recognition closer to being useful for in real world applications. In this endeavor, our efforts have been directed towards the following issues

1. Illumination-invariant recognition of faces,
2. 3D facial tracking in uncalibrated videos,
3. Recognizing faces in low resolution videos,
4. Cohort analysis to improve matching (both verification and identification) performance, and
5. Privacy issues concerning face recognition.

1.2 Biometric perspective

'Biometrics' refers to the measurement and analysis of physical or behavioral traits of humans. More often than not, such an analysis is directed towards the goal of verifying or determining personal identity. Though identity can be established using means like PINS or passwords, such cues can be forgotten, stolen and passed on to others fairly easily. Thus having the secret code/PIN cannot safely be used to validate the identity of the person. A biometric characteristic should have the following characteristics for it to be truly useful in authentication related applications

- Universality (every person should have the biometric),
- Uniqueness (every person's biometric signature should be different from others),

- Permanence (the biometric should be invariant over time),
- Collectibility (there should exist an easy, quick, inexpensive, non-intrusive way to acquire the biometric),
- Acceptability (it should be acceptable to people),
- Difficult to circumvent (it should be spoof-proof), and
- low underlying system errors (it should result in low False Accept Rate (FAR), False Reject Rate (FRR), etc).

No matter how good a matching algorithm is, it may not be possible for a single biometric to have all the mentioned desirable properties. This has led to the rise of research in multi-biometric systems that rely on fusing information from multiple biometric evidences.

The advancement and popularity of biometric systems has brought concerns of *biometric-theft*. Unlike PINs or passwords, which can be changed at will when compromised, biometric traits are unique and permanent. This leads to the observation that though biometrics are authentic, they are not secure (or private like passwords). If compromised, biometric signatures cannot be revoked or canceled. It allows for rogue establishments to track subjects across databases and institutions without consent.

1.2.1 Why face as a biometric?

A strong requirement of coming up with secure and user-friendly ways to identify people to safeguard their rights and interests has probably been the guiding force behind biometrics research. The various physical and behavioral human characteristics that have been explored to achieve this goal include fingerprints, faces, voice, gait, irises, retinas and hand geometry. These human traits can be further characterized based on their universality, uniqueness, permanence, collectability, performance, acceptability and circumvention. Though biometrics like fingerprints, irises and retinas invariably outperform the rest in terms of permanence, performance and circumvention, they are not only intrusive but also expect cooperation on the part of the user. Ease of collectability and acceptability are probably the reasons that face has emerged as a popular biometric. Another factor that has contributed to popularity of face recognition research is the fact that face recognition algorithms find use in non-critical applications like automatic tagging and indexing of personal albums and videos.

1.3 Challenges in automatic face recognition

The excitement/concerns masses have about the deployment of automatic face recognition systems in public arena probably justifies the kind of attention these problems have received in the field of computer vision. Face images show a great deal of variability. Interestingly, no two captured face images are exactly identical to each other. The variations in two face images of the same person may arise due

to facial appearance changes or differences in imaging environments. Human faces appear different at different instants due to the following factors

- **Facial deformation:** Human faces being non-rigid undergo deformations due to changes in stress, mood, facial expressions, etc. Though the deformations are not arbitrary and are guided by the underlying muscle and tissue structure, it is very difficult to analyze or model these variations from normal images.

- **Aging:** Human faces undergo considerable variation in appearance due to aging. Faces of different individuals age differently depending on health, stress, habits, race, climate, etc., that makes the task of recognizing faces across age progression very difficult.

- **Cosmetic changes:** Other than the mentioned natural variations, facial appearance can deliberately be changed by makeup, surgery, growing or shaving facial hair, etc. Sometimes even humans find it difficult to generalize across these variations.

In addition to changes in physical appearance of faces, images may appear different due to changes in conditions under which the images are captured. First of all, images may appear different due to difference in capturing device. Other than that, photometric and geometric characteristics of the environment affect image appearance. Photometric characteristics describe the illumination conditions like number, size, strength (intensity), color, placement, etc. of light sources. Geometric characteristics pertain to the geometry of the capturing device with respect to the

face being captured including distance and orientation. Fig. 1.1 illustrates how face appearance changes with variations in pose and illumination.

A lot of work has been done on the problems of constrained and unconstrained face based human identification. By constrained face recognition we mean recognizing faces from images captured in controlled canonical pose and lighting conditions with neutral facial expression. The performance of the current state-of-the-art algorithms is very good [98] as far as recognition in controlled conditions is concerned. On the other hand, a lot still needs to be done to achieve similar performance in more realistic scenarios with not much control over the environmental conditions or/and hardly any co-operation from the user.

Figure 1.1: Effect of pose and illumination variations on face images from the PIE dataset [79].

Given a face image, a face recognition algorithm aims at determining the identity of the person. It either generates a set of features or transforms it to a desirable form and then compares it with the images in the database of enrolled subjects. If the goal was to return the most similar looking images from the database, a simple correlation-based measured would have sufficed. Such a measure will probably assign higher similarity score to two face images of different persons in the same pose and illumination as compared to two images of the same person in different pose and/or illumination conditions. In identification tasks, one needs to determine facial similarity independent of these external *nuisance* factors that makes the problem hard. In most practical scenarios, there is just one image (or a few images) to generalize across these nuisance factors making the problem of automatic face recognition even more difficult.

1.4 Contributions of this dissertation

In this section, we highlight the main contributions of this dissertation. In this dissertation, we propose algorithms to 1) recognize faces across illumination variations, 2) improve verification and identification performance using cohort analysis, 3) perform cancelable face matching, and 4) track and recognize faces in low quality videos. Note that unless otherwise stated, we address the most difficult scenario of recognizing faces using just a single image (or video) throughout this dissertation.

1.4.1 Illumination-insensitive matching of faces

Given two images taken under different illumination conditions, there always exists a family of physically realizable objects which is consistent with both the images. In fact, Jacobs *et al.* [42] show that the ambiguity exist even under the hard constraints of Lambertian reflectance and known single point light sources placed at infinity. The lack of information about the geometry and reflectance of the scene makes the problem of illumination-invariant matching in its generality, ill-posed. The result, though a setback to the goal of achieving illumination-invariant matching, has not been too devastating for the task of illumination-insensitive face recognition. For example, given two face images, the physically realizable object that can account for the two, need not be face-like at all. In other words, the class-specific constraints present in the task of face recognition, makes the problem of face matching across lighting variations, somewhat tractable.

Following are a few constraints that can be used to address the intractability of the problem of illumination-invariant matching of faces

- Geometric constraints (e.g., bilateral symmetry of faces)
- Modeling constraints (e.g., linear Lambertian object, morphable model, etc.)
- Photometric constraints (e.g., low-dimensional linear subspace constraint for Lambertian reflectance)
- More samples (e.g., photometric stereo)

In this dissertation, we use geometric, statistical and photometric constraints

as described in the following subsections.

1.4.1.1 Face recognition in the presence of multiple light sources

The susceptibility of traditional face recognition algorithms to changes in pose and illumination has led to the rise of analysis-by-synthesis approaches. Though these approaches are reasonably successful in achieving this goal, most of them assume that the given face is illuminated by a single distant light source which is usually not true in realistic conditions. In contrast, we propose an algorithm to perform recognition using faces illuminated by multiple illumination sources. The following two assumptions are the backbone of our algorithm in addressing this otherwise intractable problem:

- The human face belongs to the class of linear Lambertian objects which means that it is linearly spanned by basis objects; and its surface obeys the Lambertian reflectance model.

- The subspace of Lambertian images is well represented using linear approximation based on fixed distant light sources.

The linear Lambertian property imposes a rank constraint on the shape and albedo of each face which allows one to model it as a linear combination of basis faces. The linear approximation to Lambertian subspaces aids in handling faces illuminated by multiple light sources without any prior knowledge about their number or placement. The algorithm models a face as a Lambertian surface. Therefore, we address the issue of significance of the often ignored hard nonlinearity in Lambert's

law. The nonlinearity accounts for the formation of attached shadows but is easily ignored (under single light source assumption) by eliminating the near-zero pixels from analysis. We show that the nonlinearity can be crucial even for a relatively simple task of estimating dominant illuminant direction. Moreover, multiple illumination scenario degenerates to that of single light source one for Lambertian surfaces if linearized Lambert's law is used.

1.4.1.2 Symmetry and illumination-invariance

Given any two images taken under different illumination conditions, there always exist a physically realizable object which is consistent with both the images even if the lighting in each scene is constrained to be a known point light source at infinity [42]. In this work, we show that images are much less ambiguous if objects are constrained to be bilaterally symmetric with Lambertian reflectance. In fact, set of bilaterally symmetric objects can be partitioned into equivalence classes such that it is always possible to distinguish between any two images of any two objects belonging to different equivalence classes.

Though we focus mainly on faces, the algorithm is applicable to any object/scene as long as the symmetry assumption is satisfied. The images are not required to be frontal though correspondence is assumed to be known. Similar to [97] [28], the unknown arbitrarily varying albedo is initially eliminated from the formulation using 3D bilateral symmetry. This leads to a linear relation involving light source direction and surface gradients. Though this is not sufficient to recover

the surface gradients, one can use this formulation for matching images across illumination variation. Given two linear relations from two different images, we solve for the surface gradients. The veracity of the gradients can be checked by substituting them back in the original irradiance equations and computing albedos for the two images separately. We show that the two albedos are identical if the corresponding pixels represent the same physical reality (same shape and albedo). If the points differ physically, the computed albedos almost always differ. The rare condition under which they are same is derived. In fact, this condition partitions the set of symmetric Lambertian objects into equivalence classes such that it is not possible to distinguish between different objects belonging to the same equivalence class based on just one image per object.

In case of multiple light sources, we do not know the correspondence between pixels and the light sources which illuminate those pixels (as a light source may not lie in front of a surface point to have any effect on the intensity of the corresponding pixel). We introduce class-specific information in the form of average shape to estimate this correspondence. Though this light source assignment can potentially be wrong for a few points, it will not affect the similarity analysis as long as most of the surface points are assigned correctly. This is usually the case for objects like faces, where an average shape represents the class well.

1.4.2 Cohort analysis for biometric matching

The performance of modeling approaches (like the ones proposed in this dissertation to perform illumination-insensitive face matching) depends heavily on how accurately data follows these models. For example, in Chapter 2, it is shown that the single light source assumption leads to very poor recognition performance on images lit by multiple light sources. The proposed multiple light source algorithm performs almost flawlessly on those images. Though one can improve performance by making sure that model fits the data closely, it is often not easy. More often than not, there are inexplicable factors not modeled by the approach, that can negatively affect the performance of a recognition algorithm. It is often difficult to take account of these factors specially with just one sample (here, image) per class (here, identity). In this dissertation, we propose a cohort analysis-based approach that makes use of a large number of non-match samples present in the database to improve verification and identification performance. The following paragraph provides a brief description of the proposed approach.

Most biometric matching approaches make verification or identification decisions based purely on the similarity of the query with the enrolled biometric samples of the claimed identity The similarity is usually determined based on the distance of the query from the enrolled biometrics as determined by matching algorithm. To perform well, such approaches expect the biometric classes to be reasonably compact (around the available sample for each enrolled identity) with respect to the inter-class distances, and similarly distributed. When the class distributions vary

across identities, the verification threshold may turn out to be too stringent for a few classes while too lenient for others. Additionally, biometric classes may not be isotropically distributed around the available sample(s) in feature space, making it difficult to even set a good threshold separately for each class. The performance of biometric systems gets particularly affected in situations when there are significant nuisance factors that are not modeled by the matching algorithm. These can occur in the form of illumination or pose variations in face, scanner quality in fingerprint, phone/microphone quality for speaker verification, etc. If the matching algorithm is unable to factor out these factors effectively, the raw similarity scores obtained are dependent on these factors. This increases inter-class similarity scores while decreasing the intra-class ones.

Potentially these situations can be dealt with if the knowledge of class distributions is available. In most practical scenarios, learning these distributions is infeasible with just a few (often just one) samples per enrolled identity. It is in these situations that one can make use of large number of non-match biometric samples already present in the database. Normalizing the raw similarity score of the query with the claimed identity using its similarity with the neighbors of the claimed identity provides a sense of class distributions and normalizes for any unwanted peculiarities involved in raw similarity computation. Such a score normalization using neighbors of the claimed identity is termed as cohort analysis. The unimodal framework is also extended to perform biometric fusion to reap the benefits of the availability of multiple evidences and the non-match templates in the database for improved matching performance.

1.4.3 Cancelable face matching

The concern of biometric privacy has led to research efforts to secure biometrics. Unlike PINs or passwords, which can be changed at will when compromised, biometric traits are unique and relatively permanent. One popular way to secure biometrics is to combine biometrics with user-provided keys or passwords. The user-specific private key is used to encrypt biometric template which is stored in the database. The encrypted template stored in the database is used for further matching. For matching purposes, the same encryption scheme is used to transform the query template to compare it with the stored secure template. Quite clearly, such an approach combines the advantages of biometric based authentication and password-based privacy and revocability.

One of the main problems in encryption-based biometric authentication approaches is that they tend to be sensitive to variability/noise in the input biometric space. Inherently, biometrics show a great deal of intra-class variability either due to natural causes or external imaging conditions. It is difficult to design an encryption scheme that can suitably transform features extracted from such input data minimizing within-class scatter as compared to the between-class scatter. Unlike input biometric space, in which one can perform some sort of learning to account for such intra-class variabilities, such learning is not easy in the encrypted space. Another drawback of encrypting feature extracted from the input biometrics is that such approaches tend to be specific to the features used. Therefore, it may not always be easy for such approaches to take advantage of the new developments in the

field of biometric matching. In this dissertation, we propose a physics-based face reconstruction approach that addresses these issues for cancelable face matching. Given an input face image, the proposed technique reconstructs a transformed face image that can be matched using any publicly available matcher. Depending on the capability of the face matcher used to compare the reconstructed face images, the variability/noise in the input biometric can be accounted for even though matching is performed in the transformed domain.

1.4.4 Face recognition and tracking in videos

Traditionally face recognition has been limited to still images. Though great leaps have been made in recognizing faces from still images, more needs to be done to achieve the goal of recognizing faces in uncontrolled scenarios. Still image-based approaches often struggle to truly generalize across variations in pose, expression, illumination, etc., leading to a not so satisfactory performance on real images. The advent of inexpensive cameras and increased processing power has made it possible to capture and store videos in real time. Videos have the advantage of providing more information in the form of multiple frames making it relatively easier to generalize across variations that has been difficult with still images. Video input allows to capture temporal signatures that can be used to characterize and hence identify faces. Moreover, video makes it is easier to track (or segment) faces which can then be fed into a recognition system. Importantly, psychological evidence indicates that dynamic information contributes to face recognition especially under non-optimal

viewing conditions [64]. These reasons form the basis of the recent interest in using videos for recognizing faces [36] [29]. Though video provides extra information, the video feeds are almost always uncontrolled making it challenging to track and hence recognize faces. In this dissertation, we develop two approaches to recognize and track faces in videos as described in the following subsections.

1.4.4.1 A system identification approach to recognize faces in videos

This work treats video-to-video face recognition as a dynamical system identification and classification problem. Video-to-video means that both gallery and probe consists of videos. We model a moving face as a linear dynamical system whose appearance changes with pose. An autoregressive and moving average (ARMA) model is used to represent such a system. The choice of ARMA model is based on its ability to take care of changes in appearance while modeling dynamics of the face. Recognition is performed using the concept of subspace angles to compute distances between the estimated ARMA models corresponding to gallery and probe video sequences. The results obtained are quite promising given the extent of pose, expression and illumination variations in the video data used for the experiments.

1.4.4.2 3D facial pose tracking

One of the main drawback of the proposed ARMA model based approach is that it does not explicitly takes the 3D structure and motion of the face into account. Therefore, the ARMA model-based representation will probably not work

when there is no/limited pose overlap between gallery and probe videos. What is desired here is an approach that can assist recognition of faces in videos even when there is hardly any overlap of poses. To address this, we propose a method to recover the 3D configuration of a face in each frame of a video. The 3D configuration consists of the three translation parameters (recoverable up to a scale factor) and the three orientation parameters which correspond to the yaw, pitch and roll of the face. The approach combines the structural advantages of geometric modeling with the statistical advantages of a particle-filter based inference. The face is modeled as the curved surface of a cylinder with an elliptic cross-section which is free to translate and rotate arbitrarily. The recovered 3D translation and rotation parameters are used to obtain a pose-invariant textural characterization of faces from the input video frames to perform recognition.

1.5 Organization of the dissertation

The rest of the dissertation is organized as follows. The first part of the dissertation describes the algorithms proposed to perform illumination-insensitive matching of face images. The details of the algorithm to recognize faces illuminated by arbitrary number of unknown light sources are provided in Chapter 2. Chapter 3 describes the role bilateral symmetry plays in disambiguating images taken under different illumination conditions. The developed theoretical formulation is backed by experimental results on real images to show the usefulness of the formulation when the assumptions are not strictly satisfied. The proposed cohort framework to

account for unmodeled nuisance parameters that affect the matching performance of a verification/recognition system, is described in Chapter 4. Chapter 5 describes the developed physics-based revocable face reconstruction algorithm to address the privacy issues concerning face recognition. Chapter 6 details the proposed system identification approach to recognize faces in low quality videos. A robust particle-filter based algorithm for 3D tracking of faces is also described in Chapter 6. The dissertation concludes with a summary, discussion and concluding remarks in Chapter 7.

Chapter 2

Face Recognition in the Presence of Multiple Light Sources

There are quite a few problems in computer vision like Shape-from-Shading (SFS), illumination-invariant image matching, etc. that are inherently ill-posed in their full generality. Quite often the intractability of the problem can be reduced significantly by restricting the domain of the problem and using appropriate constraints. For example, the symmetry constraint reduces the inherent ambiguities in the traditional SfS solution to a large extent [97]. In this chapter, we deal with such an intractable problem of illumination-invariant matching with a focus on human faces. In particular, we propose a solution to this problem for a class of objects under the assumption that the class consists of linear Lambertian objects. The case of multiple light sources is handled by combining the linear Lambertian assumption with a linear approximation to the subspace of Lambertian images. This aids in recognizing faces illuminated by multiple unknown light sources using just a single image.

2.1 Organization of the chapter

The following section provides a survey of the existing approaches that address the problem of illumination-insensitive matching of images. Section 2.3 illustrates the importance of the nonlinearity in Lambert's law. Section 2.4 describes the con-

cept of linear Lambertian object. The importance of the nonlinearity is further highlighted in Section 2.5 by showing improvements in the recognition accuracy when the nonlinearity is taken into account. The face recognition algorithm to handle faces illuminated by multiple light sources is presented in Section 2.6. Section 2.7 describes several challenging experiments performed to rigorously test the approach. The chapter concludes with a brief summary and discussion in Section 2.8.

2.2 Literature survey

Recent improvements in the accuracy of the face recognition algorithms for images taken under controlled conditions has shifted the focus to more challenging tasks of achieving the same performance for uncontrolled scenarios. A detailed survey of various face recognition algorithms is presented in [98]. Several researchers attempt to achieve invariance to illumination by using image processing techniques like histogram equalization [82]. Some subspace based methods try to counter illumination variations by discarding the first few principal components [8]. These techniques do improve the accuracies of the respective algorithms but are usually ineffective in the case of a non-trivial change in illumination conditions.

The inability of such heuristics to handle illumination variation has led to the rise of generative (or analysis-by-synthesis) approaches for face recognition [13, 99, 94, 33, 77]. Broadly speaking, these techniques try to model the physical process of image formation by taking into consideration quantities like surface albedo, surface normals and illumination source direction. Though the recovery of shape and surface

properties (reflectivity or albedo) from image(s) has been studied for a long time, its application to the problem of face recognition is fairly recent. An example in this category is the application of SFS algorithms. SFS research typically assumes a constant albedo across an object which is usually not true and thus limits the use of the approach. Since then, there have been several advances which have led to the application of SFS for face recognition and rendering. Zhao *et al.* [97] present an SFS approach to recover both shape and albedo for a symmetric object from a single image. [94] uses singular value decomposition (SVD) to learn generative models of objects from a set of images taken under different, and unknown illuminations. Shashua *et al.* [77] perform recognition across varying illumination under an ideal-class assumption. All objects belonging to the ideal class are assumed to have the same shape. [33] uses illumination cone models for illumination-invariant face recognition. They require a small number of training images of each face under different illuminations to recover the shape and albedo of the face. Basri *et al.* [6] propose methods for recovering surface normals in a scene using images taken under general illumination conditions. Their work is based on [7, 68] which prove that the set of all Lambertian reflectance maps obtained with arbitrary distant illumination sources approximately lie in a 9D linear subspace. In [13], Blanz *et al.* perform face recognition across pose and illumination by fitting a 3D morphable model to images. They use a set of textured 3D scans of heads for learning the model. [95] uses harmonic image exemplars to perform face recognition under varying lighting. Zhou *et al.* [99] generalize the traditional photometric approach to handle all the appearances of all the objects in a class. They impose a rank constraint on shapes

and albedos in a class to separate the two from illumination using the factorization approach.

Despite the advances made, most of the cited approaches have not been applied for the face recognition problem using a large database. This might be because many techniques require multiple, independently illuminated images of each face which are usually not present in most face datasets. Moreover, most of the approaches make single light source assumption which does not hold in most real conditions. The assumption might be driven by the unavailability of suitable datasets to test approaches which can potentially handle images illuminated by multiple number of light sources.

In this chapter, we propose an algorithm to perform face recognition across varying illumination (using a single face image) for images illuminated by multiple number of light sources. The algorithm does not need any prior information about the number or placements of the light sources. We are not aware of any standard controlled dataset containing faces illuminated by two or more number of light sources, which can be used to study the effect of the single light source assumption on face recognition algorithms. We generate such data using faces from PIE [79] and Yale Face Database B [33]. Experimental results are presented to confirm the efficacy of the approach. The proposed algorithm performs much better than its counterpart that makes the single light source assumption.

2.3 How important is the nonlinearity in Lambert's law?

The algorithm models a face as a Lambertian surface. Therefore, it is worthwhile to address the issue of the often ignored nonlinearity in Lambert's law before explaining the algorithm. In fact the performance of the algorithm can get adversely affected if the nonlinearity is ignored (thereby allowing pixel intensities to take negative values).

The diffuse component of the reflection of a surface is often modeled using the popular Lambert's Law [41]. For example, Blanz and Vetter[13] use the following equation to model the diffuse component

$$L_{r,k} = R_k . L_{r,dir} . < \boldsymbol{n}_k, \boldsymbol{l} >, \tag{2.1}$$

where R_k is the red component of the diffuse reflection coefficient, $L_{r,dir}$ is the red channel of the directed light, $\boldsymbol{n}_k$ is the surface normal and $\boldsymbol{l}$ is the light source direction. Similarly, the generalized photometric stereo method [99] uses

$$h = \rho \boldsymbol{n}^T \boldsymbol{s}, \tag{2.2}$$

where ρ is the surface albedo, $\boldsymbol{n}$ is the surface normal, and $\boldsymbol{s}$ is the light source direction (multiplied by the intensity), as the rule for image formation. A close look at these equations reveals that a linear approximation to the Lambert's law is assumed in both these models. If used in its pure form, the nonlinearity in the Lambert's law would have made (2.2) to be

$$h = \rho \max(\boldsymbol{n}^T \boldsymbol{s}, 0) \tag{2.3}$$

Quite clearly, the linearity assumption is perfectly valid as long as the directed light source is in front of the surface for all its points. In general, objects like faces do not have all the surface points facing the illumination source which leads to the formation of shadows (commonly known as form/attached shadows). The cast and attached shadows are often ignored from the analysis to keep the subspace of the observed images in a three [76] or with the addition of an ambient component [94], four dimensional linear subspace. Therefore, several generative approaches either ignore this nonlinearity completely or try to somehow ignore the shadow pixels. Here we present a simple illustration to highlight the importance of the nonlinearity in the Lambert's law.

2.3.1 Illustration 1

Suppose the goal is to estimate the illumination source direction from a single face image given the shape and albedo of the face (assuming the image has been illuminated by a single distant illumination source). We explore three approaches for this task: the first approach ignores the nonlinearity completely, the second one uses the linear rule but ignores the shadow pixels and the last one uses the Lambert's law in its pure form. The accuracy of the global minimum and its ambiguity on the error surface is taken as the criterion for the goodness of the method. The analytical expressions for the error function using the three options can be written as :

$$\text{Completely linear:} \qquad \mathcal{E}(\boldsymbol{s}) = \| \boldsymbol{h} - \boldsymbol{\rho}\boldsymbol{n}^T\boldsymbol{s} \|^2 \tag{2.4}$$

$$\text{Shadow pixels ignored:} \qquad \mathcal{E}(\boldsymbol{s}) = \| \boldsymbol{\tau} \circ (\boldsymbol{h} - \boldsymbol{\rho}\boldsymbol{n}^T\boldsymbol{s}) \|^2 \tag{2.5}$$

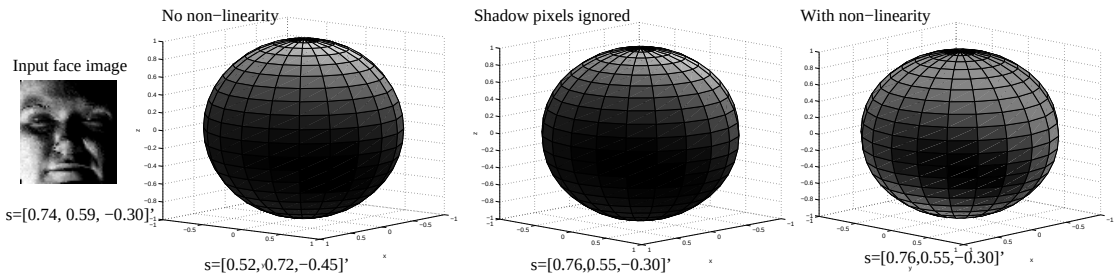

Figure 2.1: The error surfaces for the estimation of the light source direction for a given face image. The plots correspond to the three approaches described in Illustration 1. The lower the error is for a particular illumination direction, the darker the error sphere looks at the point corresponding to that direction. The true and estimated values of the illumination direction are listed along with the plots.

$$\text{Non-linear rule:} \qquad \mathcal{E}(\boldsymbol{s}) = \| \boldsymbol{h} - \max(\boldsymbol{\rho}\boldsymbol{n}^T\boldsymbol{s}, 0) \|^2 \tag{2.6}$$

where, $\mathcal{E}(\boldsymbol{s})$ is the error with $\boldsymbol{s}$ as the illumination source direction, $\boldsymbol{h}_{d\times 1}$ is the vectorized input image, $\boldsymbol{\rho}_{d\times 1}$ is the albedo vector, $\boldsymbol{n}_{3\times d}$ contains the surface normals, and $\boldsymbol{\tau}_{d\times 1}$ is the shadow indicator vector which is 0 for the shadow pixels and 1 for the rest. Clearly, the linear method penalizes the correct illumination at the shadow pixels by having non-zero error values for those pixels. On the other hand, when shadows are ignored, the illuminations which produce wrong values for the shadow pixels do not get penalized there. As the set of all possible normals lies on the surface of a unit sphere, we use a sphere to display the computed error functions. Figure 2.1 shows the error surfaces for the three methods for a given face image. The lower the error is for a hypothesized illumination direction s, the darker the surface looks at the corresponding point on the sphere. The global minimum is far from

the true value using the first approach but is correct up to a discretization error for the second and third approaches. In fact, the second and third methods will always produce the same global minimum (assuming $\boldsymbol{\tau}$ is correct), but the global minimum will always be less ambiguous in the third case because several wrong hypothesized illumination directions do not get penalized enough in the second approach due to the exclusion of the shadow pixels (Figure 2.1).

2.3.2 The case of multiple light sources

The analysis in Illustration 1 implicitly assumes that there is only one distant light source illuminating the face. Though the assumption is valid for datasets like PIE, it does not hold for most realistic scenarios. We now explore the impact of using the *linear* Lambert's law for images illuminated by multiple light sources. Using the *linear* Lambert's law, an image illuminated by k light sources can be represented as:

$$\boldsymbol{h} = \sum_{i=1}^{k} \boldsymbol{\rho}\boldsymbol{n}^T\boldsymbol{s}_i = \boldsymbol{\rho}\boldsymbol{n}^T \sum_{i=1}^{k} \boldsymbol{s}_i = \boldsymbol{\rho}\boldsymbol{n}^T\boldsymbol{s}^\star \tag{2.7}$$

where, $\boldsymbol{s}^\star = \sum_{i=1}^{k} \boldsymbol{s}_i$. This shows that under the linear assumption, multiple light sources can be replaced by a suitably placed single light source without having any effect on the image. This is a bit counter-intuitive as can be seen in a simple two source scenario where $\boldsymbol{s}_1 = -\boldsymbol{s}_2$

$$\boldsymbol{h} = \boldsymbol{\rho}\boldsymbol{n}^T(\boldsymbol{s}_1 - \boldsymbol{s}_2) = 0 \tag{2.8}$$

Thus the linear assumption can make the effect of light sources interfere in a destructive manner and give bizarre outcomes. Quite clearly, the harm done by the

linearity assumption is proportional to the angle subtended by the light sources at the surface.

Though the discussion in Illustration 1 suggests that Lambert's law in its pure form is better suited for illumination estimation than the other variants, it is only of academic interest if inclusion of the nonlinearity does not improve the recognition results. The following sections describe the proposed algorithm for illumination-insensitive matching of faces that incorporates the often ignored nonlinearity in Lambert's law. The improvement in the recognition accuracy over existing approaches highlights the importance of including the attached shadows in the analysis.

2.4 Linear Lambertian object

Definition: A *linear Lambertian object* is defined as a visual object *simultaneously* obeying the following two properties:

- It is linearly spanned by basis objects.
- It follows the Lambertian reflectance model with a varying albedo field.

While each of the above two properties has been widely studied in the literature for various tasks, the concept of linear Lambertian object which captures both the characteristics, is effective for illumination-invariant matching as shown in this chapter. An example of linear Lambertian object is human face[1], which is the focus of

[1]One may argue that specular properties of skin and eyes and the reflectance properties of hair violate the Lambertian assumption. However, the hair is excluded by preprocessing and the pixels in specular regions, unless significantly large, do not have disastrous effect on the results as

this work. The linearity [86, 8, 88] characterizes the appearances of an image ensemble for a class of objects. It assumes that an image $\boldsymbol{h}$ is expressed as a linear combination of basis images $\boldsymbol{h}_i$, i.e.,

$$\boldsymbol{h} = \sum_{i=1}^{m} f_i \boldsymbol{h}_i, \tag{2.9}$$

where f_i's are blending coefficients. In other words, the basis images span the image ensemble. Typically, the basis images are learned using the images not necessarily illuminated under the same lighting condition. This forces the learned basis images to inadequately cover variations in both identity and illumination.

The Lambertian reflectance model [38, 76, 7] with a varying albedo field is widely used in the literature to depict the appearance of certain matte objects such as faces. It assumes that a pixel h is represented as

$$h = \max(\rho \boldsymbol{n}_{3\times1}^T \boldsymbol{s}_{3\times1}, 0) \tag{2.10}$$

where $[.]^T$ denotes the transpose, ρ is the albedo at the pixel, $\boldsymbol{n}$ is the unit surface normal vector at the pixel, and $\boldsymbol{s}$ (a 3×1 unit vector multiplied by its intensity) specifies a directional light source. When the pixel is not directly facing the light source, the attached shadow is generated, i.e., the zero intensity is achieved. Another kind of shadow is the cast shadow that is generated when the light source is blocked by other pixel due to object geometry.

An image $\boldsymbol{h}$ is a collection of d pixels $\{h_i, i = 1, ..., d\}$ [2]. By stacking all the

observed in the experiments.

[2]The index i corresponds to a spatial position $x = (x, y)$. We will interchange both notations. For instance, we might also use $x = 1, ..., d$.

pixels into a column vector, we have

$$\boldsymbol{h}_{d\times 1} = [h_1, h_2, \ldots, h_d]^T = \max([(\rho_1 \boldsymbol{n}_1^T)\boldsymbol{s}, \ldots, (\rho_d \boldsymbol{n}_d^T)\boldsymbol{s}]^T, 0) = \max(\boldsymbol{T}_{d\times 3}\boldsymbol{s}_{3\times 1}, 0), \tag{2.11}$$

where the $\boldsymbol{T} = [\rho_1 \boldsymbol{n}_1, \rho_2 \boldsymbol{n}_2, \ldots, \rho_d \boldsymbol{n}_d]^T$ matrix encodes the product of the albedos and the surface normal vectors for all d pixels. Evidently, the Lambertian model is specific to the object and consequently, we call the $\boldsymbol{T}$ matrix as an *object-specific albedo-shape* matrix.

The process of combining the above two properties is equivalent to imposing the restriction of the same light source on the basis images, with each basis image expressed as $\boldsymbol{h}_i(\boldsymbol{s}) = \max(\boldsymbol{T}_i\boldsymbol{s}, 0)$. Therefore, Eq. (2.9) becomes

$$\boldsymbol{h} = \sum_{i=1}^{m} f_i \max(\boldsymbol{T}_i s, 0). \tag{2.12}$$

This is the *generative* model for the linear Lambertian object. It is evident that the key difference between linear Lambertian object and conventional subspace analysis is in the basis image $\boldsymbol{h}_i$: For a linear Lambertian object, it is now a function of the light source $\boldsymbol{s}$ through the matrix $\boldsymbol{T}_i$. We denote all the matrices $\boldsymbol{T}_i$ compactly by $\boldsymbol{W} = [\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_m]$. Since the $\boldsymbol{W}$ matrix encodes all albedos and surface normals for a class of objects, we call it a *class-specific albedo-shape* matrix.

The linear Lambertian object concept provides a unique opportunity to study the appearances of an image ensemble for a class of objects under illumination variations and opens the door to many applications like generalized photometric stereo [99] and illumination-insensitive matching of faces.

2.5 Face recognition across varying illumination (single light source)

In this section, we use the linear Lambertian assumption on human faces to devise an approach to perform face recognition across illumination variations. Though the main focus of our research is to recognize faces illuminated by multiple unknown light sources, we start with single light scenario to show the importance of incorporating the nonlinearity in Lambert's law in addition to the usefulness of the proposed linear Lambertian formulation for the task of illumination-insensitive face matching. From the linear Lambertian generative model in (2.12), if the nonlinearity is ignored, one can omit the max function as follows

$$\boldsymbol{h}_{d\times 1} = \boldsymbol{T}\boldsymbol{s} = \sum_{i=1}^{m} f_i \boldsymbol{h}_i \tag{2.13}$$

$$= \sum_{i=1}^{m} f_i \boldsymbol{T}_i \boldsymbol{s} = \boldsymbol{W}_{d\times 3m}[\boldsymbol{f}_{m\times 1} \otimes \boldsymbol{s}_{3\times 1}], \tag{2.14}$$

where '$\otimes$' denotes the Kronecker (tensor) product. Because $\boldsymbol{s}$ is a free parameter, Eq. (2.14) is equivalent to imposing a rank constraint on the $\boldsymbol{T}$ matrix: any $\boldsymbol{T}$ matrix is a linear combination of some basis matrices $\{\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_m\}$ coming from some *m basis objects*.

$$\boldsymbol{T}_{d\times 3} = \sum_{j=1}^{m} f_j \boldsymbol{T}_j. \tag{2.15}$$

Obviously, the blending coefficients f_j characterize the identity of the object and are invariant to illumination by construction. Note that the difference between these relations and linear Lambertian model is that these relations ignore the hard nonlinearity in Lambert'a law that accounts for the formation of attached shadows

by omitting the max function.

Given shape-albedo matrix $\boldsymbol{W}$, the recovery of the identity vector $\boldsymbol{f}$ and illumination $\boldsymbol{s}$ can be posed as an optimization problem as follows [99]

$$[Problem\ \ A]\ \ \min_{\boldsymbol{f},\boldsymbol{s}} \mathcal{E}(\boldsymbol{f},\boldsymbol{s}) = \|\boldsymbol{\tau} \circ (\boldsymbol{h} - \boldsymbol{W}\ (\boldsymbol{f} \otimes \boldsymbol{s}))\|^2 + (\mathbf{1}^T\boldsymbol{f} - 1)^2, \qquad (2.16)$$

Here $\boldsymbol{\tau}$ denotes the shadow indicator variable. Such a variable is useful to omit shadow pixels from the analysis when nonlinearity responsible for attached shadow formation is not incorporated in the formulation. On the other hand, if the nonlinearity is incorporated in the analysis, the recovery of the identity vector $\boldsymbol{f}$ can be posed as the following optimization problem

$$[Problem\ \ B]\ \ \min_{\boldsymbol{f},\boldsymbol{s}} \mathcal{E}(\boldsymbol{f},\boldsymbol{s}) = \| \boldsymbol{h} - \sum_{i=1}^{m} f_i \max(\boldsymbol{T}_i\boldsymbol{s}, 0) \|^2 + (\mathbf{1}^T\boldsymbol{f} - 1)^2 \qquad (2.17)$$

As the formation of attached shadows is already accounted for, there is no need of any shadow indicator variable. As shown later, such a strategy significantly outperforms the one that ignores the nonlinearity in Lambert's law. The second term is included in the error function to take care of scale ambiguity between $\boldsymbol{f}$ and $\boldsymbol{s}$. Please note that $\boldsymbol{s}$ is not a unit vector as it contains the intensity of the illumination source also.

The formulation in [99] follows a similar approach solving Problem A that uses the *linear* version of Lambert's law. Given n different objects under different (and unknown) illumination conditions, Zhou *et al.* [99] estimate $\boldsymbol{W}$ (up to an invertible matrix) by solving a rank $3m$ problem using the factorization approach. The ambiguity is resolved using symmetry and integrability constraints. The interested reader is referred to [99] for the complete derivation. The average recognition results

reported in [99] improve from 67% to 93%, when the $\boldsymbol{W}$ matrix is estimated from Vetter's 3D dataset [13] instead of the approach mentioned above.

Our goal here is to highlight the importance of the nonlinearity in Lambert's law in addition to usefulness of the linear Lambertian formulation for the task of face recognition across illumination variations. Therefore, we generate the shape-albedo matrix $\boldsymbol{W}$ using Vetter's 3D dataset for all our experiments. As opposed to [99], we take into account the inherent hard nonlinearity present in Lambert's law.

The minimization in Problem B is performed using an iterative approach, fixing $\boldsymbol{f}$ for optimizing $\mathcal{E}$ w.r.t. $\boldsymbol{s}$ and fixing $\boldsymbol{s}$ for optimization w.r.t. $\boldsymbol{f}$. In each iteration, $\boldsymbol{f}$ can be estimated by solving a linear least-squares (LS) problem but a non-linear LS solution is required to estimate $\boldsymbol{s}$. The non-linear optimization is performed using the *lsqnonlin* function in MATLAB which is based on the interior-reflective Newton method. For most faces, the function value did not change much after 4-5 iterations. Therefore, the iterative optimization was always stopped after 5 iterations. The whole process took about 5-7 seconds per image on a standard desktop.

We perform recognition experiments across illumination using the frontal faces from the PIE dataset. The correlation coefficient of the identity vectors is taken as the measure of the similarity between face images. Table 2.1 shows the recognition results obtained using this approach. Recognition is performed across illumination with images from one illumination condition from the PIE dataset forming the gallery while images from another illumination condition forming the probe set. Each gallery/probe set contains one frontal image per subject taken in the presence

of a particular light source (there are 68 subjects in each gallery/probe set). Each entry in the table shows the recognition rate achieved for one such choice of gallery and probe set. The averages from [99] are shown for comparison. For fair comparison, we show results only across the illumination scenarios displayed in [99]. The recognition performance with the inclusion of the non-linearity in Lambert's law is almost always better or same. The overall average performance is up from 93% to 96%. The improvement is significant in cases involving difficult illumination conditions (with lots of shadows) like the flash f_{17} in the PIE dataset. This shows that though the estimation becomes slightly more difficult, the recognition rate improves with inclusion of the non-linearity.

2.6 Illumination-insensitive face recognition in the presence of multiple light sources

One of the issues in handling multiple illumination case is the prior knowledge of the number of light sources. In the absence of this knowledge, one can hypothesize several different cases and choose the one with minimum residual error. This can potentially be done in a manner very similar to the approach described for the single illumination source case with the following change in the objective function

$$[Problem\ \ C]\quad \min_{\boldsymbol{f},\boldsymbol{s}} \mathcal{E}(\boldsymbol{f},\boldsymbol{s}) = \| \boldsymbol{h} - \sum_{i=1}^{m} f_i \sum_{j=1}^{k} \max(\boldsymbol{T}_i \boldsymbol{s}_j, 0) \|^2 + (\mathbf{1}^T \boldsymbol{f} - 1)^2 \quad (2.18)$$

where k is the hypothesized number of light sources. The objective function can be minimized repeatedly for different values of k and the one with minimum error can be taken as the correct hypothesis. Figure 2.2 shows the variation of the error with

Gallery	f_{08}	f_{09}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{20}	f_{21}	f_{22}	Avg	Avg from [99]
Probe														
f_{08}	-	F	F	F	96	97	81	72	50	F	97	84	90	88
f_{09}	F	-	F	F	F	99	97	96	75	F	F	97	97	94
f_{11}	F	F	-	F	F	97	94	78	63	F	99	94	94	93
f_{12}	F	F	F	-	F	F	F	99	90	F	F	F	99	97
f_{13}	97	F	F	F	-	F	F	F	96	F	F	F	99	99
f_{14}	94	F	F	F	F	-	F	F	99	F	F	F	99	99
f_{15}	88	97	97	F	F	F	-	F	F	97	F	F	98	96
f_{16}	74	90	81	93	F	F	F	-	F	76	97	F	93	89
f_{17}	59	74	63	87	99	99	F	F	-	71	94	F	87	75
f_{20}	99	F	F	F	F	99	96	82	71	-	F	97	95	93
f_{21}	97	F	F	F	F	F	F	99	96	F	-	F	99	98
f_{22}	93	F	99	F	F	F	F	F	99	99	F	-	99	98
Average	92	97	95	98	100	99	97	94	87	95	99	98	96	-
Average from [99]	89	93	92	96	98	99	96	91	80	91	96	98	-	93

Table 2.1: Recognition results on the PIE dataset. The averages from [99] that ignores the nonlinearity in Lambert's law, are included for comparison. $\boldsymbol{f}_i$ denotes images taken with a particular flash ON as labeled in PIE. Each $(i, j)^{th}$ entry in the table shows the recognition rate obtained with the images from $\boldsymbol{f}_j$ as gallery while from $\boldsymbol{f}_i$ as probes. F denotes perfect recognition score.

k, for an image illuminated by three different light sources. As can be seen, the error more or less stabilizes for $k \geq 3$. As the parameter spaces are nested, ideally the error plot should be a non-increasing function of k, but the increase in complexity of the non-linear optimization can make the plot behave otherwise. Please note that for the *linear* Lambert's law, such a curve will look more or less horizontal due to the equivalence of the single and multi-light source scenarios (Equation 2.7) under the linear assumption.

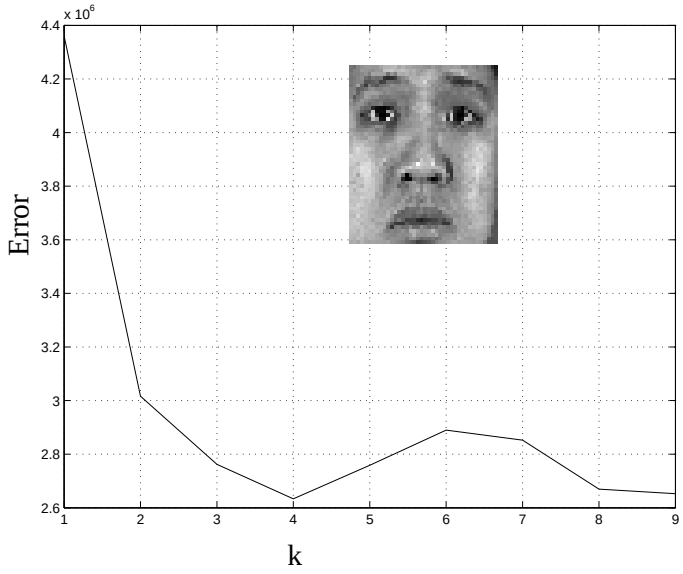

Figure 2.2: The error obtained for different hypothesized number of light sources. The face was illuminated using three light sources.

Though one can use this approach by varying k, it is not elegant and computationally intensive. In our approach, we avoid the extra computations by making the following assumption. We assume that an image of an arbitrarily illuminated face can be approximated by low dimensional linear subspace [7] that can be generated by a linear combination of the images of the same face in the same pose, illuminated by nine different light sources placed at pre-selected positions. Lee *et al.* [49] show that this approximation is quite good for a wide range of illumination conditions. Hence, a face image can be written as

$$\boldsymbol{h} = \sum_{i=1}^{9} \alpha_i \boldsymbol{h}_i \tag{2.19}$$

$$\text{where,} \qquad \boldsymbol{h}_i = \max(\boldsymbol{\rho n}^T \hat{\boldsymbol{s}}_i, 0) = \max(\boldsymbol{T}\hat{\boldsymbol{s}}_i, 0) \tag{2.20}$$

$\{\hat{\boldsymbol{s}_1}, \hat{\boldsymbol{s}}_2, \ldots, \hat{\boldsymbol{s}}_9\}$ are the pre-specified illumination directions. As proposed in [49],

we use the following directions for $\{\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, \ldots, \hat{\boldsymbol{s}}_9\}$:

$$\begin{aligned} \boldsymbol{\phi} &= \{0, 49, -68, 73, 77, -84, -84, 82, -50\}^\circ \\ \boldsymbol{\theta} &= \{0, 17, 0, -18, 37, 47, -47, -56, -84\}^\circ \end{aligned} \tag{2.21}$$

Under this formulation, (2.18) changes to

$$[Problem\ \ D]\ \ \min_{\boldsymbol{f},\boldsymbol{s}} \mathcal{E}(\boldsymbol{f}, \boldsymbol{s}) = \| \boldsymbol{h} - \sum_{i=1}^{m} f_i \sum_{j=1}^{9} \boldsymbol{\alpha}_j \max(\boldsymbol{T}_i \hat{\boldsymbol{s}}_j, 0) \|^2 + (\mathbf{1}^T \boldsymbol{f} - 1)^2 \tag{2.22}$$

This way one can potentially recover the illumination-free identity vector $\boldsymbol{f}$ without any prior knowledge of the number of light sources or any need to check different hypotheses for the same.

Now the objective function is minimized with respect to $\boldsymbol{f} = [f_1, f_2, \ldots, f_m]$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_9]$. This gives us the illumination-free identity vector $\boldsymbol{f}$ which is used for recognition. The optimization is done in an iterative fashion by fixing one parameter and estimating the other and vice-versa as shown below.

By Defining a $d \times m$ matrix $\boldsymbol{W}_f$ as

$$\boldsymbol{W}_f = \left[\sum_{j=1}^{9} \alpha_j \max(\boldsymbol{T}_1 \hat{\boldsymbol{s}}_j, 0), \sum_{j=1}^{9} \alpha_j \max(\boldsymbol{T}_2 \hat{\boldsymbol{s}}_j, 0), \ldots, \sum_{j=1}^{9} \alpha_j \max(\boldsymbol{T}_m \hat{\boldsymbol{s}}_j, 0) \right]_{d \times m}, \tag{2.23}$$

it is easy to show that

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{W}_f \\ \mathbf{1}^T \end{bmatrix}^\dagger \begin{bmatrix} \boldsymbol{h} \\ 1 \end{bmatrix}. \tag{2.24}$$

where $\boldsymbol{h}_{d \times 1}$ is the vectorized input face image, $[.]^\dagger$ is the Moore-Penrose inverse, and $\mathbf{1}_{1 \times m}$ is the m-dimensional vector of ones, included to handle scale ambiguity between $\boldsymbol{f}$ and $\boldsymbol{\alpha}$.

Looking carefully at the objective function (2.22), one can easily observe that $\boldsymbol{\alpha}$ too can be estimated by solving a linear LS problem (as $\{\hat{\boldsymbol{s}}_1\hat{\boldsymbol{s}}_2 \ldots \hat{\boldsymbol{s}}_9\}$ is known). This avoids the need for any nonlinear optimization here. Recall that a nonlinear LS method was required to estimate $\boldsymbol{s}$ in the approach proposed for the single light source case. The expression for $\boldsymbol{\alpha}$ can be written as:

$$\boldsymbol{\alpha} = \boldsymbol{W}_{\alpha}^{\dagger} h \tag{2.25}$$

where,

$$\boldsymbol{W}_{\alpha} = \left[\sum_{i=1}^{m} f_i \max(\boldsymbol{T}_i\hat{\boldsymbol{s}}_1, 0), \sum_{i=1}^{m} f_i \max(\boldsymbol{T}_i\hat{\boldsymbol{s}}_2, 0), \ldots, \sum_{i=1}^{m} f_i \max(\boldsymbol{T}_i\hat{\boldsymbol{s}}_9, 0)\right]_{d\times 9}.$$

For most of the face images, the iterative optimization converged within 5-6 iterations. As there is no non-linear optimization involved, it took just 2-3 seconds to recover $\boldsymbol{f}$ and $\boldsymbol{\alpha}$ from a given face image on a normal desktop. As the identity variable is estimated from an image by separating the effect of all the light sources in the form of $\boldsymbol{\alpha}$, it is used as the illumination-invariant representation for recognition across varying illumination. The correlation coefficient of the identity vectors is used as the similarity measure for recognition experiments.

2.7 Experiments and results

To begin with, we test this algorithm by running the same experiment as we do for the single light source approach. Though the PIE dataset is not suited to test the ability of this algorithm to handle arbitrarily illuminated images, a good performance here can be considered as a proof of concept. The overall average

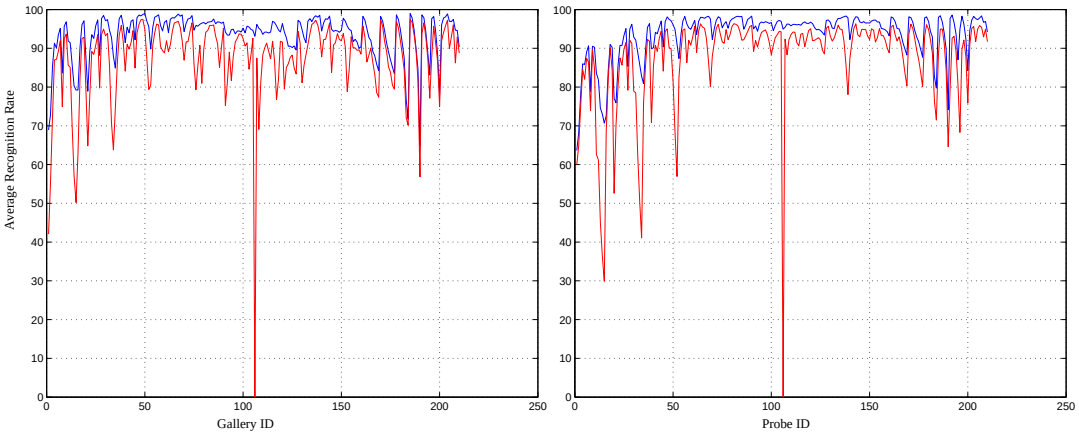

Figure 2.3: The per-gallery and per-probe set average recognition rates on the 210 doubly-illuminated scenarios generated from the PIE dataset. The blue curve (the darker one on the top for monochrome version) shows the performance of the proposed approach while the red curve (the lighter one for monochrome version) corresponds to the linear single light source approach [99].

recognition rate for the experiment obtained using this algorithm is 95% which is higher than [99].

Due to the unavailability of a standard dataset containing face images with multiple light sources ON at a time, we generate such a dataset using the PIE and Yale datasets. Due to the controlled nature of the datasets, multiple images of a subject under different illuminations but same pose, are more or less aligned. If we ignore any camera gain, this allows us to add multiple images of a person taken under different illuminations to get one with the effect of an image captured with multiple lights ON. The images generated this way look reasonably realistic (see Figure 2.4).

We perform experiments on the dataset created by adding images from two illumination conditions from PIE at a time. As PIE has 21 different illumination scenarios, we get a total of ${}^{21}C_2 = 210$ different *doubly-illuminated* scenarios. Recognition was done across all 210 scenarios by taking one as the gallery and another one as the probe set at a time to get 210×209 recognition scores. As it is difficult to show the recognition scores by drawing a 210×210 table, we show only the aggregated per-gallery and per-probe set recognition rates (similar to the averages in Table 2.1) in Figure 2.3. The blue curve (the darker one on the top for monochrome version) on the top shows the averages obtained by the proposed approach. For comparison, we show the recognition rates obtained on this dataset using Zhou *et al.*'s linear method [99] that ignores shadow pixels under the single light source assumption (red curve - the lighter one for monochrome version). For ease of use, we will call this method as ISP-SLS (Ignores Shadow Pixels under Single Light Source

Figure 2.4: The doubly-illuminated images of a subject from the Yale database. Each image is generated by adding 2 images of the same subject illuminated by different light sources.

Assumption). There exist a zero in the red curve because for one gallery/probe set, the method ended up ignoring most of the pixels as shadows and thus was unable to recover the identity variable. The recognition rates obtained using the proposed approach are always better or same as compared to ISP-SLS. The increase in the recognition accuracy is more prominent for the cases where the two illumination sources combined to generate the doubly-illuminated scenario were far apart. This happens because the destructive interference of two light sources (due to the linearity assumption in ISP-SLS) increases with an increase in the angle between the two.

We further test the algorithm by generating a similar *doubly-illuminated* data using Yale Face Database B [33]. Figure 2.5 shows the six challenging illumination conditions used to generate fifteen different scenarios (shown in Figure 2.4) by pairing two at a time. The average recognition rate achieved on this difficult data (Figure 2.4 shows images of one subject under the 15 illumination conditions) using our algorithm is 77%. This is up by more that 25% compared to the accuracy achieved both by ISP-SLS method and the method which takes the non-linearity into account under the single light source assumption.

Figure 2.5: The 6 illumination conditions from the Yale face Database B used to generate the doubly-illuminated data.

All the above experiments implicitly assume that the faces in the gallery and probe set are illuminated by the same number of light sources. Clearly, the proposed algorithm does not impose any such restriction. Therefore, we perform another experiment to test the ability of the proposed approach to generalize across varying number of light sources. We generate five illumination scenarios using the PIE dataset with the number of light sources (added to create each scenario) ranging from 1-5. To avoid any bias, the combinations of the light sources are selected randomly from the 21 illumination sets in the PIE dataset. Recognition is performed across the five scenarios by considering one among them as the gallery and another one as the probe set at a time. As before, each gallery/probe set contains one image for each of the 68 subjects present in the PIE dataset. While the ISP-SLS approach performs badly in this experiment, the proposed approach does a perfect job as shown in Table 2.2. Figures 2.6 and 2.7 show the reconstructed surfaces for a face illuminated in the presence of the five illumination scenarios using the two approaches. The quality of the reconstructions explains the difference in the recognition accuracy obtained using these two methods.

The difference between the ISP-SLS and the proposed multiple light source

Figure 2.6: The reconstructed shapes of a face using the ISP-SLS approach. Each column displays the 3 components of the reconstructed surface normals. Columns 1-5 correspond to the five illumination scenarios with the number of light sources varying from 1-5, respectively. The quality of the reconstructed surface degrades as the number of light sources increase.

method can also be highlighted using image relighting examples as shown in Figure 2.8, where there are five illumination scenarios with increasing number of light sources from top to bottom. To confirm the authenticity of the results, we perform another similar experiment with 10 different scenarios with the number of randomly selected light sources (added to generate the 10 scenarios) ranging from 1-10. Here, the proposed approach achieves average recognition accuracy of 99.7% (The average recognition rate achieved by ISP-SLS here is 54%).

2.8 Summary and conclusion

In this chapter, we proposed an algorithm to recognize *multiply*-illuminated faces from just one image. The algorithm has the capability to generalize across

Gallery	$\{f_{20}\}$	$\{f_{05}, f_{22}\}$	$\{f_{20}, f_{06}, f_{18}\}$	$\{f_{21}, f_{06}, f_{07}, f_{03}\}$	$\{f_{03}, f_{15}, f_{06}, f_{19}, f_{05}\}$
Probe					
$\{f_{20}\}$	- / -	100 / 100	100 / 66	100 / 26	100 / 26
$\{f_{05}, f_{22}\}$	100 / 100	- / -	100 / 62	100 / 28	100 / 25
$\{f_{20}, f_{06}, f_{18}\}$	100 / 93	100 / 91	- / -	100 / 72	100 / 74
$\{f_{21}, f_{06}, f_{07}, f_{03}\}$	100 / 62	100 / 66	100 / 90	- / -	100 / 93
$\{f_{03}, f_{15}, f_{06}, f_{19}, f_{05}\}$	100 / 66	100 / 66	100 / 93	100 / 93	- / -

Table 2.2: Recognition results on the multiply-illuminated data generated from the PIE dataset. The various scenarios differ in the number of light sources. The flash Ids from PIE randomly selected to generate each scenario are shown in curly braces. The 1st number shows the recognition rate obtained using our approach while the 2nd number shows the performance of the ISP-SLS method.

Figure 2.7: The reconstructed shapes of a face using our approach. As in Figure 2.6, each column shows the 3 components of the reconstructed surface normals. There is hardly any difference in the reconstructed surfaces across different illuminations scenarios.

face images taken in the presence of varying number of unknown light sources. The approach performs well and outperforms the ISP-SLS approach which uses the linear version of the Lambert's law. Though the comparison with the ISP-SLS approach might not seem fair as it does not try to model multiple light sources, the comparison does reflect the effect, the single light source assumption might have, for recognizing faces in real conditions. Moreover, we illustrated that if the Lambert's law is assumed to be linear, the single and multiple light scenarios are equivalent. Therefore, one can infer that the relaxation of the non-linearity in the Lambert's law is harmful for recognizing faces illuminated by an arbitrary number of light sources. Almost perfect performance of the proposed approach in experiments involving large number of light sources conforms to the belief that face recognition gets easier with an increase in number of light sources.

(a)

(b)

Figure 2.8: Image relighting/rendering results using (a) ISP-SLS and (b) multiple light source algorithms. For each row, the left image is the reference image used for estimating the surface normals and albedos and the remaining nine images are rendered ones corresponding to the nine lighting conditions in Eq. (2.21).

Chapter 3

Symmetric Objects are Hardly Ever Ambiguous

Given any two images taken under different illumination conditions, there always exist a physically realizable object which is consistent with both the images even if the lighting in each scene is constrained to be a known point light source at infinity [42]. In this work, we show that images are much less ambiguous for the class of bilaterally symmetric Lambertian objects. In fact, the set of such objects can be partitioned into equivalence classes such that it is always possible to distinguish between two objects belonging to different equivalence classes using just one image per object. The conditions required for two objects to belong to the same equivalence class are very restrictive, thereby leading to the conclusion that images of symmetric objects are hardly ambiguous. The observation leads to an illumination-invariant matching algorithm to compare images of bilaterally symmetric Lambertian objects. Experiments on real data are performed to show the implications of the theoretical result even when the symmetry and Lambertian assumptions are not strictly satisfied.

3.1 Introduction

The problem of matching images of an arbitrary scene/object under different illumination conditions has been quite elusive. Lack of information about the

geometry and reflectance map makes this problem in its generality, ill-posed. In fact, Jacobs *et al.* [42] show that this problem cannot be solved even under hard constraints of Lambertian reflectance and known single point light sources placed at infinity.

Quite often in vision problems, the intractability of the problem can be reduced significantly by restricting the domain of the problem and using appropriate constraints. In this chapter, we analyze the problem of matching symmetric objects across illumination variations. In particular, we show that unlike general objects, it is almost always possible to distinguish between two bilaterally symmetric objects using just one image per object.

The symmetry assumption eliminates the unknown albedo in the SFS formulation, thereby allowing us to deal with arbitrarily varying albedo maps. Moreover, symmetry leads to a linear constraint on the values of the unknown surface gradients for each point of the object. Though the constraint makes the SFS problem more tractable, it is still not sufficient to recover the surface gradients for general unknown albedo maps.

Unlike the existing work on symmetric SFS, our goal here is illumination-invariant matching rather than shape recovery. We use the linear constraint provided by symmetric SFS to prove the well-posedness of the matching problem for the class of bilaterally symmetric objects. Given two linear constraints from two different images, we solve for the surface gradients. The correctness of the gradients can be checked by substituting them back in the original image irradiance equations for the images and independently computing albedos from the two images. We show

that the two albedo estimates are identical if the corresponding pixels represent the same physical reality (same shape and albedo). If the points differ physically, the computed albedos almost always differ. We derive the rare condition under which they are same. In fact, the condition partitions the set of symmetric Lambertian objects into equivalence classes such that it is always possible to distinguish between two different objects belonging to different equivalence classes based on just one image per object.

The theoretical analysis leads to an algorithm that can be used to match images of real objects where the symmetry and Lambertian assumptions are not strictly satisfied. Given an image, an illumination-invariant representation is derived that can be used for matching. If the assumptions are strictly satisfied, the algorithm is provably correct (up to the described ambiguity). Experimental results show the usefulness of the approach on real images.

3.2 Organization of the chapter

The rest of the chapter is organized as follows. Section 3.3 discusses the related work. The SFS formulation utilizing the 3D bilateral symmetry is described in Section 3.4. The theoretical analysis to prove that the images of symmetric objects are hardly ambiguous is outlined in Section 3.5. In Section 3.6, we propose an algorithm to perform illumination-invariant matching of such objects. Experiments performed to evaluate the performance of the matching algorithm are described in Section 3.7. The chapter concludes with a summary and discussion in Section 3.8.

3.3 Related work

There has been a lot of work on the problem of illumination-invariant matching and recognition. Brooks *et al.* [16] discuss the existence and uniqueness of shapes consistent with a given intensity pattern. In [40], a given image is filtered to suppress the lighting effects in order to recover the object reflectance. A method to recover intrinsic properties of an object using multiple images is proposed in [76]. Jacobs *et al.* [42] describe a matching algorithm based on the observation that the ratio of two images of the same object is simpler than that of two different objects. Chen *et al.* [19] utilize the insensitivity of the direction of image gradients to changes in illumination direction in a probabilistic framework to recognize faces across illumination.

Other than these generic methods, a lot of research has been directed towards recognizing faces across illumination variations. Quite often face-specific methods physically model the image formation process which involves illumination sources, albedo and shape. Class specific properties of faces have been utilized to perform reliable reconstruction or recognition in spite of the ill-posed nature of the problem. [13][99][94][33][77][6] are a few remarkable works in this direction.

Yuille *et al.* [94] use singular value decomposition (SVD) to learn generative models of objects from a set of images taken under different unknown illuminations. Shashua and Raviv [77] perform recognition across varying illumination under an ideal-class assumption. All objects belonging to the ideal class are assumed to have the same shape. [33] uses illumination cone models for illumination-invariant face

recognition. They require a small number of training images of each face under different illuminations to recover the shape and albedo of the face. Basri and Jacobs [6] propose methods for recovering surface normals in a scene. Result in [7] and [68] forms the basis of their work, which proves that the set of all Lambertian reflectance maps obtained with arbitrary distant illumination sources approximately lie in a 9D linear subspace. In [13], Blanz and Vetter perform face recognition across pose and illumination by fitting a 3D morphable model to the images. Zhou *et al.* [99] generalize the traditional photometric approach to handle all appearances of all objects in a class. They impose a rank constraint on shape and albedo in a class to separate the two from illumination.

Though SFS approaches for the recovery of shape and albedo have been studied for a long time, it is only recently that attempts have been made to use them for real matching problems. Due to the ill-posed nature of the problem, the SFS research typically makes uniform albedo assumption which often limits the applicability of the approaches. In a recent work [96][97], Zhao and Chellappa present an SFS approach to recover both shape and albedo for a symmetric object from a single image under piecewise constant constraint on albedo. In [96], they use the same approach for generating frontally illuminated prototype images to perform face recognition. They use partial gradient information from a generic 3D model to perform this task. Using the same formulation, Dovgard and Basri [28] make use of class-specific constraints by writing the unknown surface gradients as a linear combination of the surface gradients of a set of known 3D face models to recover the shape.

Though our work is partly motivated by Zhao and Chellappa's work [96][97],

we differ in the following aspects

1. We derive precise conditions under which images of two different objects are ambiguous.

2. Our approach for illumination-invariant matching is provably correct for symmetric Lambertian objects.

3. We do not use any class-specific information like generic 3D model as used in [96].

3.4 Symmetric shape from shading

Under the assumptions of orthographic projection and Lambertian reflectance, the perceived intensity of a surface point of an object can be written as

$$I = L\rho \frac{1 - pl - qk}{\sqrt{p^2 + q^2 + 1}\sqrt{l^2 + k^2 + 1}} \tag{3.1}$$

where ρ is the surface albedo, $\frac{(p,q,1)}{\sqrt{p^2+q^2+1}}$ is the surface normal, L is the intensity of the light source and $\frac{(l,k,1)}{\sqrt{l^2+k^2+1}}$ is the illuminant direction. As done normally in SFS formulations, we assume that the image intensity I is normalized by the known light source intensity to eliminate L from the expression.

The albedo ρ_- and surface normals $\{p_-, q_-\}$ of the bilaterally symmetric point are characterized as follows

$$\rho_- = \rho \qquad \{p_-, q_-\} = \{-p, q\} \tag{3.2}$$

Therefore, its intensity I_- can be written in terms of the albedo and surface normals

of its symmetric counterpart as follows

$$I_- = \rho \frac{1 + pl - qk}{\sqrt{p^2 + q^2 + 1}\sqrt{l^2 + k^2 + 1}} \tag{3.3}$$

Using (3.1) and (3.3), the albedo can be eliminated leading to the following linear constraint on the surface gradients

$$\frac{I_-}{I} = \frac{1 + pl - qk}{1 - pl - qk} \tag{3.4}$$

$$(I_- - I) - (I_- + I)pl - (I_- - I)qk = 0 \tag{3.5}$$

$$Slp + Dkq = D \tag{3.6}$$

where $S = I_- + I$ is the sum of the intensities of the symmetric points and $D = I_- - I$ is the difference of the two. The linear relation implies that the set of possible surface gradients $\{p, q\}$ lie on a straight line in the pq-space, parameterized by the perceived intensity and the lighting condition. Note that the regular reflectance map provides a quadratic constraint on the values surface gradients can take, given the pixel intensity, albedo and illumination conditions. Figure 3.1 shows the regular quadratic reflectance map and the corresponding linear constraints (3.6). Even if the albedo is known, there are two possible solutions for the unknown surface gradients. Though enforcing integrability [32] helps in removing the ambiguity completely for constant and piece-wise constant albedo maps, the problem is still ill-posed for the more general case of unknown arbitrary albedo map [97]. However, the formulation is quite useful for illumination-invariant matching as discussed in the following sections.

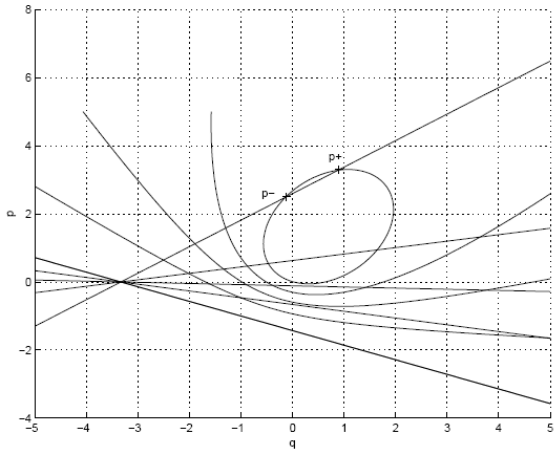

Figure 3.1: Regular and symmetric reflectance maps [97].

3.5 Role of symmetry in illumination-invariant matching

In this section, we use the symmetric SFS formulation to analyze the problem of illumination-invariant matching for the class of bilaterally symmetric objects. Given an image of a bilaterally symmetric object, each pair of symmetric points results in a linear constraint of the form (3.6). Given a second image of the same surface, we obtain another linear relation for each corresponding point pair which leads to the following Lemma.

Lemma 3.5.1 *The linear relations for a point with surface gradients* $\{p_0, q_0\}$*, derived from images taken under different light sources, are concurrent with* $\{p_0, q_0\}$ *as the point of concurrence.*

Proof Line $Slp + Dkq = D$ in the pq-space has to pass through the point $\{p_0, q_0\}$. This is true for all such lines derived from all possible images of the point under various illumination conditions. As two lines can intersect at only one point, the

lines are concurrent with $\{p_0, q_0\}$ as the point of concurrence, which proves the lemma.

Therefore, if two images come from the same object, the corresponding lines intersect at their true surface gradient. Interestingly, even if the two points are not physically same (i.e., they have different surface gradients), the two lines still intersect in the pq-space unless they are parallel. As the points have different surface gradients, the point of intersection can not be the true surface gradient for both of them. These observations help us prove that it is possible to distinguish between two symmetric Lambertian objects using just one image per object as described in the following subsection.

3.5.1 The ambiguity in matching

In a matching scenario, the goal is to determine if the two images come from the same physical object or not. Given two images taken under different illumination conditions, we get an intersection point in the pq-space for each corresponding symmetric point pair, which is a possible solution for the unknown surface gradients.

For each pair of corresponding points from the two image, we get two linear constraints as follows

$$S_1 l_1 p + D_1 k_1 q = D_1 \tag{3.7}$$

$$S_2 l_2 p + D_2 k_2 q = D_2 \tag{3.8}$$

where the subscripts 1 and 2 distinguish the quantities corresponding to the two images. Unless they are parallel, the two lines intersect at a point (say $\{\bar{p}, \bar{q}\}$) in the

pq-space. Substituting the intersection point back in the image irradiance equations (3.1) for the two images, following two albedo estimates are obtained

$$\hat{\rho_1} = \frac{\sqrt{\bar{p}^2 + \bar{q}^2 + 1}\sqrt{l_1^2 + k_1^2 + 1}}{1 - \bar{p}l_1 - \bar{q}k_1} I_1 \qquad (3.9)$$
$$\hat{\rho_2} = \frac{\sqrt{\bar{p}^2 + \bar{q}^2 + 1}\sqrt{l_2^2 + k_2^2 + 1}}{1 - \bar{p}l_2 - \bar{q}k_2} I_2$$

From Lemma 3.5.1, if the two points have same surface gradients and albedo, then the two lines intersect at their true surface gradient. Substituting the true surface gradient back in the irradiance equation will always produce the same true albedo. Though not intuitive, it is possible to get $\hat{\rho_1} = \hat{\rho_2}$ even when the two points are physically different (i.e., they differ either in surface gradients or albedo). The condition on the two points for this to happen is derived in the following theorem.

Theorem 3.5.2 *The two albedos $\hat{\rho_1}$ and $\hat{\rho_2}$ are same if the following condition is satisfied*

$$\frac{\rho_1}{\rho_2} = \frac{p_2\sqrt{1 + p_1^2 + q_1^2}}{p_1\sqrt{1 + p_2^2 + q_2^2}} \qquad (3.10)$$

where ρ_1 and ρ_2 are the true albedos for the two points and $\frac{(p_1,q_1,1)}{\sqrt{1+p_1^2+q_1^2}}$ and $\frac{(p_2,q_2,1)}{\sqrt{1+p_2^2+q_2^2}}$ are the corresponding true surface normals.

Proof Suppose $\frac{(l_1,k_1,1)}{\sqrt{l_1^2+k_1^2+1}}$ and $\frac{(l_2,k_2,1)}{\sqrt{l_2^2+k_2^2+1}}$ are the illuminant directions for image 1 and 2 respectively. For image 1, the true surface gradients $\{p_1, q_1\}$ satisfy (3.7), i.e.,

$$S_1 l_1 p_1 + D_1 k_1 q_1 = D_1 \qquad (3.11)$$

Using (3.11) and (3.7), we get

$$q = \frac{1}{k_1} - \frac{1 - k_1 q_1}{k_1 p_1} p \qquad (3.12)$$

Similarly, for image 2, we have

$$q = \frac{1}{k_2} - \frac{1 - k_2 q_2}{k_2 p_2} p \tag{3.13}$$

These lines intersect at the following point $\{\bar{p}, \bar{q}\}$ in the pq-space

$$\bar{p} = \frac{p_1 p_2 (k_1 - k_2)}{p_1 k_1 (1 - k_2 q_2) - p_2 k_2 (1 - k_1 q_1)} \tag{3.14}$$

$$\bar{q} = \frac{p_1 (1 - k_2 q_2) - p_2 (1 - k_1 q_1)}{p_1 k_1 (1 - k_2 q_2) - p_2 k_2 (1 - k_1 q_1)} \tag{3.15}$$

Now the two albedos obtained by substituting $\{\bar{p}, \bar{q}\}$ back in the image irradiance equations for the two points are same if

$$\begin{aligned} &\frac{\sqrt{\bar{p}^2 + \bar{q}^2 + 1}\sqrt{l_1^2 + k_1^2 + 1}}{1 - \bar{p} l_1 - \bar{q} k_1} I_1 \\ &= \frac{\sqrt{\bar{p}^2 + \bar{q}^2 + 1}\sqrt{l_2^2 + k_2^2 + 1}}{1 - \bar{p} l_2 - \bar{q} k_2} I_2 \end{aligned} \tag{3.16}$$

i.e.,

$$\frac{1 - \bar{p} l_1 - \bar{q} k_1}{1 - \bar{p} l_2 - \bar{q} k_2} . \frac{\sqrt{l_2^2 + k_2^2 + 1}}{\sqrt{l_1^2 + k_1^2 + 1}} = \frac{I_1}{I_2} \tag{3.17}$$

Substituting $\bar{p}$ and $\bar{q}$ from (3.14) and (3.15), the left hand side of (3.17) simplifies to

$$\frac{p_2}{p_1} . \frac{1 - l_1 p_1 - q_1 k_1}{1 - l_2 p_2 - q_2 k_2} . \frac{\sqrt{l_2^2 + k_2^2 + 1}}{\sqrt{l_1^2 + k_1^2 + 1}} \tag{3.18}$$

Also, the right hand side of (3.17) can be written in terms of the true surface gradients and albedos as follows

$$\frac{\rho_1}{\rho_2} . \frac{1 - l_1 p_1 - q_1 k_1}{1 - l_2 p_2 - q_2 k_2} . \frac{\sqrt{l_2^2 + k_2^2 + 1}\sqrt{p_2^2 + q_2^2 + 1}}{\sqrt{l_1^2 + k_1^2 + 1}\sqrt{p_1^2 + q_1^2 + 1}} \tag{3.19}$$

From (3.18) and (3.19), the condition in (3.17) is true if

$$\frac{p_2}{p_1} = \frac{\rho_1}{\rho_2} . \frac{\sqrt{p_2^2 + q_2^2 + 1}}{\sqrt{p_1^2 + q_1^2 + 1}} \tag{3.20}$$

which proves the theorem.

Theorem 3.5.2 leads to a few interesting observations which are described in the following corollaries.

Corollary 3.5.3 *The condition in Theorem 3.5.2 is trivially satisfied if the two points have the same surface gradients and albedo.*

Corollary 3.5.4 *The condition in Theorem 3.5.2 can be true for points even if they differ either in surface gradients or albedo. This essentially means that the point characterized by surface gradients* $\{\bar{p}, \bar{q}\}$ *and albedo* $\hat{\rho}_1 = \hat{\rho}_2$ *can account for both the images, i.e., it is not possible to distinguish between the two points using just one image (of each point) even under hard constraints of bilateral symmetry, Lambertian reflectance and known distant point light sources.*

Corollary 3.5.4 establishes the ambiguity on a per-point basis. If this is true for all visible points of the two objects, then the two objects are indistinguishable given just one image per object taken under different illumination conditions. As chances of such a condition being satisfied by all the corresponding points of two objects are low, it can be concluded that symmetry helps in disambiguating images across illumination. Note that the condition is on the surface gradients and albedo maps of the objects and not on their particular images.

3.5.2 Equivalence classes of bilaterally symmetric objects

We consider the condition in Theorem 3.5.2 as a relation $R(i, j)$ relating two objects i and j (assuming the condition is satisfied for all corresponding point pairs).

Hence, $R(1,2)$ means that the condition is satisfied for all corresponding points of objects 1 and 2. It is interesting to see that relation R is

1. reflexive, i.e., $R(i,i)$ holds,

2. symmetric, i.e., $R(i,j)$ implies $R(j,i)$, and

3. transitive, i.e., $R(i,j)$ and $R(j,k)$ implies $R(i,k)$.

Therefore, the condition in Theorem 3.5.2 induces an equivalence relation on the set of all possible bilaterally symmetric objects. In other words, such a set can partitioned into equivalence classes such that any two objects belonging to the same equivalence class cannot be distinguished using just one image per object. This follows directly from Corollary 3.5.4. On the other hand, two objects belonging to two different equivalence classes do not satisfy the condition in Theorem 3.5.2 and thus can always be distinguished using just one image per object.

3.6 Illumination-invariant Matching

If the assumptions of Lambertian reflectance and bilateral symmetry are reasonably adhered to, the formulation in Section 3.5.1 can directly be used to reliably match images across illumination. As the chance of getting images of two different objects that belong to the same equivalence class is very low, the algorithm should not make any error in matching.

Unfortunately, in most practical applications, the objects are neither Lambertian nor perfectly symmetric. From Section 3.5.1, two images are recognized as

belonging to the same physical object, if the two estimated albedos $\hat{\rho}_1$ and $\hat{\rho}_2$ are same. $\hat{\rho}_1$ and $\hat{\rho}_2$ depend non-linearly on the estimated surface gradients $\{\bar{p}, \bar{q}\}$. Estimation of surface gradients $\{\bar{p}, \bar{q}\}$ in turn depends on how strictly the assumptions are adhered to. Deviations from the assumptions make the estimation of surface gradients $\{\bar{p}, \bar{q}\}$ and hence $\hat{\rho}_1$ and $\hat{\rho}_2$ quite unstable. The instability in the estimation makes the scheme unsuitable for real data.

Here, we propose a novel algorithm to match images of symmetric objects across illumination which follows naturally from Theorem 3.5.2. The algorithm does not involve estimation of $\{\bar{p}, \bar{q}\}$ or $\hat{\rho}_1$ and $\hat{\rho}_2$, and thus degrades quite gracefully when the assumptions are not strictly satisfied.

From Theorem 3.5.2 and Corollaries 3.5.3 and 3.5.4, two objects appear similar (given one image per object) iff

$$\frac{\rho_1}{\rho_2} = \frac{p_2\sqrt{1+p_1^2+q_1^2}}{p_1\sqrt{1+p_2^2+q_2^2}} \tag{3.21}$$

That is, iff

$$p_1\frac{\rho_1}{\sqrt{1+p_1^2+q_1^2}} = p_2\frac{\rho_2}{\sqrt{1+p_2^2+q_2^2}} \tag{3.22}$$

From the given images, we have the following image irradiance relation for each point on the object

$$I = \rho\frac{1-pl-qk}{\sqrt{p^2+q^2+1}\sqrt{l^2+k^2+1}} \tag{3.23}$$

Substituting for ρ_1 and ρ_2 from the image irradiance equations for the two objects in (3.22)

$$I_1\frac{\sqrt{1+l_1^2+k_1^2}}{1-p_1l_1-q_1k_1}p_1 = I_2\frac{\sqrt{1+l_2^2+k_2^2}}{1-p_2l_2-q_2k_2}p_2 \tag{3.24}$$

For each image, symmetry provides a linear constraint of the form (3.7) which has to be satisfied by the true surface gradients $\{p_1, q_1\}$, i.e.,

$$S_1 l_1 p_1 + D_1 k_1 q_1 = D_1 \tag{3.25}$$

For pixels with $D_1 \neq 0$,

$$\frac{S_1}{D_1} l_1 p_1 + k_1 q_1 = 1 \tag{3.26}$$

From (3.26) and (3.24), the condition for the corresponding points of the two objects to appear similar becomes

$$I_1 \frac{\sqrt{1 + l_1^2 + k_1^2}}{l_1(\frac{S_1}{D_1} - 1)} = I_2 \frac{\sqrt{1 + l_2^2 + k_2^2}}{l_2(\frac{S_2}{D_2} - 1)} \tag{3.27}$$

Interestingly, the condition in (3.27) involves only light source directions and image intensities. Thus, given two images, one can use this simple condition for each corresponding pixel to decide whether they come from the same object or not. If the symmetry and Lambertian assumptions are strictly adhered to, the matching decision is provably correct up to the ambiguity in Corollary 3.5.4. As the condition in (3.27) does not involve any unstable estimation of surface gradients or albedo, the algorithm degrades gracefully with deviations from the assumptions.

The two sides of the condition in (3.27) can be treated separately as the illumination-invariant representation of the respective objects as follows

$$I_{1r} = I_1 \frac{\sqrt{1 + l_1^2 + k_1^2}}{l_1(\frac{S_1}{D_1} - 1)} \tag{3.28}$$

$$I_{2r} = I_2 \frac{\sqrt{1 + l_2^2 + k_2^2}}{l_2(\frac{S_2}{D_2} - 1)} \tag{3.29}$$

Two images can be easily compared by generating these *virtually* relighted images.

3.7 Experiments

The main contribution of this work is the theoretical statement that unlike general objects, it is possible to distinguish between bilaterally symmetric Lambertian objects using just one image. In this section, we present the results of experiments performed on simulated and real data to evaluate the practical implications of the work.

3.7.1 Experiments on simulated data

First, we use simulated data to verify the correctness of the proposed theoretical result. We use the 3D face models used by Blanz and Vetter in their morphable model [13]. We generate several images of 100 subjects in the database under randomly selected illumination conditions. Here, the faces are made bilaterally symmetric and the images are generated using Lambertian reflectance. As the assumptions made in the theoretical formulation are strictly adhered to, the matching algorithm does not make any error.

3.7.2 Experiments on real data

We also test the performance of the algorithm on PIE dataset [79]. The PIE dataset has 68 subjects with images of each subject in 21 different illumination conditions. The images show deviations from Lambertian and symmetry assumptions. Moreover, the light source direction needs to be estimated which involves some error. Figure 3.2 shows the *virtually* relighted images obtained from different images

of a subject in the dataset. The light source direction in an image is estimated using a simple algorithm recently proposed by Lee and Moghaddam [51]. The relighted images look like flattened frontally illuminated images. As desired, the illumination effects in the original images mostly disappear in the relighted images.

Though the relighted images are not perfect (as the assumptions are not strictly satisfied), they seem promising to be used for matching images across illumination variations. We perform a face recognition experiment using the PIE dataset. A set of commonly used challenging illumination conditions from the PIE dataset is chosen to test our simple relighting based scheme (see Figure 3.3). In this setting, all images in one illumination scenario are used to form the gallery and another one to form the probe set. Thus, both the gallery and the probe set have one image per subject. The recognition experiment is repeated for all combinations of gallery and probe sets. Similarity between a gallery and a probe image is measured using a simple cross correlation between the corresponding relighted images as follows. Suppose f_g and f_p are two vectorized relighted images, then the similarity of the images is given by

$$S(g,p) = \frac{< f_g, f_p >}{|f_g||f_p|} \tag{3.30}$$

where $< f_g, f_p >$ denotes the scalar product of the two vectors. This is a very simple measure and fits well with the goal of stress testing the practical usefulness of the theoretical results. Table 1 shows the recognition results obtained in the experiment. The proposed approach using the relighted images works quite well even with such a simple distance measure. Unlike most face recognition methods, we do not

make use of any face-based statistics (like Eigenfaces, 3D morphable models, etc.). Recognition performance using the intensity images directly is also shown for comparison. Intensity images are normalized before computing the similarity. For most gallery-probe scenarios, relighted images perform better than the normalized intensity images. The improvement is quite significant when the illumination conditions for the gallery and probe scenarios are very different.

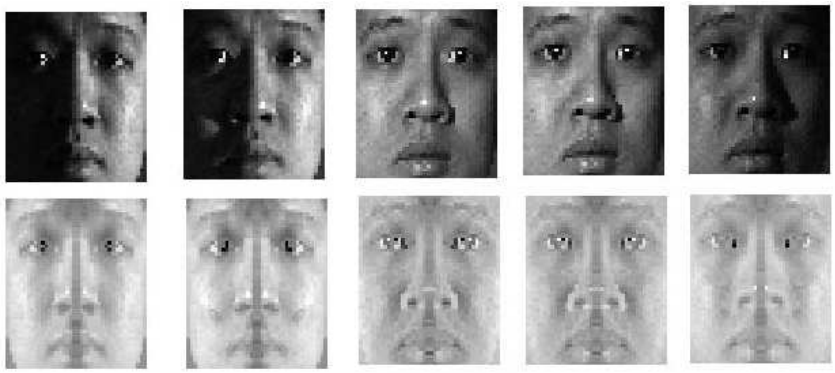

Figure 3.2: Virtually relighted image examples using images from the PIE dataset.

3.8 Summary and discussion

We showed that two bilaterally symmetric objects can almost always be distinguished using just one image per object taken under different illumination conditions. The condition under which they cannot be distinguished, partitions the set of symmetric Lambertian objects into equivalence classes. In practice, it is difficult for two objects to satisfy the condition leading to the conclusion that bilaterally symmetric objects are hardly ambiguous.

Based on the theoretical formulation, we proposed a virtual relighting algo-

Probe	f_{09}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{21}	f_{22}
Gallery									
f_{09}	-/-	99/99	97/97	97/94	75/63	60/44	56/34	99/99	85/84
f_{12}	99/99	-/-	99/99	100/99	81/74	62/46	59/31	100/100	85/96
f_{13}	99/94	100/97	-/-	100/100	100/100	94/78	81/54	100/100	100/100
f_{14}	99/91	100/97	100/100	-/-	99/100	94/79	82/59	100/100	100/100
f_{15}	94/35	100/49	100/100	100/100	-/-	99/100	99/96	100/68	100/100
f_{16}	97/38	100/49	100/94	100/96	100/100	-/-	100/100	100/65	100/99
f_{17}	85/37	91/44	97/63	99/71	100/100	100/100	-	94/50	100/90
f_{21}	99/99	100/100	100/100	100/100	87/79	76/51	69/44	-/-	100/97
f_{22}	97/54	100/81	100/100	100/100	100/100	97/96	97/72	100/96	-/-

Table 3.1: Recognition results on the PIE dataset. f_i denotes images taken with i^{th} flash ON as labeled in the PIE dataset. Each $(i,j)^{th}$ entry in the table shows the recognition rate obtained with the images from f_i as gallery and from f_j as probes. The first number is the rank-1 recognition performance using the relighted images while the second number is the performance using the intensity images directly.

rithm to recognize real objects that do not strictly satisfy the assumptions made. The algorithm is provably correct for symmetric Lambertian objects up to the ambiguity described in Theorem 3.5.2. The relighted images obtained on real images seem to be free of any illumination effects. Face recognition experiments using the relighted images showed excellent performance without using any sophisticated classifier or class-based statistics.

There exist a few specific cases where symmetric SFS analysis may not be

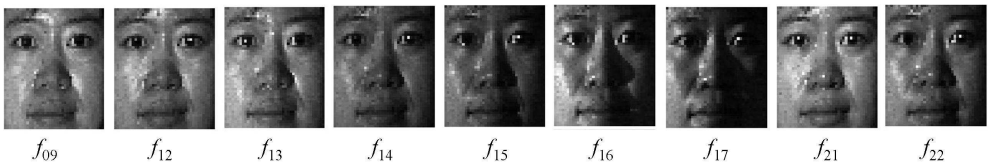

Figure 3.3: Illumination conditions from the PIE dataset used in the face recognition experiment.

effective. Shadow pixels do not reveal much information about the surface gradients and have to be excluded from the formulation. Moreover, if $l = 0$ or $p = 0$, two symmetric points have same image intensity, thereby providing no additional information due to symmetry.

Chapter 4

Cohort Analysis for Biometric Matching

Most biometric matching algorithms make decisions based solely on a score that represents the similarity between the query biometric and enrolled biometric of the claimed identity. Though there have been attempts to perform score-level fusion, the emphasis has been on multi-classifier and multi-sample fusion. The commonly adopted fusion techniques, however, rarely make use of the large number of non-match biometric samples present in the enrollment database. In this chapter, we describe algorithms that make use of these often ignored non-match biometric samples to improve biometric verification and identification performance. For each enrolled subject, a cohort (set of similar biometric samples) is identified from the available non-match samples based on a simple match score-based criterion. The final (consolidated) similarity score of a query biometric is estimated using its similarity not only with the claimed identity but also with the cohort of this identity. The similarity scores are fused using two different approaches: a likelihood ratio based normalization scheme and a Support Vector Machine based classifier. Experiments on face and fingerprint biometrics using multiple match algorithms show the performance of the approaches. The FVC 2002 data set and a private IBM data set are used for fingerprint experiments; while the PIE illumination data set is used for face recognition experiments. The results show that the cohort-based algorithms

significantly improve the verification and identification performance at the expense of estimating a few extra matches. Incidentally, any existing biometric matcher can be used.

4.1 Introduction

By definition, verification (or authentication) is the confirmation of a claimed truth. In the context of biometrics, verification amounts to validating if a given query biometric is similar to the enrolled biometric of the claimed identity. On the face of it, this appears to be a straight-forward binary classification problem based on a similarity score s between the query biometric and that of the claimed identity. The claim of the query is validated if the similarity is greater than a pre-set threshold T. Figure 4.1 shows a typical verification system. To perform well, such a system expects the classes to be reasonably compact with respect to the inter-class distances, i.e., samples from the same class are much closer to each other than the ones from different classes.

In an identification task, given a query biometric, the goal is to return the most similar biometric from the enrolled database. Unlike the verification task, the query needs to be compared against all enrolled biometrics to rank-order them and return the most similar ones (Figure 4.2). The decision is made based on the similarity of the query biometric with the enrolled biometrics. As in the verification task, the similarity score of the query biometric with an enrolled biometric is often purely a measure of how close the query is from the enrolled samples (usually just one) of

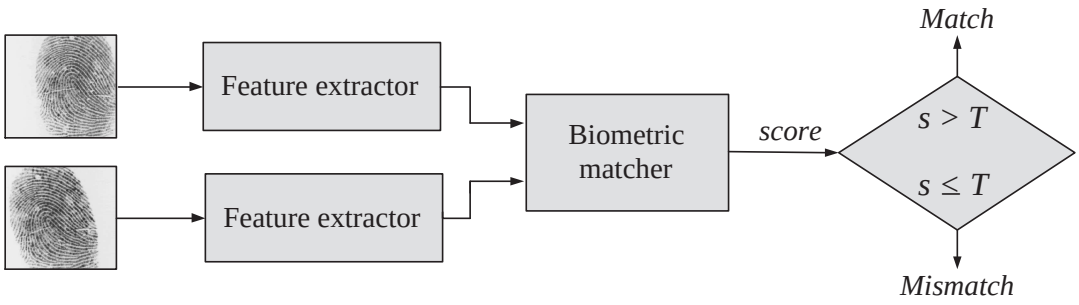

Figure 4.1: A typical verification system. A matcher determines the similarity score s between two biometrics. The decision is made by comparing the similarity score with a suitable pre-set threshold T.

the identity in terms of similarity.

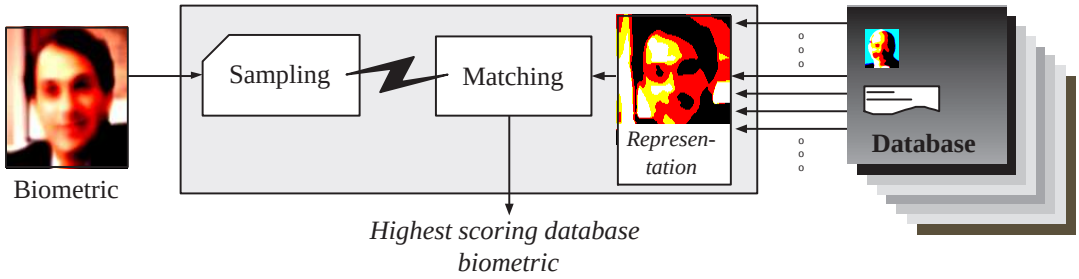

Figure 4.2: A typical identification system. A matcher determines the similarity of the given query with the enrolled identities to rank-order them and return the most similar ones.

4.1.1 Issues in using raw similarity scores

A verification system based on a fixed threshold on the raw similarity scores works well when the classes have similar distributions and the intra-class expanse (spread) is smaller than the inter-class distances. When classes are non-identically

distributed in the feature space, the threshold may turn out to be too stringent for a few classes while too lenient for others. This results in many false accepts and false rejects, adversely affecting the overall verification performance of the system. Such situations are more realistic than the ones where classes are identically distributed and spread apart. In theory, one could set class-specific thresholds to avoid such errors. This requires the system to have *a priori* knowledge about the distribution of each class in the feature space. More often than not, very few (usually just one) samples per identity (class) are available in the database which provide hardly any information about the class distributions.

Figure 4.3 illustrates a distribution of classes in a feature space. Typically biometric classes will not have similar distributions. The classes may not even be isotropically distributed about their centers. Moreover, as the number of classes increases, they tend to overlap leaving no choice of threshold that achieves error-free verification. In addition, when dealing with realistic noisy images, the similarity score of even a genuine query may be lower than normal, making it difficult to set a single threshold to validate both noisy and noiseless images as shown in Figure 4.4. Quality assessment as a pre-processing step could help choose a different threshold depending on the amount of noise in the query input, this may not always be possible. Most verification systems have to deal with such situations, irrespective of the features they use.

In an identification system, the raw similarity scores of the query biometric with the available samples of the enrolled biometrics are directly used to rank-order the enrolled identities and to identify the query biometric. However, if the class

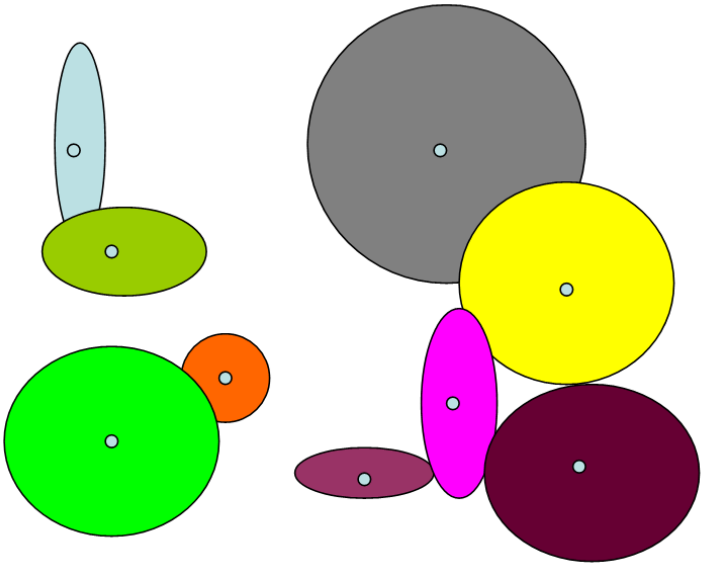

Figure 4.3: The figure illustrates a typical feature space containing overlapping classes with different distributions.

distributions differ greatly, direct ranking of the enrolled identities based on the raw similarity of their available samples with the query biometric may not be optimal. Though there is no threshold used here as in the verification case, the rank-ordering does depend on how the similarities are computed. Raw similarity scores computed using the few available (usually just one) enrolled biometric samples for each identity do not take into account the variations in class distributions because of the lack of training samples.

4.1.2 Class neighborhoods in feature space

In most realistic scenarios, there is a large number of enrolled (and training) biometrics other than the biometric of the claimed identity. The large number of

non-match biometrics (the biometrics that do not belong to the claimed identity) are almost always ignored by the traditional verification systems while validating the claim of a given query biometric. Though reasonable when treating each enrolled subject in isolation, such strategies do not utilize the possibly useful class distribution information present in the form of the large number of biometrics present in the database. The neighboring, nearby biometric classes can provide information about the class distributions of biometric representations in the neighborhood of the claimed biometric. The information can potentially be useful to determine class-specific thresholds that has proved elusive, especially with just one example biometric per class. For example, the similarity scores of the query biometric with the neighbors of the claimed biometric can help in reducing false accepts and rejects. The neighborhood information even has the potential to provide resilience to noise in validating a genuine query to a certain extent (Figure 4.4).

This motivates us to look beyond the raw similarity score between the query biometric and just the claimed biometric. We show how one can effectively utilize the similarity of the query biometric with the preselected *neighbors* of the claimed biometric in validating the identity claim. From a pure machine learning point of view, inclusion of multiple similarity scores increases the dimensionality of the features used to perform verification. Note that the traditional verification schemes use a one-dimensional feature in the form of raw similarity score between the query and the claimed identity. Suitably selected higher dimensional features may result in greater class separation (genuine vs. impostor) leading to increased verification performance.

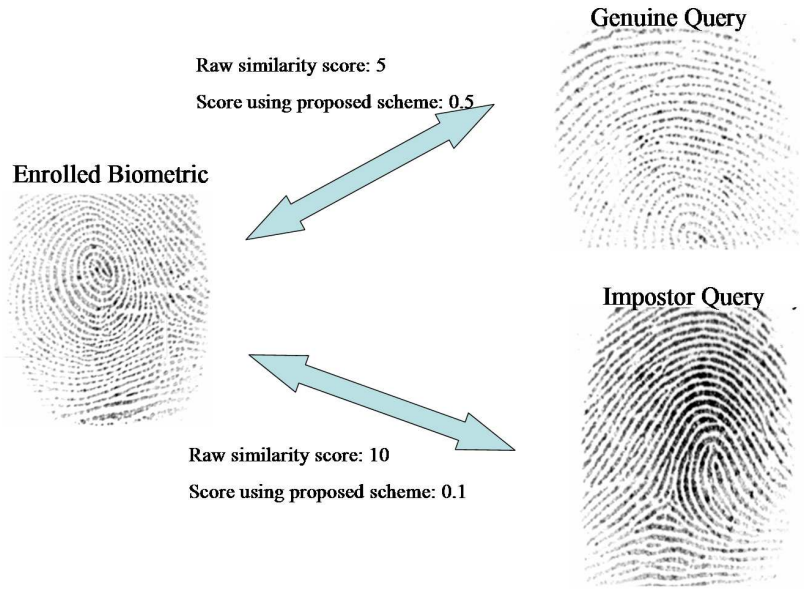

Figure 4.4: An illustration to highlight the effectiveness of the proposed cohort-based normalization scheme. As the raw similarity score of the genuine biometric is lower than that of the impostor biometric, the traditional raw similarity score-based threshold strategy is bound to make an error. In contrast, the proposed approach increases the score of the genuine query.

For identification scenarios, though each query is compared to all enrolled biometrics, the variations in the class distributions are often not taken into account. It is not possible to learn class distributions with the limited number of samples available per biometric. As with the verification task, neighborhood information of a biometric can be useful when the classes are differently distributed. From a machine learning point of view, addition of neighborhood scores to characterize the similarity of a query with an enrolled biometric increases the chance of better separation of biometric classes. We propose to use such information to normalize the variations in class distribution and to improve identification performance. This is achieved by comparing the query against the cohort (neighbors) of the claimed identity and using those similarity scores to account for differences in class distributions.

The unified fusion method makes use of mated and non-mated samples to improve performance. Figure 4.5 illustrates the advantage of using neighborhood information as proposed in this chapter. The correct match is the top match using the proposed approach that accounts for variations in class distributions. In contrast, directly using the raw similarity scores, the correct match is not present in the top five matches. In this example, variations are due to differences in illumination conditions under which the images are captured.

These intuitions are the basis for the cohort-based matching schemes described in this chapter. We use cohort as the term to denote a compact set of neighboring biometric samples. For each enrolled class, its cohort is selected to get an idea about how crowded the feature space is around that class. The proposed algorithms utilize the similarity of the query biometric with the cohort (of the enrolled biometric) to

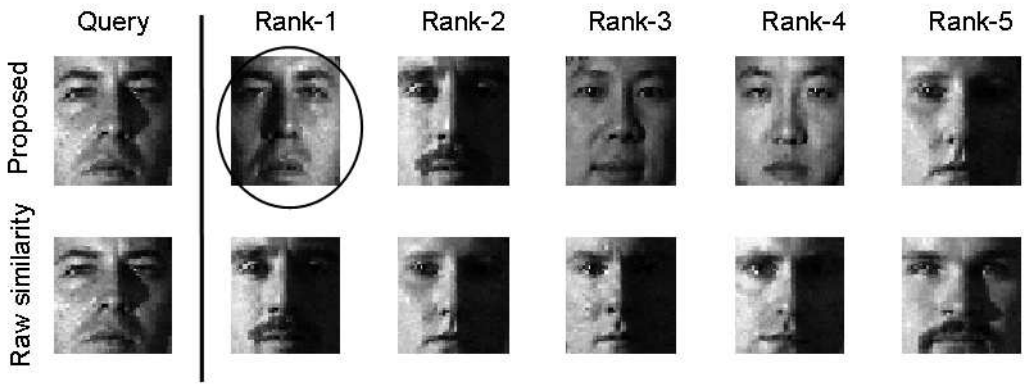

Figure 4.5: The top row shows the top five matches obtained using the proposed cohort-based approach; the bottom row shows the top five matches obtained using the raw similarity scores. The correct match is encircled in the top row while it is missing from the bottom row.

normalize for the variations in class distributions, thereby improving the verification and identification performance.

The main contributions of this work are as follows–

1. We propose a novel score-based cohort selection strategy to address cohort analysis for matching biometrics. In contrast, most earlier attempts to cohort selection (normally done for speaker verification) depend on some sort of statistical model. To the best of our knowledge, no such attempt has been made for other biometrics (face, fingerprint, etc.) for which such a model is not known (or learning such a model is not easy).

2. A machine learning approach is developed to effectively utilize cohort scores. A Support Vector Machine (SVM) based classifier is used that fuses cohort

scores with the raw similarity score of the query biometric and the enrolled biometric. This returns a final cohort-based similarity measure, which is used to perform verification and identification tasks. Most earlier works view this as a score-normalization problem instead of classifier training.

3. The approach is generalized to scenarios where multiple biometric samples per identity are present in the database. The similarity scores of the query biometric with the biometric of the claimed identity and its cohort are fed into an SVM-based classifier to arrive at the final decision.

4.2 Organization of the chapter

The rest of the chapter is organized as follows. The following section briefly reviews related published work. Section 4.4 provides a theoretical justification for the proposed approaches, which are described in detail in Section 4.5. Section 4.5.1 describes the score-based approach for cohort selection. Two techniques proposed to utilize cohort information to improve biometric matching performance are described in Section 4.5.2. A useful extension of the approach for fusing multiple biometrics per enrolled identity in the cohort framework is also discussed. Results of extensive experiments performed to evaluate the usefulness of the proposed algorithms are in Section 4.6. Section 4.7 concludes with a brief summary and discussion.

4.3 Related work

Much research has been done on the problem of biometric matching [14], but for most popular biometrics, little has been done to look beyond the raw similarity scores to improve the matching performance. Here we group the related works into three categories to summarize the progress made so far to normalize for the variations in class distributions for the task of biometric matching. Most earlier techniques address the problem as one of score normalization using some appropriate score-based statistics. Unlike our work, most of them concentrate mainly on the verification task.

4.3.1 Score normalization for speaker verification

Score normalization for text-independent speaker verification is popular practice [53][5][62][57][61][72][39][21][24]. Li and Porter [53] propose normalization technique that improve speaker recognition accuracy using short uncontrolled speech samples. The normalization depends on the mean and variance of the scores of the query samples with the enrolled biometric. Auckenthaler *et al.* [5] review the world model and use the zero normalization techniques using Bayes' theorem. They propose a novel normalization technique called "test normalization," which shows an improvement over other standard techniques. Normalization involves calculating impostor log-likelihood scores for a test utterance to estimate mean and variance parameters. Bengio and Mariethoz [57] introduce various score normalization techniques applied to text-independent speaker verification systems.

A second approach to score normalization is cohort normalization [39][72] that uses a set of cohort speakers close to an enrolled speaker. The cohort selection is done during training by comparing the speaker model to cohort models. Ariyaeeinia and Sivakumaran [4] propose an approach that finds a cohort set of speakers during operation. Higgins *et al.* [39] propose using cohort speakers to use a likelihood ratio as the basis for verification. The authors suggest that the denominator likelihood is dominated by the density of the nearest reference speakers called cohort. Reynolds [71] shows that a speaker-independent universal background model provides better normalization as compared to a speaker-dependent one for the task of speaker verification. Both the speaker and background are modeled using Gaussian Mixture Models (GMMs). A fruitful extension of such an approach to other biometrics like fingerprints and faces has not been discussed before. Estimation of the background distribution in the absence of a suitable statistical model to represent these biometrics makes the problem hard.

4.3.2 Score normalization for other biometrics

Fierrez-Aguilar *et al.* [31] discuss the advantages of score normalization for signature verification. They conclude that class-dependent thresholds improve the verification performance. This happens because the enrolled and query biometric distributions are not aligned for various enrolled biometrics. Appropriate score normalization not only implicitly assigns different thresholds for different classes but also accounts for other non-biometric variations, leading to better verification

performance. In [85], Tulyakov and Govindaraju combine the top two scores for each query in an identification framework. The authors point out the benefit of using a combination of scores, instead of only the best score. They draw parallels of such a technique with the score normalization for speaker verification and identification. Zorita *et al.* [80] use a global fingerprint population to form a universal cohort set for all enrolled fingerprints. The normalization is done using the maximum score attained by the query biometric against the cohort set. A large population of fingerprints is required to form a useful global cohort set. As the query fingerprint needs to be matched with all the fingerprints in the cohort set, the technique requires considerable computation. In contrast, we propose to select separate small cohort sets for each enrolled biometric. This provides the benefits of using non-match examples in performing the verification task at little extra computational cost.

4.3.3 Other fusion-based approaches for biometric matching

Cohort or neighborhood-based normalization schemes are relatively unexplored for more general class of biometrics that are hard to model statistically. Instead, attempts have been made to fuse multiple biometrics per subject in the enrolled database. Uludag *et al.* [87] propose a similarity score-based approach to select and fuse multiple biometrics for each enrollee to improve the performance of a fingerprint authentication system. Verification is done based on the mean (or minimum) of the similarity scores of the query with the biometrics of the claimed identity. Ryu *et al.* [73] propose to generate a super-fingerprint by incorporating only the highly

reliable minutiae based on multiple fingerprint images. A successive Bayesian estimation approach is applied on a sequence of prints to determine the highly likely minutiae. Online improvement of the enrolled biometrics during the verification process has also been proposed [46]. These methods utilize only positive examples (available biometric samples of the enrolled identity) to improve the enrolled biometrics. In contrast, we investigate the usefulness of incorporating non-match samples for verification and identification tasks.

4.4 A probabilistic perspective

The matcher used in most authentication systems estimates a similarity measure to evaluate the hypothesis that the query x belongs to the class of an enrolled biometric w. In probabilistic terms, this can be written as follows

$$s(x, w) = \psi(p(x|w)) \tag{4.1}$$

Here, $p(x|w)$ denotes the likelihood that x comes from the distribution of w and $\psi(.)$ is an increasing function that maps this likelihood into a similarity score. Matchers differ in the choice of $\psi(.)$ and the way the likelihood is computed but the underlying concept remains the same.

For tasks in which each enrolled biometric w is represented using a statistical model, computing the score $s(x, w)$ in (4.1) is quite straight-forward. Unfortunately, for most biometrics it is not easy to do so. For example, no such statistical representation is known for fingerprint and face biometrics. In such situations, matchers estimate the similarity of the query with the available samples of the enrolled bio-

metric. The hypothesis that the query belongs to the class w is evaluated based on a suitable function of the similarity score (distance) of the query with the available samples of class w. Such a scheme is reasonable if a large number of samples (that represent the overall distribution of the class) per enrolled identity are present in the database. Clearly, this is not the optimal basis to make verification or identification decisions when only a few (or just one) samples per class are available because the similarity of the query with the available sample(s) do not account for the variations in class distributions of the enrolled biometrics.

4.4.1 Similarity ratio

One can treat the problem of estimating the similarity of the query with an enrolled biometric as a basic hypothesis testing problem between the following two hypotheses

- H_1: The query x belongs to the claimed identity w.
- H_0: The query x does not belong to the claimed identity w. In other words, x belongs to the complement (or background) class $\bar{w} = (U - w)$, with U being the universal set of biometrics.

The optimal test to decide between the two hypotheses is the following likelihood ratio test

$$\frac{\psi(p(x|w))}{\psi(p(x|\bar{w}))} \begin{cases} > \theta & \text{accept } H_1 \\ \leq \theta & \text{accept } H_0 \text{ (or reject } H_1) \end{cases}$$

The *similarity ratio*

$$S(x, w) = \frac{s(x, w)}{s(x, \bar{w})} = \frac{\psi(p(x|w))}{\psi(p(x|\bar{w}))} \tag{4.2}$$

takes into account the similarity of the query with the rest of the classes along with the samples of the claimed identity. Such a score has a much better potential as compared to raw similarity $s(x, w)$, to normalize for the variations in class distributions even when only one sample per class is present in the database. A fixed threshold θ on the similarity ratio $S(x, w)$ can be seen as a dynamic threshold in terms of the raw similarity score for verification tasks. For identification tasks, the similarity ratio based approach essentially transforms the raw similarity scores to a space in which the differences in class distributions have been accounted for.

4.4.2 Noise resiliency

Interestingly, the idea of likelihood ratio based measure is useful even to handle noise in queries. This follows from the observation that such a measure normalizes for the prior $p(x)$ on the query. The prior on the query essentially encodes its quality or some sort of bias not accounted for in the raw similarity score. It can be mathematically written as follows

$$p(x) = p(x|w)p(w) + p(x|\bar{w})p(\bar{w}) \tag{4.3}$$

The prior $p(x)$ is much lower for noisy queries than the good quality ones. For example, a random image claiming to be a fingerprint will have a very low value of $p(x)$. Assuming there are lots of classes in the database, we get $p(\bar{w}) \gg p(w)$. Hence, $p(x) \approx p(x|\bar{w})$. Therefore, the similarity ratio based measure implicitly accounts for

the quality of the query, making it desirable for the matching tasks.

Interestingly, one can consider noise to be another aspect of accounting for class distributions during the matching process. Presence of noisy samples, essentially leads to non-isotropic expansion of identity classes in the feature space. Note that noise can be anything that does not contribute to the identity as seen by a matching algorithm. For example, in case of face matching, bad illumination or extreme pose can be considered noise when the matching algorithm is not capable of modeling such variations. Fig. 4.5 shows an example in which the similarity score based measure is able to correct the mistakes made by a face matcher due to extreme illumination.

4.4.3 Background modeling

An issue in using a similarity ratio-based measure is the choice of the background model $\bar{w} = (U - w)$ for each class w. Again, U is the universal set representing the database of biometrics. Given a reasonable model for the background, computing the similarity ratio is quite straightforward. Even if there are good models to represent individual biometrics, the development of a universal background class is a challenge for most biometric modalities. This has led to speaker verification approaches that attempt to normalize the raw similarity scores using a chosen set of impostors (cohort) for each enrolled identity [39]. The complement likelihood in the denominator (4.2) is replaced by a suitable function (e.g., mean or maximum) of the likelihood that the query template belongs to an impostor class. One can potentially choose all the available (enrolled or not) biometrics to be the impostors; however,

that will increase the computational complexity of processing each query when the database is large. Therefore, only those biometrics that have good resemblance with the claimant biometric are chosen as impostors to represent the complement class $\bar{w}$.

The idea that the denominator of the likelihood ratio is dominated by the nearest biometric was suggested by Higgins *et al.* [39] in the context of speaker verification. The goal is to approximate the conditional density of the given query in the neighborhood of the claimed biometric. Even doing this requires a good representative model for each biometric. In this work, we focus on biometrics for which such models are either not known or not easy to learn from just a few samples. For such biometrics, we propose a very simple similarity based method to select a set of impostors (cohort) to represent the background class.

4.5 Proposed approach

We propose to use mated and non-mated biometric samples effectively to improve biometric verification and identification performance. Given a query, its similarity with an enrolled biometric is based not only on the similarity score with the available samples of the enrolled biometric but also the neighborhood set (cohort) of the enrolled biometric. This requires determining the neighbors of each enrolled biometric in the database. Note that neighbors can either be other enrolled biometrics or can belong to a training set representative of the biometric.

Neighboring classes can easily be determined if one can establish a suitable

generative/statistical model given the biometrics of the enrolled entity. For most biometrics (like face, fingerprint, etc.), learning such models, given a single (or a few) biometric samples is not easy. We propose an extremely simple (yet effective) scheme where the neighbors of each biometric are selected based on the raw similarity scores. The approach does not assume anything about the biometric modality or the way a matcher computes similarities. Though the matcher can be used to robustly determine similarity, this is not the focus of this work. In fact, we show that even a simple score-based scheme provides the benefits of cohort analysis.

Given a query, we use the available matcher to determine the similarity scores with the claimed identity and its cohort[1]. The scores are fused using two different algorithms for verification and identification. In the first algorithm, the raw similarity score is normalized using the maximum cohort score. Though simplistic, the method performs quite well across biometrics using several different matchers. In the second method, the feature vector comprising of the raw similarity score and the cohort scores is fed to a standard SVM classifier that validates the genuineness of the claim. For the identification task, the output of the SVM (distance from the separating hyper-plane) is used as the final similarity measure to rank-order the enrolled entities. The proposed cohort-based approach is further extended for situations when multiple templates per identity are enrolled in the database. Superior performance is observed for different biometrics using raw similarity scores from several different matchers in both verification and identification tasks.

[1] Cohort is plural in itself and refers to the entire set of neighbors of an enrolled identity

4.5.1 Cohort selection

Selecting close-by impostors (cohort set) for each enrolled biometric is a crucial step in cohort analysis. In the speaker verification case, individual or universal background speaker sets are often modeled using Hidden Markov Models (HMMs) or a mixture of Gaussians. Given just one (or a few) fingerprint or face samples, learning statistical models is hard. This makes the derivation of a universal background of face background quite difficult to realize in practice. The task of choosing subject-dependent sets of impostor fingerprints/faces seems equally impregnable in the absence of a model. Here we propose a simple similarity score-based strategy for selecting a cohort. As shown later in the experiment section, though simple, such a strategy does very well in improving the overall verification and identification performance.

A cohort is selected for each enrolled biometric from other biometrics available in the database. The biometric samples can either belong to the enrolled set or a separate training set. No assumption about the training set is made, though it is expected to be a good representative of the biometrics at hand. In the absence of such a training set, one can choose cohort from the target (enrolled) set itself as is done in most experiments described in this chapter. Given a biometric matcher, we compute the (raw) similarity of each enrolled target subject with the templates in the training set. The training templates with high similarity scores are chosen to form its cohort. Intuitively, these training samples constitute the neighborhood of the enrolled identity in the class space.

4.5.1.1 Effectiveness of the cohort selection scheme

Biometric descriptors are multi-dimensional. Therefore, a purely score-based approach to choose similar biometric samples that lie in a high-dimensional space may seem hard. It is worthwhile to note that traditionally, these raw similarity scores are used to make the final verification/identification decisions. So these scalar scores do indicate something about the proximity of the biometrics even though the biometrics themselves lie in a high-dimensional feature space. For example, in the context of fingerprint verification, matchers often produce the number of paired minutiae as the similarity score which should be a reasonable measure of the similarity of the fingerprints for cohort selection. There is little doubt that one may be able to choose better impostors by comparing the feature vectors instead of depending purely on the similarity score of a matcher. However such a simple score-based technique helps us in evaluating the effectiveness of incorporating impostor scores in matching tasks without going into the details of the biometric, feature extraction process or matcher. Moreover, this allows us to easily demonstrate the usefulness of our fusion strategy across different matchers and biometrics.

Fig. 4.6 shows a few cohort sets selected using the proposed approach. A set of 68 face images (one random image per identity in the PIE data set [79]) is used as the training set. The matching algorithm in [1] is used to compute the similarity scores. It is quite interesting to see that the cohort images share resemblance with the target face image. In a few cases, it is because of the facial features like mustache or skin color. In others, the resemblance is purely because of the illumination

Figure 4.6: Cohort selection for face images. In each row, columns 2-6 show the automatically chosen cohort set for the face in the first column. The matching algorithm in [1] is used to compute the similarity scores. The selected cohort images seem to share some resemblance with the corresponding claimant images in the form of mustache, illumination conditions, etc.

conditions. This happens because most matchers are unable to completely separate the external factors like illumination effects of the facial features, often resulting in higher similarity scores for two different faces with similar illumination conditions as compared to two face images of the same person under different lighting. When such a cohort set is used to make the verification decision (as described in the following section), it has the potential to normalize for these nuisance external factors and produce similarity scores that are better correlated to the true identity.

Fig. 4.7 shows the selected cohorts for a few fingerprints. A set of 100 fingerprints from the FVC 2002 [56] data set is used as the training set. The Bozorth 3 matcher [89] is used to generate raw similarity scores. In spite of the small size of the training set, the cohort fingerprints seem to share some resemblance with the corresponding target fingerprints (first fingerprint in each row). As the Bozorth matcher is a minutia-based matcher, the members of the cohort sets need not always share perceptual similarity with the target fingerprint.

4.5.1.2 Issues

Given the cohort for each enrolled subject, the following issues need to be addressed to judge its usefulness and effectiveness for the task of biometric matching.

1. Can the cohort information be used to improve the verification and identification performance?

2. Can the background class be modeled effectively using a small cohort for each enrolled biometric?

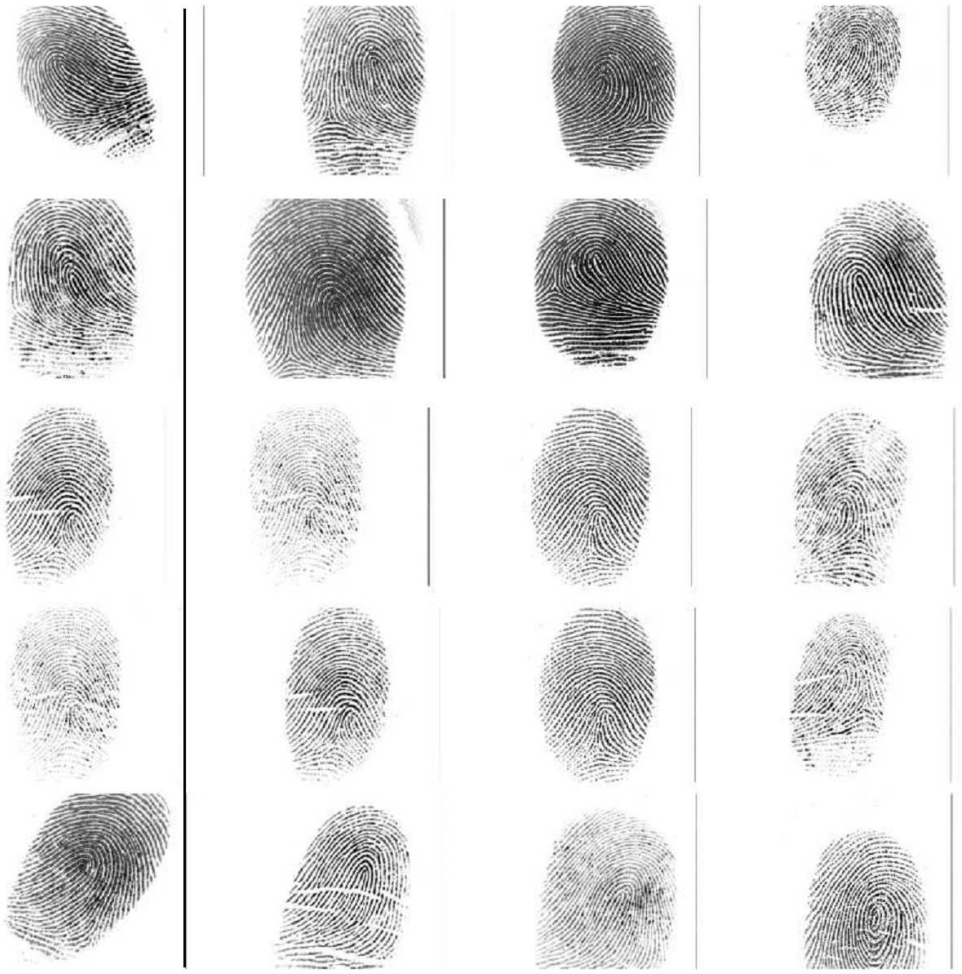

Figure 4.7: Cohort selection for fingerprints. In each row, columns 2-4 display the automatically selected cohort sets for the fingerprint in the first column. The Bozorth 3 matcher [89] used to generate raw similarity scores is a minutiae-based matcher. Therefore, the cohort fingerprints are not always perceptually similar to the corresponding enrolled fingerprint.

3. How large a cohort do we need to improve matching performance? How does the performance vary as the cohort size is increased?

4. How effective is the proposed simple approach in selecting a compact cohort?

5. Is the selection strategy general enough to work seamlessly across different biometrics and matchers?

6. Is cohort-based matching effective in scenarios with multiple enrolled samples per identity?

7. How should the cohort scores be fused with raw similarity scores?

4.5.2 Cohort analysis for biometric verification and identification

In this section, we describe algorithms that utilize cohort sets to improve biometric verification and identification performance. Rigorous experiments are performed to show how the proposed techniques utilize the neighborhood class information present in the form of a cohort for improving the performance. The score-based strategy proposed in the preceding section, is used for cohort selection which in a way stress-tests the proposed cohort-based biometric matching algorithms.

4.5.2.1 Normalization-based cohort analysis

Here we propose a technique that is motivated by the similarity score ratio described in Section 4.4. The similarity of a query with the claimed identity is computed as the ratio of its (raw) similarity with the claimed identity divided by

the (raw) similarity with the complement class $\bar{w}$, i.e.,

$$S(x, w) = \frac{s(x, w)}{s(x, \bar{w})}. \tag{4.4}$$

Here $s(x, \bar{w})$ is the similarity score of the query with the complement class. The raw similarity with the claimed identity can directly be determined using the available matcher. In the absence of a statistical model, the likelihood score to represent the complement class needs to be computed based on a suitable function of the raw similarity scores of the query with the similarities of the cohort of the claimed identity. Assuming the cohort set to be of size k, this is achieved using the following max-rule

$$s(x, \bar{w}) = \max\{s(x, w^1), s(x, w^2), \ldots, s(x, w^k)\}, \tag{4.5}$$

where $s(x, \bar{w})$ is the similarity score with the background and

$$\{s(x, w^1), s(x, w^2), \ldots, s(x, w^k)\} \tag{4.6}$$

is the set of similarity scores of the query with the cohort w^j's for the enrolled biometric w.

The cohort set for each enrolled class is pre-computed off-line using the proposed cohort selection approach. For the verification task, the claim of the query is validated through an appropriate threshold on the normalized similarity score. For identification, the enrolled templates are rank-ordered based on the normalized scores as opposed to the raw similarity scores as done traditionally.

4.5.2.2 SVM-based cohort analysis

As motivated in Section 4.4, a cohort set is selected to represent the background class. Given the cohort, the likelihood of the query biometric generated by the background class has to be evaluated. Though collectively the cohort represents the background class, there is no obvious optimal way to compute the required likelihood measure $s(x, \bar{w})$ or to compute the score ratio $S(x, w)$ in (4.4). The approach based on max-rule based is quite effective (as empirically shown in Section 4.6), but it is not clear if this is the best way to utilize the additional information provided by the cohort. This has motivated us to look for ways that are more effective in using the cohort scores.

In this section, we propose a machine-learning approach to fuse cohort scores. Experiments show that this novel strategy significantly outperforms the proposed max-rule based scheme. The usefulness of this approach does not end at just fusing the cohort scores. We extend this approach in the following section for scenarios when multiple biometrics samples per identity are available in the database.

Given the raw similarity scores of a query with the enrolled identity and its cohort, we need a suitable function to combine these scores to obtain a consolidated similarity score that appropriately takes the class distributions, biometric quality, etc. into account. This is viewed as a binary classification problem in the sense that the goal is to separate the genuine and impostor classes as far apart as possible. We use a linear SVM to perform this task. The SVMs are learning systems that use a hypothesis space of linear functions in a (usually high-dimensional) feature space,

trained with a learning algorithm from optimization theory [22]. Note that the two classes here are the genuine (similar) biometrics and the impostor (dissimilar) biometrics and not the identity classes themselves.

The raw similarity scores of the query with the available sample of an enrolled identity and its cohort form a feature vector as follows

$$F(x, w) = \big[s(x, w), s(x, w^1), s(x, w^3), \ldots, s(x, w^k)\big], \tag{4.7}$$

where $s(x, w)$ is the raw similarity score of the query with the enrolled biometric w, while $s(x, w^j)$ are the raw similarity scores of the query with the members of its cohort. Fig. 4.8 illustrates the essence of cohort analysis in the SVM framework. In this illustration, the raw similarity score $s(x, w)$ between the query and the enrolled identity is essentially a vector in the input space while the vector $F(x, w)$ consisting of both raw and cohort similarity scores belongs to the feature space. As shown in the figure, the input space may not be linearly separable leading to poor matching performance. A suitable transformation to the feature space has the potential to separate the two classes better, resulting in improved matching performance. Interestingly, the traditional way of using raw similarity scores for verification is equivalent to using a linear SVM with one-dimensional feature vector consisting of the raw similarity of the query with the enrolled biometric.

For data in an n-dimensional space, the SVM tries to find the maximum-margin hyper-plane separating the two classes during training. Though an SVM is normally used as a classifier, we use it here to compute the similarity of the query with the enrolled biometrics taking the cohort information into account. Therefore,

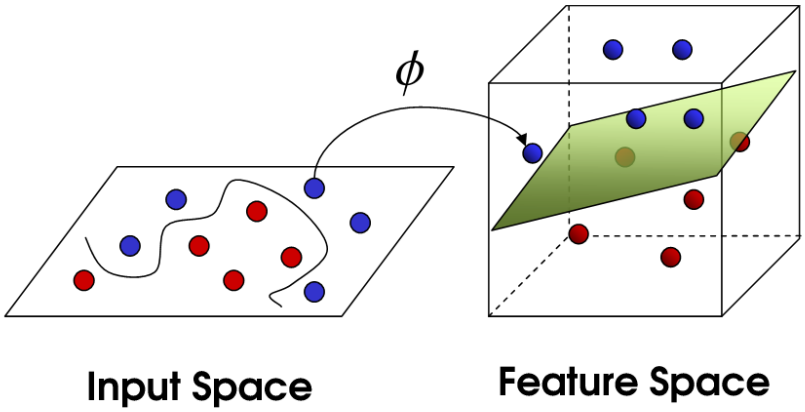

Figure 4.8: SVM framework for cohort-based verification. The raw similarity scores form the input space where the genuine and impostor classes are not linearly separable. The feature vectors $F(x, w)$ form the feature space where the classes tend to be linearly separable.

the distance of each query feature vector from the separating hyperplane (learnt during training) is used as the (dis)similarity measure that it belongs to the genuine class or not. These consolidated similarity scores are used to authenticate queries in the verification scenarios. For identification, these final similarity scores are used to rank-order the biometrics.

4.5.2.3 SVM-based biometric fusion in the cohort framework

The algorithms presented so far in this chapter have implicitly assumed that there is just one biometric sample for each enrolled identity in the database. Though the proposed approaches work well in this verification setting, they will definitely be more valuable if they can be extended to more general scenarios in which multiple

samples are present to characterize each identity.

In this section, we extend the proposed SVM cohort analysis to maximize the benefit when there are multiple enrolled samples per subject. Traditionally such a sample fusion has been restricted to either—

1. score-based cold fusion [87] that uses simple functions ($\max, \min$, etc.) to fuse the similarity scores of the query with all the available samples of the enrolled identity in the database.

2. feature-level biometric fusion schemes [73] in which multiple samples are fused to generate a more reliable template which is used for matching.

The fusion algorithm we propose here belongs to the first category of score-based fusion with the following two major differences from the traditional schemes. First, there is no need to select a function to fuse scores. Instead this is automatically learnt from the training data during the SVM learning phase. Second, cohort scores are also fused with raw similarity scores of the query with the database biometrics to perform verification.

The formulation we propose here is inspired by the one proposed for cases when just one biometric per enrolled identity is present. A linear SVM, as described in the previous section, is used to fuse the multiple (raw) similarity and cohort scores. The only difference lies in the feature vector used to train the SVM and perform verification. Here the feature vector consists of the raw similarity scores of the query with available (multiple) samples of the claimed identity and with the members of its cohort set. Assuming there are m biometrics per identity enrolled in

the database, the feature vector is of the following form

$$F(x,w) = \big[s(x,w_1), s(x,w_2) \ldots s(x,w_m), \\ s(x,w^1)s(x,w^2) \ldots s(x,w^k)\big], \tag{4.8}$$

where $s(x,w_i)$ represent the similarity score of the query with the ith biometric of the claimed identity while $s(x,w^j)$ represent the query similarity with the jth member of the cohort set of the claimed identity. The final similarity measure is derived based on the distance of such a feature vector from the hyper-plane separating the two SVM classes (genuine and impostor). Verification is performed by selecting a suitable threshold on the final similarity score while in identification scenarios, the final consolidated score is used to rank-order the enrolled identities.

4.6 Experiments

We present the results of rigorous experiments performed to evaluate the proposed cohort analysis. The experiments are designed to address the issues described in Section 4.5.1.

4.6.1 Database and matcher description

For fingerprint experiments, we use FVC 2002 [56] DB1 (Set A) database. The data set consists of eight fingerprints each of 100 different subjects. There is a significant variation in the quality of the eight copies of the same print. For most experiments, the raw similarity scores are computed using the NIST Fingerprint Image Software 2 [89]. The Bozorth 3 matcher included in the software is a minutiae-

based matcher that computes the similarity of the fingerprints using the similarity of their minutiae and their relative positions in the fingerprint. Such a matcher depends on the reliable extraction of fingerprint minutiae and therefore, produces low scores when minutiae points are missed or incorrectly located due to noise. Experimental results show that the proposed cohort-based matching algorithms are able to correct a few of such mistakes made by the Bozorth matcher.

For most experiments, one fingerprint per subject is randomly selected and enrolled (out of the eight copies available) to form the enrolled set. The remaining 700 fingerprints are used as queries leading to 700 mated pairs and 700×99 non-mated pairs in verification experiments. Identification performance often degrades with an increase in the number of enrolled identities. Therefore we also use a private IBM dataset consisting of 1000 unique fingerprints with 2 copies each to test the proposed cohort-based identification approaches.

For verification and identification experiments on faces, we use the PIE data set [79]. The PIE database consists of 68 subjects with variations in illumination, pose and expression. We use only the illumination part of the PIE dataset in our experiments. There are 21 images of each subject in 21 different illumination conditions. Figure 4.9 shows the 21 images of a subject from the PIE dataset. In each experiment, one randomly chosen image per subject is enrolled. The remaining 68×20 images are used as queries leading to 68×20 matching pairs and $68 \times 67 \times 20$ non-matching pairs. Two different illumination-insensitive techniques [99] [1] are used to generate the similarity scores. The algorithms in [99] and [1] model a face as a *Linear Lambertian Object*. Given a face image, its illumination-invariant representation in

the form of shape-albedo information is obtained. Though [99] deals mainly with single light scenarios, [1] incorporates the inherent non-linearity in *Lambert's law* to handle more general lighting scenarios. Illumination-invariant matching being an extremely ill-posed problem, there are cases in which these algorithms are unable to output similarity scores that are truly invariant to changes in illumination. We show examples in which the proposed cohort-based methods are able to correct the mistakes made by these matchers.

Figure 4.9: The 21 illumination conditions in the PIE dataset.

4.6.2 Performance metrics

For verification experiments, we use Receiver Operator Characteristic (ROC) curves to evaluate performance. The ROC curve consists of the system False Reject Rate (FRR) plotted against False Accept Rate (FAR) obtained for various verification thresholds (e.g., Figure 4.10). We use logarithmic scale along both FRR and FAR axes to highlight the performance difference between different ROC curves. Other than the visual difference between different ROCs, we also use Equal Error

Rate (EER) and FRR at low FAR to compare different approaches. The EER indicates the point on an ROC at which FRR becomes equal to FAR. FRR at low FAR measure is particularly interesting for high security scenarios.

For identification experiments, we use standard Cumulative Match Characteristic (CMC) curves to evaluate performance. A CMC curve plots the cumulative distribution of rank at which a correct match occurs for a query set. Often performance at rank 1 is used to judge the goodness of the approach.

4.6.3 Max-normalization approach

4.6.3.1 Fingerprint verification performance

Figure 4.10 shows the improvement in verification performance obtained using the proposed normalization scheme. The cohort is of size 10 for each enrolled fingerprint in this experiment. From the available eight fingerprints per subject in the data set, we randomly select one per subject to form the database. The remaining 7 prints per identity are used for querying. Therefore, there are 700 genuine and 700×99 impostor comparisons. The normalization scheme reduces the FRR at 0.001 FAR by about 25% (from approximately 0.062 to 0.045) as compared to the FRR obtained using the traditional raw similarity score based verification. The reduction in FRR at 0.1 FAR using the proposed scheme is around 40% (from approximately 0.023 to 0.014).

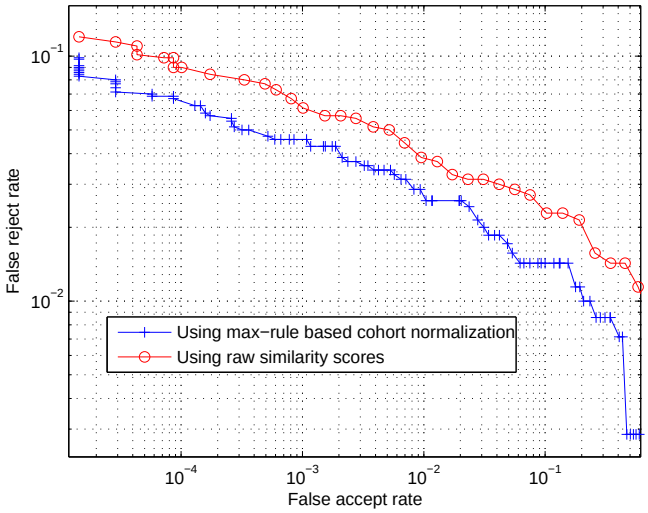

Figure 4.10: The ROC plot shows the verification performance on the FVC 2002 data set. The proposed normalization scheme reduces the False Reject Rate (FRR) by about 25% at 0.001 False Accept Rate (FAR).

4.6.3.2 Statistical significance

To show that this is not a chance improvement in performance because of a peculiar selection of the database, we repeat the experiment using several random choices of enrolled set (again having just one enrolled fingerprint per subject). As before, there are 700 genuine and 700×99 impostor comparisons. Figure 4.11 shows the improvement in the performance using the EER and FRR at low FAR (useful for high-security scenarios). Though the amount of improvement varies with the choice of enrolled set, the proposed approach consistently performs much better than just using the raw similarity scores. On average, there is a reduction of over 25% in both FRR at 0.001 FAR and EER.

We further use the *t-test* to evaluate the statistical significance of the improvement in performance. The *t-value* computed to evaluate the significance is essentially the ratio of difference in mean to the variability of the two sets of performances obtained using raw similarity scores directly and the proposed cohort-based normalization. The expression for *t-value* is as follows

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2_{X_1} + \sigma^2_{X_2}}{n}}}. \tag{4.9}$$

Here n is the number of samples (here, trials), $\bar{X}_1$ and $\bar{X}_2$ are the means while $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$ are the variances of the two performance distributions obtained using the traditional raw similarity score and the cohort scheme. Using the expression, we obtain *t-value* of 10.3 for EER while 15.4 for FRR at FAR=0.001, both of which pass the test of significance by large margins.

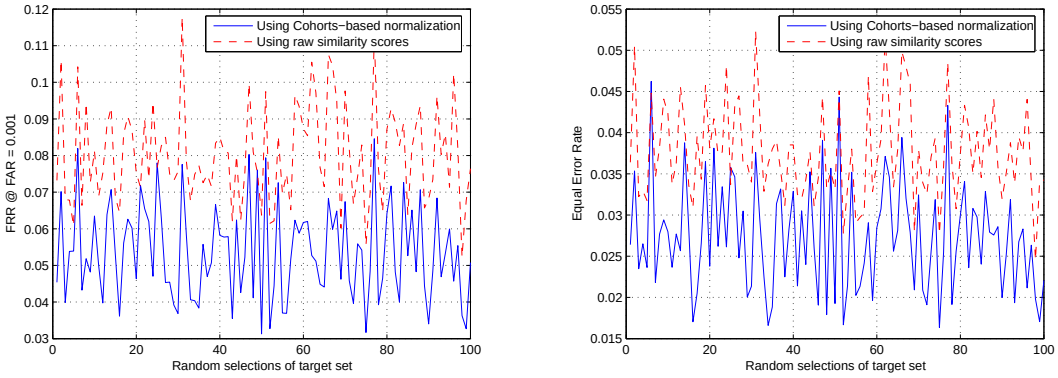

Figure 4.11: The plots show improvement in the FRR at 0.001 FAR and EER using the cohort-based normalization for various random selections of the target set. The variation in performance across various sets is because of the difference in similarity and/or quality of the chosen biometrics with respect to the query ones.

4.6.3.3 Face verification performance

To evaluate the portability of the proposed scheme to other biometrics and matchers, we test the approach on the PIE (face) data set using two different algorithms to compute the raw similarity scores. One face image per subject is randomly chosen to form the data set of face images (gallery). The remaining 20×68 images are used as queries. There are 68×20 genuine and $68 \times 20 \times 67$ impostor comparisons. Figure 4.12 shows the performance on the face data set using the matcher proposed in [1]. The cohort scheme significantly outperforms the traditional raw similarity score based approach. In this experiment, cohort-based normalization reduces the FRR by around 45% (from approximately 0.38 to 0.21) at 0.1 FAR. Similar improvement in verification is observed using the other matcher [99].

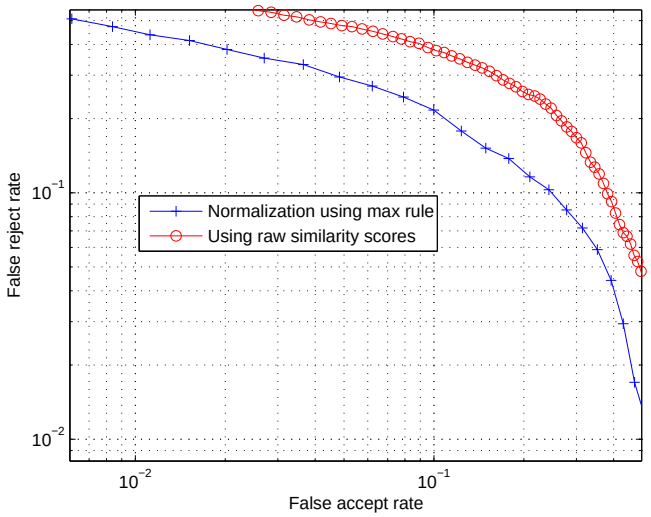

Figure 4.12: The ROC plot shows the improvement in verification performance achieved by normalizing the similarity scores using the selected cohort sets on the PIE data set.

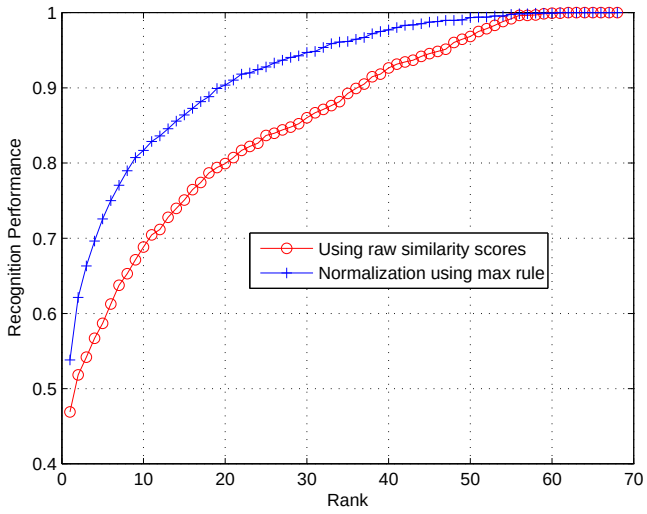

Figure 4.13: The CMC plot shows the improvement in identification performance on the PIE data set.

4.6.3.4 Face identification performance

Figure 4.13 shows the improvement in performance obtained in an identification experiment. The experiment is performed on the images from the PIE data set with one randomly selected face image for each of the 68 subjects enrolled forming the gallery. The remaining 20×68 images taken under different challenging illumination conditions form the query set. The raw similarity scores are computed using the approach in [99]. The cohort approach compares favorably to using raw similarity scores.

In Figure 4.14, we show the top five matches for 5 queries to illustrate the reason for the improved identification performance. Using the raw similarity scores directly, there seems to be a strong correlation in the query illumination condition and the illumination conditions of the top matches. In contrast, the pro-

posed normalization-based cohort analysis does much better in returning the correct match, even when the illumination conditions are quite different.

4.6.3.5 The choice of cohort

Though we notice improvement in the performance using the automatically selected cohort sets, so far it is not clear if the chosen set is compact and is better than any other set. To investigate this issue, we conduct an experiment to observe the verification performance by varying the cohort size. The FVC 2002 DB1 data set is used to analyze the trend. As shown in Figure 4.15, the performance improves till the cohort size is about 20 but does not change much by increasing the cohort size beyond 20. Compared to the raw similarity based verification, the proposed normalization scheme reduces the ERR and FRR at 0.001 FAR by more than 25% using 20 cohorts for each enrolled identity. This shows that one can get good performance using the proposed approach without having to compare the query with a large universal set of fingerprints.

Figure 4.16 compares the performance of the proposed approach with random selection of the cohort sets. For each enrolled subject, a cohort is chosen at random and is used to perform the normalization. As one random selection may not illustrate the true picture, we repeat this experiment 100 times, using a new random cohort set (for each subject) in each run. The distribution of EER and FRR for 0.001 FAR obtained on the FVC 2002 data set using the random cohort selection is shown in Figure 4.16. The EER and FRR obtained with cohort sets selected

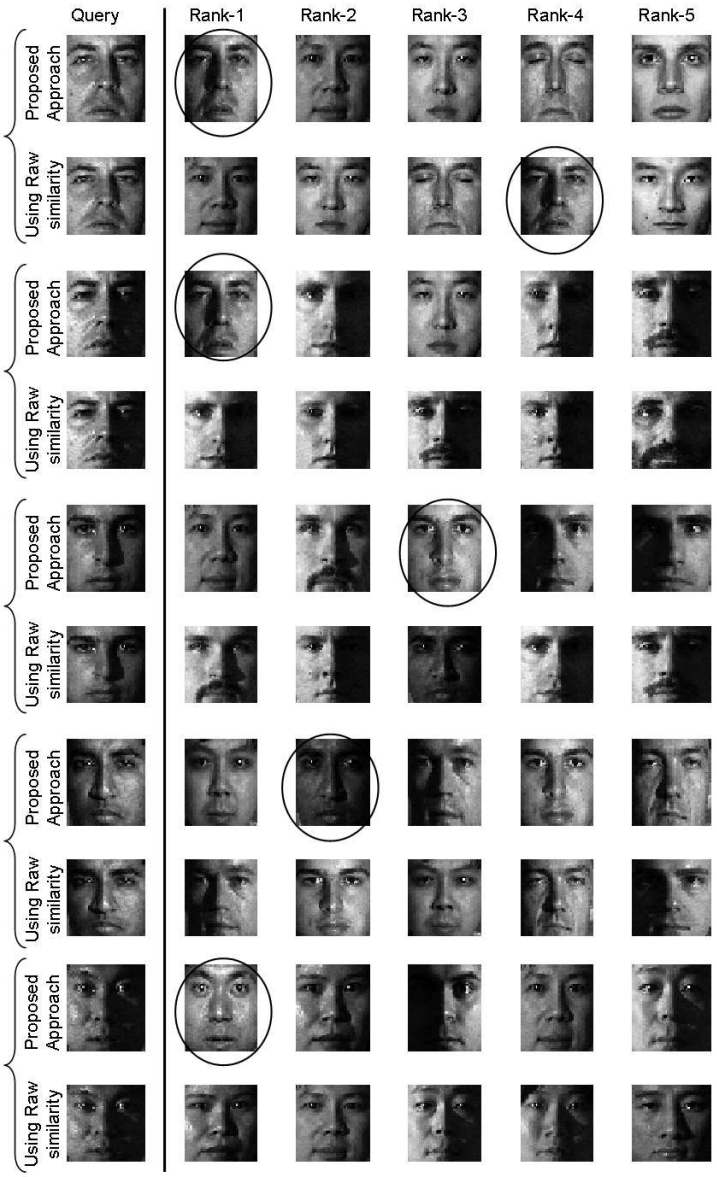

Figure 4.14: Top matches returned in an identification experiment on the PIE data set. Each pair of rows compares the top matches obtained for a given query, using the cohort scheme and raw similarity scores. The correct match is encircled. The matches obtained using the raw similarity show strong correlation with the query in terms of illumination conditions. In contrast, cohort-normalized scores do quite well in returning the correct match.

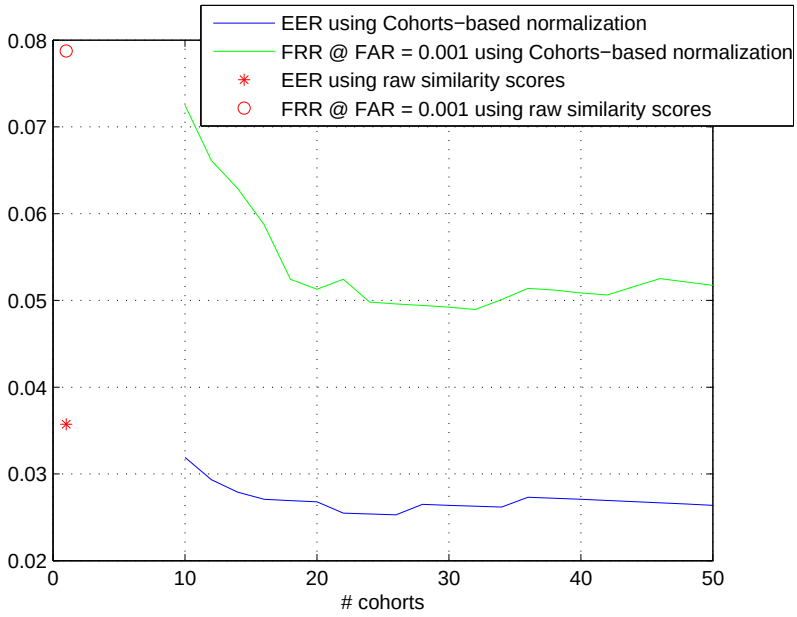

Figure 4.15: The variation of EER and FRR at 0.001 FAR with the size of cohort set to compute the normalized score. As desired, the performance improvement saturates around cohort of size 20. The proposed normalization technique reduces the EER and FRR at 0.001 FAR by over 25%.

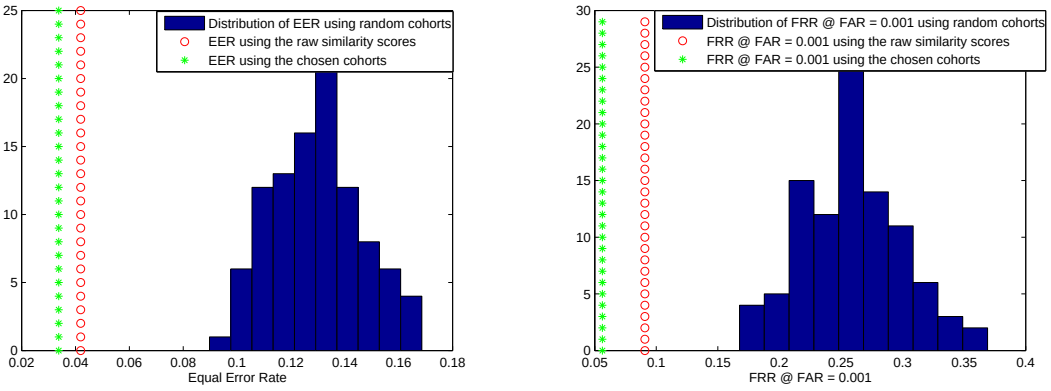

Figure 4.16: The plots show the usefulness of including cohort scores for the verification task. The left plot shows the ERR distribution while the right one shows the FRR at 0.001 FAR. The vertical *starred* line (green) shows the performance using the cohort selected using the proposed score-based scheme while the vertical *circled* line (red) shows the performance using the traditional way of using raw similarity scores. Normalization using random cohort sets makes the performance worse.

using the proposed approach are significantly lower than the ones obtained using random selection. In fact, even using raw similarity scores for verification gives better performance as compared to normalization using random cohort sets.

The observation illustrates the importance of the cohort selection process to perform verification. As illustrated earlier, a cohort is chosen to gain some knowledge about the class distribution in the vicinity of the class at hand. A random cohort selection is not the best way to model the background class effectively. Therefore, the normalization using a randomly selected cohort, randomly boosts or reduces the genuine and impostor scores leading to a performance that can potentially be even worse than using the raw similarity scores. This justifies the poor performance obtained using randomly selected cohort sets in all 100 trials of the experiment.

4.6.4 SVM-based approach

4.6.4.1 Fingerprint and face verification performance

Figure 4.17 shows the verification performance on the FVC 2002 fingerprint data set achieved using the SVM-based approach. We use the free SVM toolbox [83] for this task. During training, the classifier tries to capture the relationships between the raw similarity scores and the cohort scores to determine the maximum-margin hyper-plane that separates the genuine and impostor classes. Therefore, it can be trained using an unrelated dataset.

Figure 4.18 shows the verification performance on the PIE face data set. The two plots correspond to the two different algorithms [99] [1] used to generate the

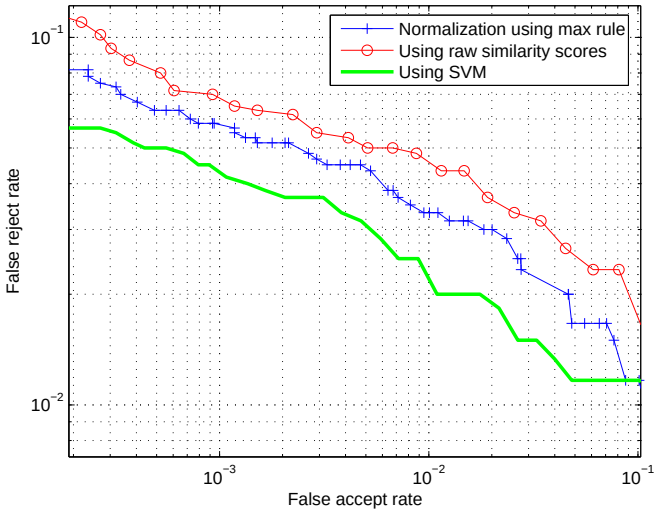

Figure 4.17: The ROC plot shows the improvement in the verification performance on the FVC 2002 data set using the proposed cohort-based schemes.

similarity scores. For each subject, the size of the cohort is five. As before, there are 68×20 genuine and $68 \times 20 \times 67$ impostor scores. The expected improvement in the performance shows that the strategy is generalizable across different biometrics and matchers.

4.6.4.2 Fingerprint and face identification performance

We perform an identification experiment on the PIE data set with one randomly selected image per subject in the gallery set and 68×20 queries. The cardinality of the cohort sets for each subject is five. Figure 4.19 shows the recognition performance obtained in this experiment. As in the verification scenario, the cohort-based approaches perform significantly better than using the raw similarity scores directly. The rank-10 performance of the SVM-based cohort approach is 90% which

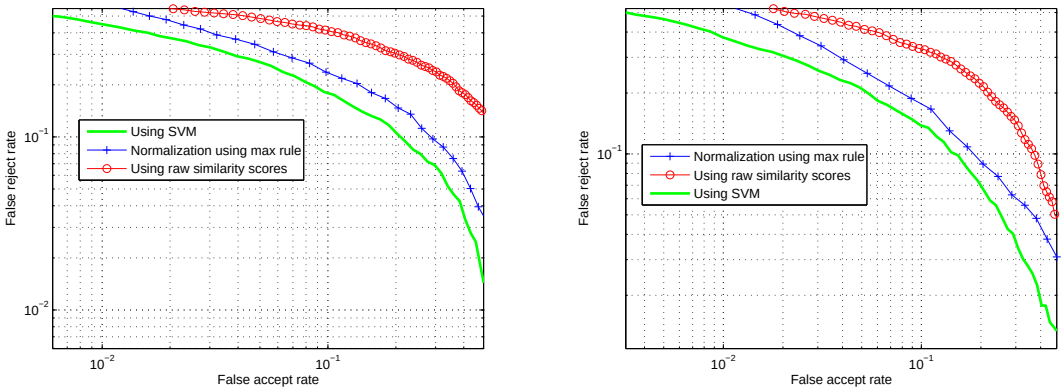

Figure 4.18: The ROC plots show the improvement in verification performance obtained using the proposed cohort-based approaches on the PIE (face) dataset. The matcher in [99] is used to generate the similarity scores for the left plot while the one in [1] is used for the right plot.

is up by 20% as compared to 70% obtained using just the raw similarity scores.

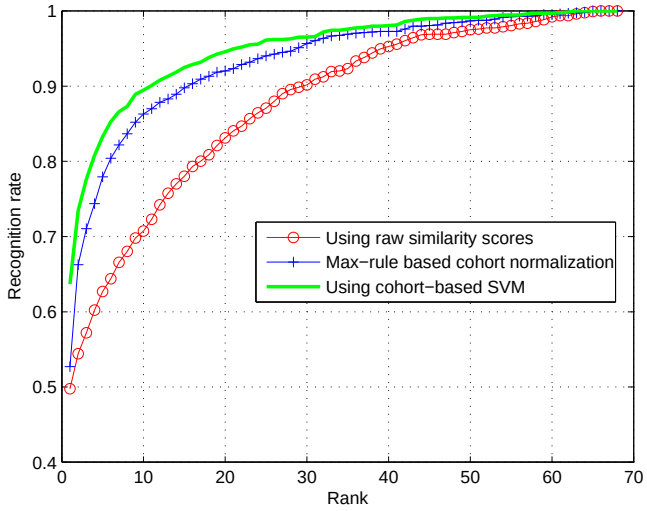

Figure 4.19: The CMC plot shows the improvement in the identification performance using the proposed cohort techniques on the PIE (face) data set.

Another identification experiment is performed using the fingerprints from a private IBM database. The data consists of 1000 fingerprints with two copies of each fingerprint. The similarity scores are computed using the approach in [70]. One fingerprint per identity is enrolled in the gallery and 200 randomly chosen fingerprints from the remaining set are taken as the probes. This is a standard open set identification setting considered to be much more difficult than a closed-set one. Figure 4.20 compares the cohort-based recognition performance against the performance of the raw similarity score based approach. Though the rank-1 performance for all the methods are quite similar, the proposed methods outperform the raw similarity score-based performance at higher ranks.

4.6.5 SVM-based biometric fusion in the cohort framework

Figure 4.21 (left) shows the performance of the cohort-based fusion approach to fuse the scores obtained by comparing the query fingerprint with the multiple enrolled samples of the claimed identity. The experiment is performed on the FVC 2002 [56] data set. For each identity, three randomly selected copies of the fingerprint are enrolled in the database, while the remaining five are used as queries. The cohort set for each identity in such a multi-sample scenario is determined based on the raw similarity scores of the training samples with all the enrolled samples of that identity. The comparison is shown with approaches that use minimum [43], mean [87], and max of the three mated scores (similarity scores with the copies of the claimed identity) as the final score for matching. As shown, the proposed

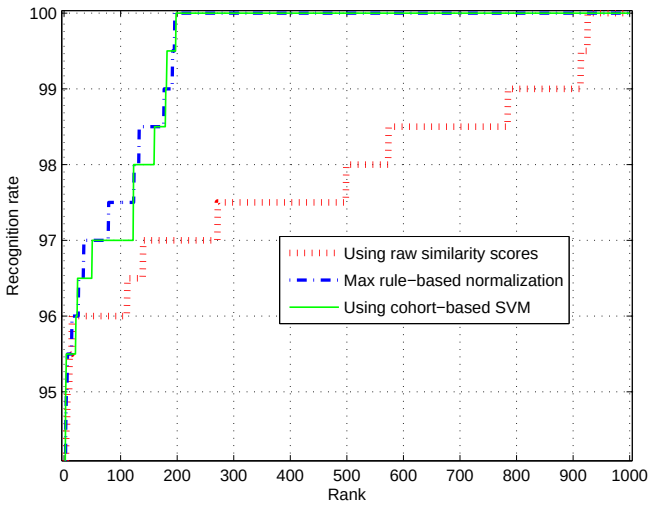

Figure 4.20: The CMC plot shows the improvement in identification performance using the proposed algorithms on a private fingerprint data set. The gallery consists of one randomly selected fingerprint for 1000 subject. 200 randomly chosen fingerprints from the rest of the database are used as queries. Though the rank-1 performance is more or less the same using the three methods, the proposed cohort-based approaches outperform the traditional one at higher ranks. Such an improvement in performance is useful for indexing and retrieval tasks.

approach comfortably outperforms the other popular fusion strategies. Figure 4.21 (right) shows the results of a similar experiment on the PIE face data set. Again, three face images per identity are enrolled and the remaining 18 per subject are used as queries.

Figure 4.21 and Figure 4.17 illustrate the advantages of including non-match samples in validating a query. The inclusion of non-match samples improves the performance even when just a single sample per identity is in the database. However,

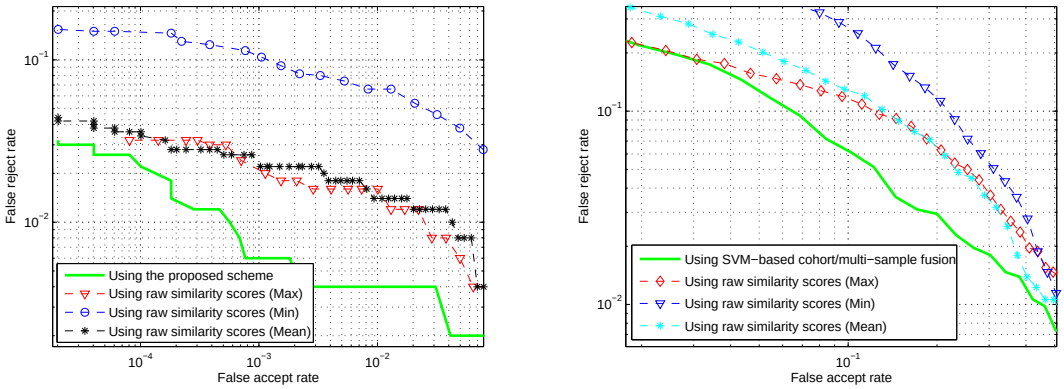

Figure 4.21: The ROC plots show the improvement in fingerprint (left) and face (right) verification performance using the proposed SVM-based approach when there are multiple enrolled samples per identity. The comparison is done with the traditional approaches of taking the max, min, or mean of the similarity scores of the query with the samples of the claimed identity.

the combination of multiple enrolled samples, cohort analysis and the SVM-based fusion strategy performs the best. Therefore, the proposed cohort analysis does not obviate the advantages of having multiple enrolled samples in the database, instead it complements such information to further improve the matching performance. Table 4.1 illustrates this point by comparing the verification performance obtained at specified FAR with and without multiple mated/non-mated samples.

4.7 Summary and discussion

We discussed the limitations of using likelihood-based raw similarity scores for the task of biometric matching. It is shown that much more can be achieved using the large number of non-match biometric samples often present in the database.

	FRR at 0.01 FAR (Fingerprint)	FRR at 0.001 FAR (Fingerprint)	FRR at 0.1 FAR (Face)
Single enrolled (without cohort)	0.045	0.07	0.41
Single enrolled (cohort-MAX rule)	0.033	0.057	0.23
Single enrolled (cohort-SVM)	0.02	0.04	0.17
Multiple enrolled (without cohort)	0.014	0.022	0.12
Multiple enrolled (cohort-SVM)	0.004	0.006	0.06

Table 4.1: Effect of multiple mated and non-mated samples on verification performance. The performance numbers are taken from Fig. 4.21 and Fig. 4.17.

A simple score-based approach is used to select cohort sets for biometrics with no suitable statistical model. The approach makes use of the raw scores returned by a matching algorithm to select the cohort sets. We make no assumption about the biometric or matcher. Neighborhood information in the form of the selected cohort is further used to perform biometric matching. Two different approaches are developed to fuse cohort scores with the raw similarity score of the query with the true biometric. The first one is a simple likelihood ratio-based normalization technique. Experiments show that such a simple approach leads to a significant improvement in both verification and identification performance. In the second approach, the cohort scores are incorporated in the final similarity using a linear SVM. The SVM-based cohort analysis for biometric matching significantly outperforms even the proposed score normalization approach. The approach is further extended to address multiple-

template scenarios. The performance improvement obtained using the cohort-based approaches is significant and consistent across multiple biometrics, data sets, and matching algorithms. It is the combination of multiple enrolled samples, non-match samples and SVM-based fusion that gives the best matching performance.

In all the experiments, the cardinality of the selected cohort set has been the same for each enrolled identity. In theory, this should depend on the individuality of the biometric and therefore should vary across biometric classes and also the quality of query and reference templates. It would be an interesting exercise to examine the impact of variable size cohort. Though the max-normalization scheme can potentially handle variable size cohort sets, the same is not true for the SVM-based approaches. We would like to address this in our future work.

The proposed cohort analysis framework does not assume anything about the data or matching algorithms. Therefore, it should be useful for any pattern matching/classification task. It will be interesting to explore its usefulness for such tasks in completely different domains.

Chapter 5

Physics-based Revocable Face Matching

We present a face reconstruction approach for revocable face matching. The proposed approach generates photometrically valid cancelable face images by following the image formation process. Given a face image, the approach estimates facial albedo followed by a subject-specific key based photometric deformation to generate a cancelable face image. The proposed approach allows for using any available face matcher to perform verification or recognition in the transformed domain, a capability missing from most existing works on cancelable face matching. Experiments are performed to evaluate the performance, privacy and cancelable aspects of the face images reconstructed using the approach. Results obtained are very promising and make a strong case for such backward compatible cancelable face representations that can seamlessly make use of advancements in automatic face recognition research.

5.1 Introduction

The advancement and popularity of biometric systems has brought concerns of *biometric-theft*. Unlike PINs or passwords, which can be changed at will when compromised, biometric traits are unique and permanent. This leads to the observation that though biometrics are authentic, they are not secure (or private like

passwords). If compromised, biometric signatures cannot be revoked or canceled. It allows for rogue establishments to track subjects across databases and institutions without consent.

The concern of biometric privacy has led to research efforts to secure biometrics [69]. One popular way is to combine biometrics with user-provided keys or passwords to make them secure. The user-specific private key is used to encrypt biometric template which is stored in the database. The encrypted template stored in the database is used for further matching. For matching purposes, the same encryption scheme is used to transform the query template to compare it with the stored secure template. Quite clearly, such an approach combines the advantages of biometric based authentication and password-based privacy and revocability.

Ratha *et al.* [69], in their pioneering work, present several one way (non-invertible) transforms for constructing multiple secure identities from a fingerprint. They show that a user can be given as many biometric identifiers as needed by issuing a new transformation key which can be canceled and replaced when compromised. Savvides *et al.* [74] extend their earlier work on correlation filter based face matching to produce cancelable biometric representations. They show that convolving the training images with any random convolution kernel before building the filter does not change the resulting correlation output peak-to-sidelobe ratios, thus preserving the authentication performance while maintaining privacy. Boult [15] introduces robust biometric transform that can be used for revocable face authentication. The transformed feature vector is separated into a fractional and an integral part where the integral part is encrypted while the other is left unsecured. Teoh

et al. [84] present a biometric-hash framework by integrating biometric and user-specific password using Random Multispace Quantization (RMQ). The process is carried out by first obtaining a fixed length feature vector from the input biometric followed by a non-invertible random subspace projection and quantization.

One of the main problems in encryption-based biometric authentication approaches is that they tend to be sensitive to variability/noise in the input biometric space. Inherently, biometrics show a great deal of intra-class variability either due to natural causes or external imaging conditions. It is difficult to design an encryption scheme that can suitably transform features extracted from such input data minimizing within-class scatter as compared to the between-class scatter. Unlike input biometric space, in which one can perform some sort of learning to account for such intra-class variabilities, such learning is not easy in the encrypted space. Another drawback of encrypting feature extracted from the input biometrics is that such approaches tend to be specific to the features used. Therefore, it may not always be easy for such approaches to take advantage of the new developments in the field of biometric matching.

In this chapter, we propose a physics-based face reconstruction approach that addresses these issues for cancelable face matching. Given an input face image, the proposed technique reconstructs a transformed face image that can be matched using any publicly available matcher. Depending on the capability of the face matcher used to compare the reconstructed face images, the variability/noise in the input biometric can be accounted for even though matching is performed in the transformed domain.

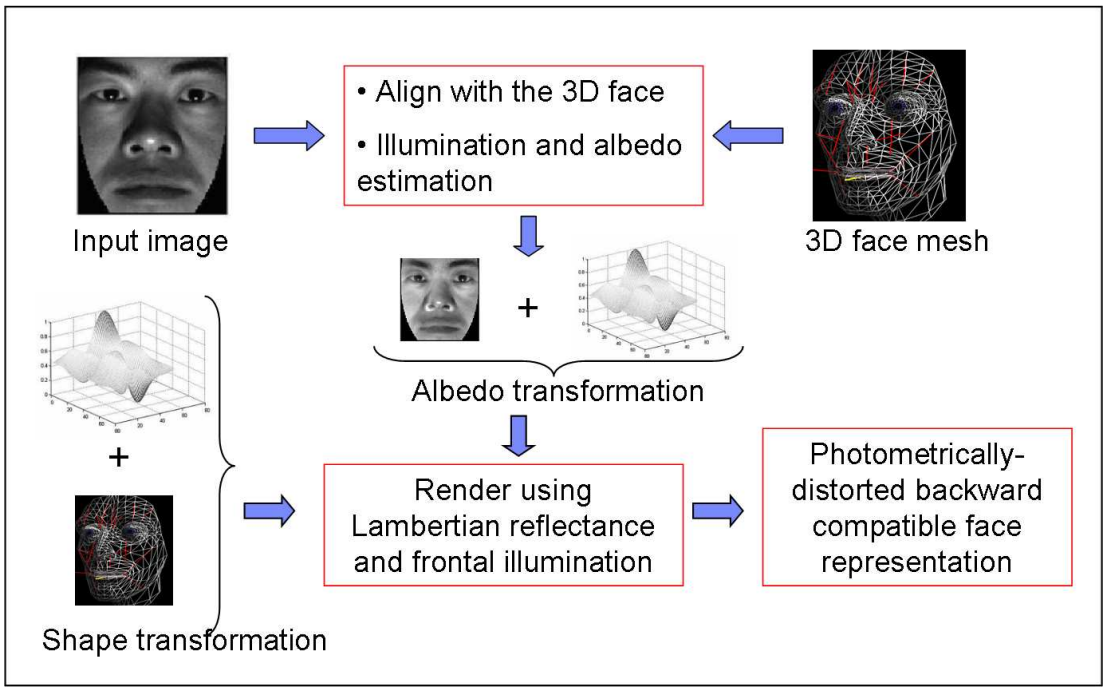

Figure 5.1: A schematic of the proposed approach.

5.2 Organization of the chapter

The chapter is organized as follows. The proposed face reconstruction approach is described in Section 5.3. Results of extensive experimental evaluations performed to validate the usefulness of the approach in terms of privacy, security and matching performance are shown in Section 5.4. The chapter concludes with a brief summary and concluding remarks in Section 5.5.

5.3 Physics-based face reconstruction

An input face image is the result of an interplay between the physical characteristics of a real 3D face, external imaging environment (illumination, view, etc.),

capturing device, etc. Our goal is to create another face image from the input image that can be used as a cancelable representation of the face, that can be matched using any available face matcher. One of the critical components of such an approach is appropriateness of the transformation from the input face to the desired cancelable face image such that the output is photometrically valid. Quite clearly, direct manipulation of the image intensity values may lead to images which are physically unrealizable.

In this work, we first estimate albedo from a single input face image. This is followed by user-specific key based transformation of albedo. Due to the absence of real 3D shape information of the input face and the difficulty in estimating shape from a single image, we use a transformed (distorted) version of the average facial 3D shape. As with albedo, the kind and amount of shape distortion is guided by the user-specific key. Once we have the distorted albedo and shape, we render a face image that does not reveal the identity of the subject in the input image. The vast range of possible transformations (or distortions) on estimated albedo and 3D facial shape provides cancelability to the approach for scenarios when the template is compromised. Figure 5.1 shows a schematic of the proposed approach. The various steps of the proposed approach are described in the following sections.

5.3.1 Albedo estimation

The first step of our approach is to estimate surface albedo from the input face image. Without loss of generality, face images are assumed to be pre-cropped and

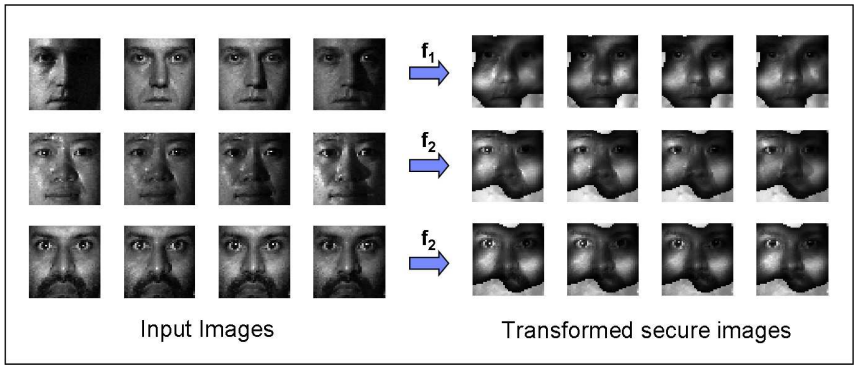

Figure 5.2: Examples of transforms applied to a few images from the PIE dataset.

pose-normalized to be in the frontal pose. Albedo estimation is performed using the non-stationary stochastic filtering framework proposed by Biswas *et al.* [11]. Given a coarse albedo map (obtained using the average facial 3D shape of humans), the approach estimates a more robust albedo map by accounting for the statistics of errors in surface normal and light source estimation in an image restoration framework [3]. Readers are encouraged to read [11] for technical details.

5.3.2 Albedo and shape transformation

In this step, the estimated albedo and the average facial 3D shape is transformed using the user-provided secure key. From the large number of available choices for such a transformation, we use one based on mixture of Gaussians. Albedo is transformed by multiplying it with a mixture of Gaussian image. The number, peak locations, and variance of the Gaussian distributions is determined using the key. For shape transformation, we generate another mixture of Gaussians surface and linearly combine it with the average 3D facial shape. As with albedo, the

user-specific key determines the specifics of the mixture of Gaussians surface.

5.3.3 Image reconstruction

The transformed albedo and shape are used to reconstruct a photometrically valid face image. Assuming Lambertian reflectance model, the desired image can easily be generated using the following relation

$$\mathbf{I}_r = \rho_r \max(\mathbf{n}_r \cdot \mathbf{s}, 0) \tag{5.1}$$

where $\mathbf{I}_r$ is the reconstructed transformed face image, ρ_r is the transformed albedo map, $\mathbf{n}_r$ is the transformed surface normal map and $\mathbf{s}$ is the light source direction which is taken to be $[0, 0, 1]^T$ for frontal lighting. Figure 5.2 shows a few images generated using this approach.

5.4 Experimental evaluation

In this section, we describe the experiments performed to evaluate the usefulness of the proposed backward-compatible cancelable face reconstruction. In our implementation, the user-defined keys are generated using a random number generator that defines the number (5-10), location and variance of Gaussian peaks required to generate distorted images. The experiments are performed on illumination part of the PIE face dataset [79] that consists of face images of 68 subjects under 21 challenging illumination conditions (Figure 5.3). Each experiment consists of matching images in one illumination scenario against another. This results in 68 genuine and 68×67 impostor pairs. All the verification results and score distributions presented

in this chapter are obtained by repeating the experiment for all $\binom{21}{2}$ pairs of illumination conditions, thereby resulting in $\binom{21}{2} \times 68$ genuine and $\binom{68}{2} \times 68 \times 67$ impostor pairs. In addition to evaluating privacy and revocability, experiments also reflect the illumination-invariance property of the approach. Illumination-invariance is a byproduct of using albedo images as opposed to direct intensity images for transformed face reconstruction. In all the experiments, similarity scores are computed using Principal Component Analysis (PCA). The PCA bases are learnt from the FRGC training data [66] that consists of 366 training face images.

5.4.1 Performance

Figure 5.4 shows the genuine and impostor score distributions obtained using the reconstructed faces. In this experiment, every subject has a different transformation key. The plot shows the distributions obtained in two different runs of the experiment using different sets of keys for each identity. The genuine/impostor distributions hardly overlap leading to almost flawless performance. Note that the proposed approach is able to account for illumination variations present in the original images, a capability missing in most previous cancelable face matching approaches.

5.4.2 Lost key scenario

We now evaluate the performance of the approach by using the same transformation key for all the subjects. This simulates the stolen/lost key scenario when an adversary somehow gets hold of a user's key and tries to break into the system

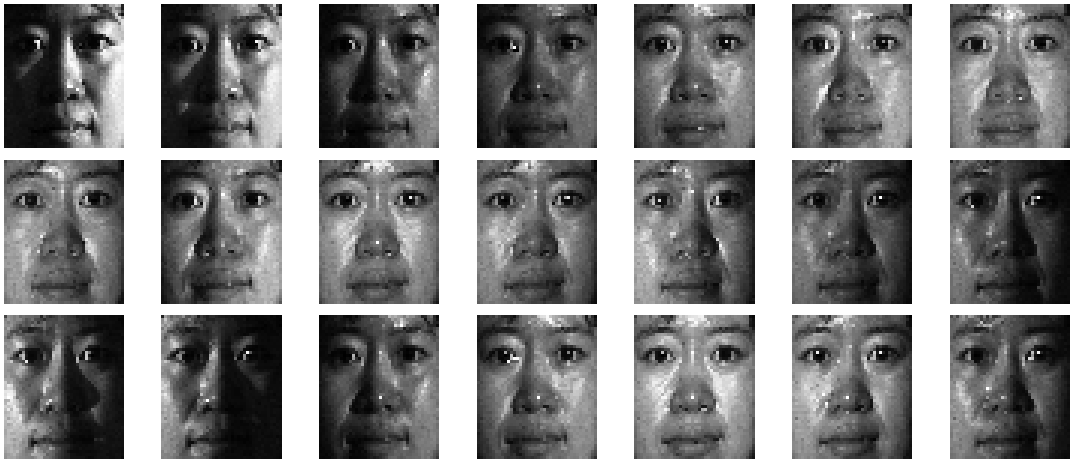

Figure 5.3: The 21 illumination conditions in the PIE dataset.

using that key. Figure 5.5 shows that the separation between the genuine/impostor distributions is preserved even when the same transformation key is used for all the subjects. In fact, the reconstructed faces perform better in a verification setting as compared to the original input images even when same transformation key is used for reconstruction (Figure 5.6).

5.4.3 Privacy and revocability

We first compare the genuine and impostor score distributions obtained while matching the reconstructed images against the original input images , i.e., using original images in the gallery while the reconstructed images as queries. The experiment is repeated by replacing the original images with another set of transformed images generated using a different set of keys. Figure 5.7 shows that the genuine and impostor score distributions have hardly any separation indicating that the reconstructed faces reveal hardly any identifying information when compared against the original

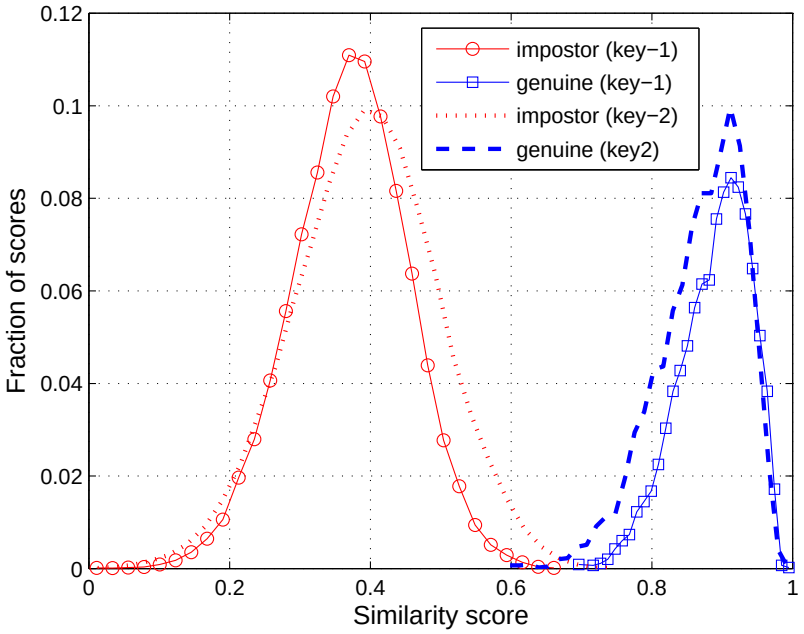

Figure 5.4: Impostor and genuine score distributions obtained using the generated face images. The plot shows results obtained in two different runs of the proposed algorithm using different set of keys.

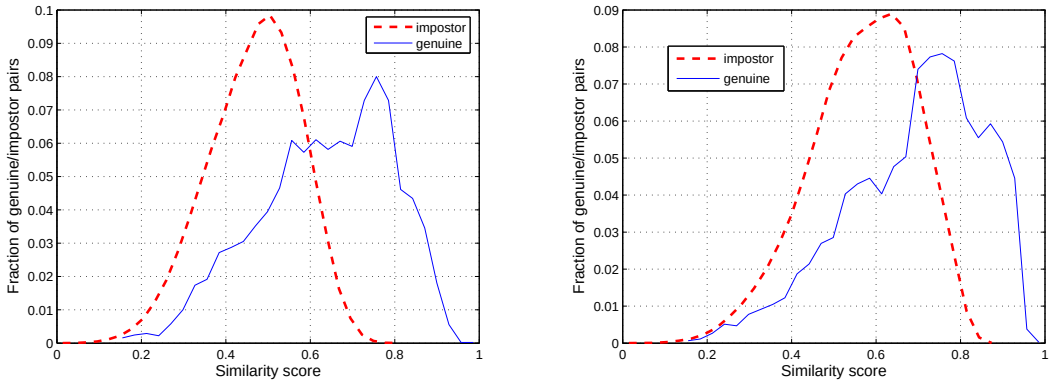

Figure 5.5: Lost key scenario: Genuine/impostor score distributions obtained in matching experiments on the face images reconstructed using the same key for all identities (left) and the original input images (right). The genuine/impostor separation is preserved even when same key is used to transform all identities.

or (differently) transformed images. To further evaluate the privacy/revocability of the proposed approach, we also compare the mated score distributions obtained while matching 1) original images against transformed images (should be low for privacy), 2) transformed images against other transformed images generated using the same key (should be high for good performance), and 3) transformed images against other transformed images generated using a different key (should be low for revocability). Figure 5.8 shows that the genuine score distributions are in fact as desired proving the privacy and revocability aspects of the proposed approach.

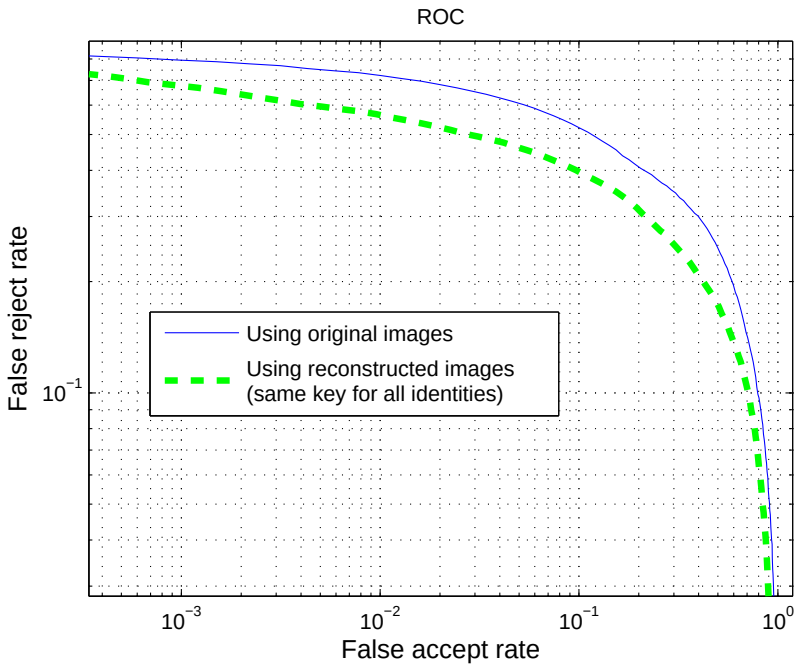

Figure 5.6: Lost key scenario: Comparison of Receiver Operator Characteristic (ROC) curves obtained in a verification experiment with the original images in the gallery while the transformed faces (generated using same key for all identities) as queries.

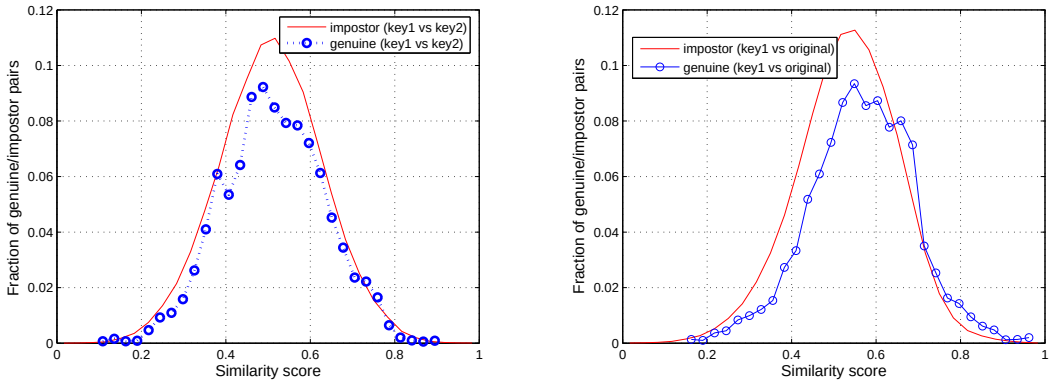

Figure 5.7: Privacy/revocability test: 1) Genuine/impostor score distributions obtained using the transformed image set 1 as the gallery and transformed image set 2 as queries (left), and 2) Genuine/impostor score distributions obtained using the original images in the gallery and the transformed ones as the queries (right).

5.5 Summary

Unlike inter-operable fingerprint templates, there is no common format for face features other than the image itself. In order to achieve backward compatibility, we proposed a physics-based face reconstruction approach for cancelable face matching. Given an input face image, the proposed technique reconstructs a new transformed face image that can be matched using any available matcher. We tested our approach using a standard database with several different transforms. The results are extremely encouraging. We will test the scalability of our approach using larger databases and publicly available face matchers in future. Note that though it is impossible for an adversary to get the original image back from just a transformed image and the corresponding key, the transform is invertible if he/she has access to

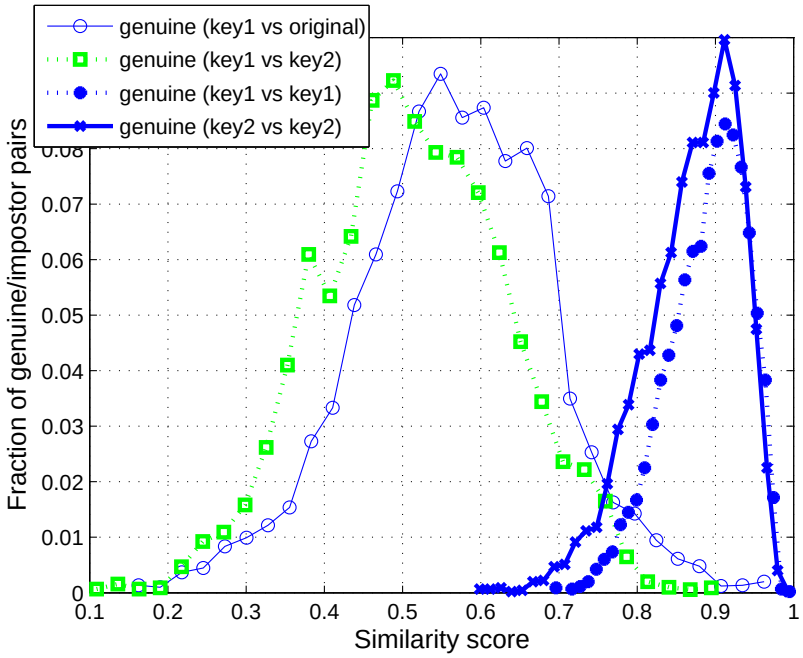

Figure 5.8: Privacy/revocability test. Comparison of distributions of mated scores: 1) Original image against transformed image (should be low for privacy), 2) Transformed image against other transformed image generated using the same key (should be high for good performance), 3) Transformed image against other transformed image generated using a different key (should be low for revocability).

the exact distortion algorithm used to obtain the transformed images.

Chapter 6

Face Recognition and Tracking in Videos

Traditionally, face recognition research has been limited to recognizing faces from still images. The advent of inexpensive high-resolution video cameras and increased processing power makes it viable to capture, store and analyze facial videos. Videos have the advantage of providing more information in the form of multiple frames. Moreover, video input allows to capture any temporal signature present that can be used to characterize and hence, identify a person. Video-based face recognition (VFR) is particularly useful in surveillance scenarios in which it may not be possible to capture a single good frame as required by most still image based methods.

6.1 Challenges in automatic video-based face recognition

Effective utilization/fusion of the information (both spatial and temporal) present in a video to achieve better generalization (for each subject) and discriminability (across different subjects) for improved identification is one of the biggest challenges faced by a VFR system. The fusion schemes can range from simple selection of good frames (which are then used for recognition in a still-image based recognition framework) to estimation of the full 3D structure of a face which can then be used to generalize across pose, illumination, etc. The choice may depend pri-

marily on the operational requirements of the system. For example, in a surveillance setting, the resolution of the faces may be too small for reliable shape estimation. The choice also limits the recognition capability of the system. A simple good frame selection scheme will not have the capability to generalize appearance across pose variations and thus requires the test video to have some pose overlap with the gallery videos. Effective modeling of subject-specific facial characteristics from video data can only be achieved if the changes in facial appearance during the course of the video are appropriately attributed to different factors like pose changes, lighting, expression variations, etc. Unlike still image based scenarios, these variations are inherent in a VFR setting and must be accounted for to reap the benefits of extra information provided by the video data. In addition, due to the nature of the input data, VFR is often addressed in conjunction with tracking problem which is a challenging problem by itself. In fact, more often than not, tracking accuracy depends on the knowledge of reliable appearance model (depends on the identity provided by the recognition module) while recognition result is dependent on the localization accuracy of the face region in input video.

In this chapter, we propose two approaches to address the problem of VFR. In the first, we learn the appearance and dynamics of a moving face given its video without explicit 3D reconstruction of the face. Face is modeled using an autoregressive and moving average (ARMA) model. Subspace angles between the learnt models are used to measure the similarity of faces. Though the results obtained using this approach are very promising, one of the main limitation of the algorithm is that it does not take the 3D shape of the face into account. This makes the approach

sensitive to the extent of pose overlap across gallery and probe videos. To address this, we propose a particle-filter based algorithm to recover 3D configuration of face in each frame of the video. The recovered 3D configuration is used to normalize for pose variation. This allows us to perform VFR even when there is limited/no pose overlap across gallery and probe videos.

6.2 Organization of the chapter

The rest of the chapter is organized as follows.

The first part of the chapter describes the proposed ARMA model-based approach. Section 6.3.1 describes a few related works on VFR. We provide an intuition for the proposed ARMA-based approach in Section 6.3.2 . This is followed by the details of the approach in Section 6.3.3. Section 6.3.4 describes various distance metrics used for comparing the generated ARMA models to estimate the degree of similarity between two face videos. We present the details of our experiments and their significance in Section 6.3.5.

The second part of the chapter describes the proposed facial tracking algorithm. Section 6.4.1 describes a few existing approaches for facial tracking. In Section 6.4.2, we discuss the geometric modeling of the face. Section 6.4.3 presents the features used for tracking. In Section 6.4.4 we discuss our particle filter-based tracking algorithm. Section 6.4.5 presents experiments on tracking and recognition. The chapter concludes with a summary and discussion in Section 6.5.

6.3 ARMA model-based approach for VFR

6.3.1 Related work

Recently, methods based on multiple images/video sequences that do not involve creating an explicit 3D model have been suggested. Such an approach is supported by many psychophysics works like [17], where authors argue that a 3D object is represented as a set of 2D images (instead of a 3D model) in our brains. Leaving out the algorithms based on simple voting, most of these methods make use of either the natural variability in a face (due to variation in pose or expression) or the information present in the temporal variation of face. In [12], Biuk *et al.* recognize a face from a sequence of rotating head images by computing the Euclidean distances between trajectories formed by face sequences in PCA feature space. The Mutual Subspace Method (MSM) proposed in [92], considers the angle between input and reference subspaces formed by the principal components of the image sequences (not necessarily ordered) as the measure of similarity. This approach discounts the inherent temporal coherence present in a face sequence that might be crucial for recognition. In [75], face recognition is cast as a statistical hypothesis testing problem, where a set of images is classified using the Kullback-Leibler divergence between the estimated density (assumed to be Gaussian) of the probe set and that of gallery sets. This method is based on the underlying assumption that face recognition can be performed by matching distributions. However, two such distributions for the same subject might look very different depending on the range of poses and expressions covered by the two sets. Moreover, this approach

is sensitive to illumination changes. In [54], Liu *et al.* learn temporal statistics of a face from a video using adaptive Hidden Markov Models to perform video-based face recognition. In [91], kernel principal angles, applied on the original image space and a feature space, are used as a measure of similarity between two video sequences. Zhou *et al.* [101] propose a tracking-and-recognition approach by resolving uncertainties in tracking and recognition simultaneously in a probabilistic framework. Lee *et al.* [50], in their recent work, represent each person by a low-dimensional appearance manifold, approximated by piecewise linear subspaces. They present a maximum a posteriori formulation for recognizing faces in test video sequences by integrating the likelihood that the input image comes from a particular pose manifold and the transition probability to this manifold from the previous frame. Among the methods mentioned, Lee *et al.* [50] method seems to be the one most capable of handling large 2D and 3D rotations.

Although many previous methods make use of temporal information present in face videos to improve recognition, there has been no attempt to model a moving face as a dynamical system. Our work can be seen as an attempt to explore this. We present a method for modeling a moving face as a linear dynamical system to perform recognition. Each frame of a video is, therefore, assumed to be the output of the dynamical system particular to the subject. Our work follows [25] and [81], where Soatto *et al.* used a very similar idea to characterize dynamic textures. In [10], they use the same approach for recognizing different types of human gait. As in [81], we also use a first order ARMA model. The difference is that here we try to capture the varying appearance (due to pose and expression variation) and dynamics of

face using this framework. Once the models are estimated, recognition is performed by computing distances between ARMA models corresponding to probe face and gallery faces. We use several distance metrics based on subspace angles between the ARMA models.

6.3.2 Motivation

Suppose we want to model a point constrained to move in xy-plane (Figure 6.1). The position of the point at any time instant is guided by its position at the previous time instant. The point has an attribute, say color, that varies with time depending on the position of the point. In this framework, color of the point

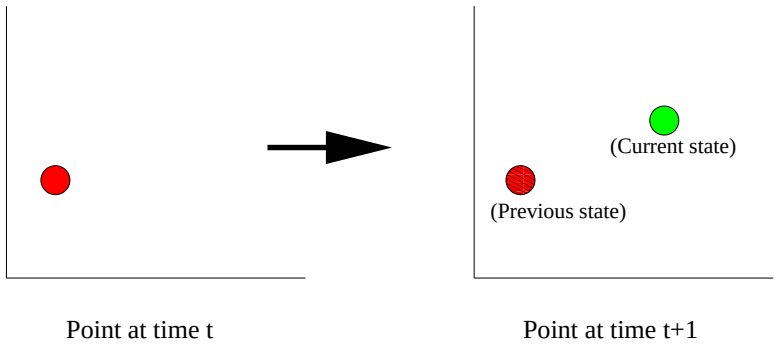

Figure 6.1: Motivation: modeling the dynamics of a moving point where color is the only observable attribute.

is the only thing that is visible to the outside world. Modeling such a phenomenon essentially requires two mappings viz.,

$$Position_{t+1} = \phi(Position_t) \tag{6.1}$$

$$Color_t = \psi(Position_t) \tag{6.2}$$

where the subscript denotes time instant. Given a sequence of observations (colors), if we can estimate ϕ and ψ, we are done. This is quite similar to the case of face videos if we think of the pose of the face as the position of the point and the 2D appearance of the face as the color of the point. The dependence of the appearance on the pose is analogous to that of the color on the position. The degree of goodness of such a model is limited by the choice of the forms of the mappings ϕ and ψ and the accuracy of their estimation. In general, these mappings can be arbitrarily complex but methods to estimate them are often not known. In our work, we get promising results by assuming them to be linear.

6.3.3 Framework for modeling

In this section, we develop a mathematical formulation that helps us in estimating the unknown parameters of the model, we use, to characterize a moving face sequence.

If the mappings ϕ and ψ are some linear operators, (6.1) and (6.2) can be written as:

$$x(t+1) = Ax(t) + v(t) \tag{6.3}$$

$$I(t) = Cx(t) \tag{6.4}$$

where, $I(t)$ is appearance of the face at time instant t, $x(t)$ is a state vector that characterizes the pose of the face, A and C are matrices representing the linear mappings and $v(t)$ is an IID realization from some unknown density $q(.)$, that takes care of the implicit assumption that the dynamical system is driven by an IID

process.

Suppose at each time instant t, we can measure only a noisy version of $I(t)$ i.e., $y(t) = I(t) + w(t)$ where $w(t)$ is an IID sequence drawn from a known distribution. This leads to a first order ARMA model as follows:

$$x(t+1) = Ax(t) + v(t) \tag{6.5}$$

$$y(t) = Cx(t) + w(t) \tag{6.6}$$

This formulation has similarities with the pioneering work by Ali [2], where he addresses the problem of estimation and prediction for stationary spatial-temporal processes. He too uses a simultaneous linear model to represent spatial-temporal processes.

At this stage, we use the closed-form solution as described in [81], where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$, $v(t) \sim \mathcal{N}(0, Q)$ and $w(t) \sim \mathcal{N}(0, R)$. This makes our model a linear dynamical system driven by zero-mean Gaussian noise. Given a video sequence (i.e., a sequence of observation vectors $y(1), \ldots, y(\tau)$), we need to estimate the parameters A, C, Q and R to model the face in the video.

6.3.3.1 Closed-form solution to estimate the parameters

Let $Y^\tau = [y(1), \ldots, y(\tau)] \in \mathbb{R}^{m \times \tau}$ with $\tau > n$, then for $\{t = 1 \ldots \tau\}$, (6.6) can be written as

$$Y^\tau = CX^\tau + W^\tau\,; \quad C \in \mathbb{R}^{m \times n} \tag{6.7}$$

where X and W are defined in a manner similar to Y. If singular value decomposition (SVD) of Y^τ is $Y^\tau = U\Sigma V^T$, where Σ is a diagonal matrix, $U \in \mathbb{R}^{m \times n}$, $U^T U = I$,

$V \in \mathbb{R}^{\tau \times n}$ and $V^T V = I$, then

$$\hat{C}(\tau) = U \tag{6.8}$$

$$\hat{X}(\tau) = \Sigma V^{\tau} \tag{6.9}$$

$$\hat{A}(\tau) = \Sigma V^T D_1 V (V^T D_2 V)^{-1} \Sigma^{-1} \tag{6.10}$$

where $D_1 = \begin{pmatrix} 0 & 0 \\ I_{\tau-1} & 0 \end{pmatrix}$ and $D_2 = \begin{pmatrix} I_{\tau-1} & 0 \\ 0 & 0 \end{pmatrix}$, and

$$\hat{Q}(\tau) = \frac{1}{\tau} \sum_{i=1}^{\tau} \hat{v}(i) \hat{v}^T(i) \tag{6.11}$$

where $\hat{v}(t) = \hat{x}(t+1) - \hat{A}(\tau)\hat{x}(t)$, give a closed-form solution (suboptimal in the sense of Frobenius).

6.3.4 Framework for recognition

Given gallery and probe face videos, the model parameters (as explained in Section 6.3.3) for each one of them are estimated. The gallery model, which is *closest* to the probe model, is assigned as the identity of the probe. We here discuss the metrics used to measure this degree of similarity.

Computing the L_2-norm of the difference between corresponding model matrices as a measure of distance will not suffice as it implicitly ignores the underlying geometry of the subspaces which can be non-Euclidean. We make use of subspace angles between ARMA models for this cause. We follow the mathematical formulation given in [20] to compute these angles. The subspace angles are defined as the principal angles between the column spaces generated by the observability matrices of the two matrices extended with the observability matrices of the corresponding

inverse models. Principal angles between two subspaces are the angles between their principal directions.

Cock *et al.* [20] convert the ARMA model as represented in (6.5) and (6.6) into a forward innovation model:

$$\hat{x}_{t+1} = A\hat{x}_t + Ke_t \tag{6.12}$$

$$y_t = C\hat{x}_t + e_t \tag{6.13}$$

where $K \in \mathbb{R}^n$ is the Kalman gain as described in [65]. The problem of computing the subspace angles between the two models can be transformed into an eigenvalue problem involving the system parameters of forward and inverse innovation models.

In order to estimate the distance between two models, we need certain distance measures based on the computed subspace angles. There are several distance metrics based on subspace angles between ARMA models. The first one is due to Martin [58] and can be written as:

$$d_M(M1, M2)^2 = ln \prod_{i=1}^{n} \frac{1}{cos^2\theta_i} \tag{6.14}$$

where M_1 and M_2 are two ARMA models and θ_i's are the subspace angles between them. Other distance measures include gap and Frobenius norm based distances defined as:

$$d_g(M1, M2) = sin\theta_{max} \quad and \tag{6.15}$$

$$d_f(M_1, M_2)^2 = 2\sum_{i=1}^{n} sin^2\theta_i \tag{6.16}$$

There is another distance described in [90] which is the largest principal angle between the two models. In our experiments, all these metrics give similar recognition

performance.

6.3.5 Experiments, results and discussion

We conducted face recognition experiments using the proposed framework on two datasets. The first one is same as the one used by Li *et al.* [52]. It has face videos for 16 subjects with 2 sequences per subject. In these sequences, the subjects arbitrarily move their heads and change their expressions. The illumination conditions for 2 sequences of each subject were quite different. For each subject, one sequence was put in the gallery while the other formed a probe. A few example images from this dataset are shown in Figure 6.2.

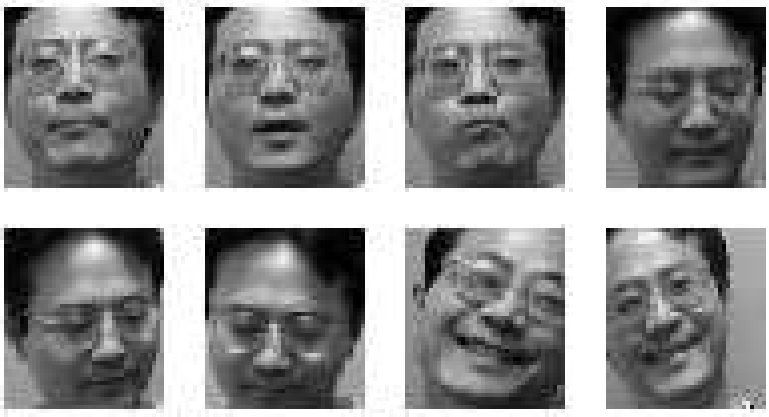

Figure 6.2: Few cropped faces from a video sequence in the first dataset.

The second dataset (obtained from UCSD/Honda) is the one used by Lee *et al.* [50]. With this dataset, we had a gallery of size 15 and probe containing 30 video sequences. In each video, subject moves his/her face in an arbitrary sequence of 2-D and 3-D rotations while changing facial expression and speed. There is even partial

occlusion in a few frames of several video sequences. The illumination conditions vary significantly among the various sequences. Although the datasets used are small, we consider them good tests for our algorithm because of the extreme pose and expression variations and varying illumination as is evident from Figures 6.2 and 6.3.

Figure 6.3: Few cropped faces from a video sequence in the UCSD/Honda dataset.

Our experiment broadly consists of three steps: preprocessing, model estimation and recognition. The preprocessing step involves cropping out the face from each frame of the video sequence. We use a variant of KL tracker [78] to track the nose tip location and an edge-based rough pose estimator. The nose tip location gives an idea about the location of the face while the pose information helps in getting the expanse of the face image relative to the nose. Figures 6.2 and 6.3 show few of the images cropped using this automatic method. Model identification involves estimating A, C and K for each face sequence using the closed form solution explained in Section 6.3.3 while recognition involves computing the principal angles

between probe and gallery models and using them to compute the distances between the models.

With both the data sets, we got recognition performance of more than 90% (15/16 for the first dataset and 27/30 for the second). These numbers are very promising given the extent of pose and expression variations in the video sequences. The results reported in [50] are on per-frame basis and are not directly comparable even though one of the datasets used is the same.

6.3.5.1 Independent evaluation

In [36], Hadid and Pietikainen present an experimental evaluation of integration of facial dynamics in video-based face recognition. The factors like sequence length and image quality are considered in the analysis. The experiments are performed on two face video datasets: MoBo (Motion of Body) [35] and Honda/UCSD [50]. The Mobo dataset consists of 96 face sequences of 24 subjects walking on a treadmill while the considered subset of Honda/UCSD dataset consists of 40 sequences of 20 subjects. The complete details of the experiment settings are described in [36].

Table 6.1 summarizes the performance obtained using spatio-temporal representations (HMM and ARMA) and their static image counterparts (PCA and LDA). All 300 frames present in each probe video are considered in this experiment. As shown in the table, HMM and ARMA-based approaches perform slightly better than PCA and LDA.

In a real application, a subject may not be in front of a camera for such a long

	MoBo Dataset	Honda-UCSD dataset
PCA	87.1%	89.6%
LDA	90.8%	86.5%
HMM	92.3%	91.2%
ARMA	93.4%	90.9%

Table 6.1: Recognition rates using all 300 probe frames using MoBo and UCSD-Honda dataset as reported in [36].

duration. Therefore, Hadid and Pietikainen [36] perform an experiment to evaluate the effect of sequence length on the four approaches. Figure 6.4 and Figure 6.5 show the results obtained in this experiment for the two datasets. As shown in the figures, the proposed ARMA model based approach performs consistently well for a large range of sequence lengths. In another experiment, the authors analyze the effect of image resolution on the recognition rates using Mobo dataset. Table 6.2 shows the results obtained in this experiment. As shown, the proposed ARMA-model based approach compares favorably against other approaches and shows least degradation in performance as image resolution is reduced.

6.4 3D facial pose tracking and recognition

Face tracking is a crucial task for several applications in computer vision. It serves as the first step in several applications like face recognition, lip reading, human

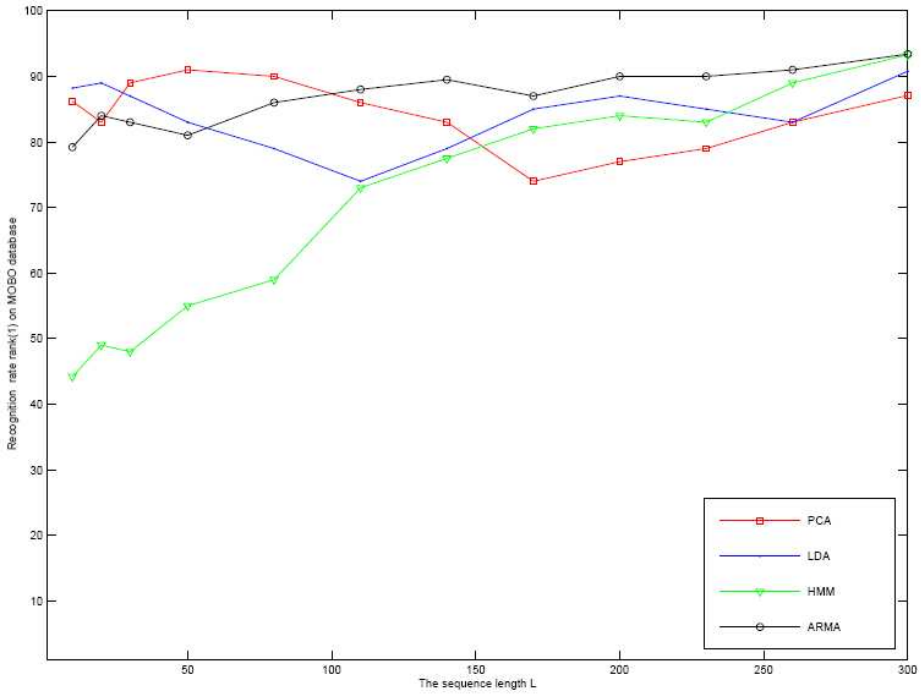

Figure 6.4: Effect of sequence length on recognition rates on MoBo dataset [36].

computer interaction and animation. Most of these applications require that actual 3D parameters of the motion of the head, like the orientation of the head, to be recovered. In this section, we describe an approach for reliable tracking of position and orientation of the face under illumination changes, occlusion and extreme poses. The usefulness of the recovered 3D configuration for the VFR problem is also shown.

6.4.1 Prior work

There has been significant work on facial tracking using 2D appearance based models. [47] [60] [93] use 2D face models based on splines or deformable templates. [100] [37] use affine and planar models, respectively to track a face. Quite clearly, such approaches based on the 2D appearances usually do not explicitly solve the

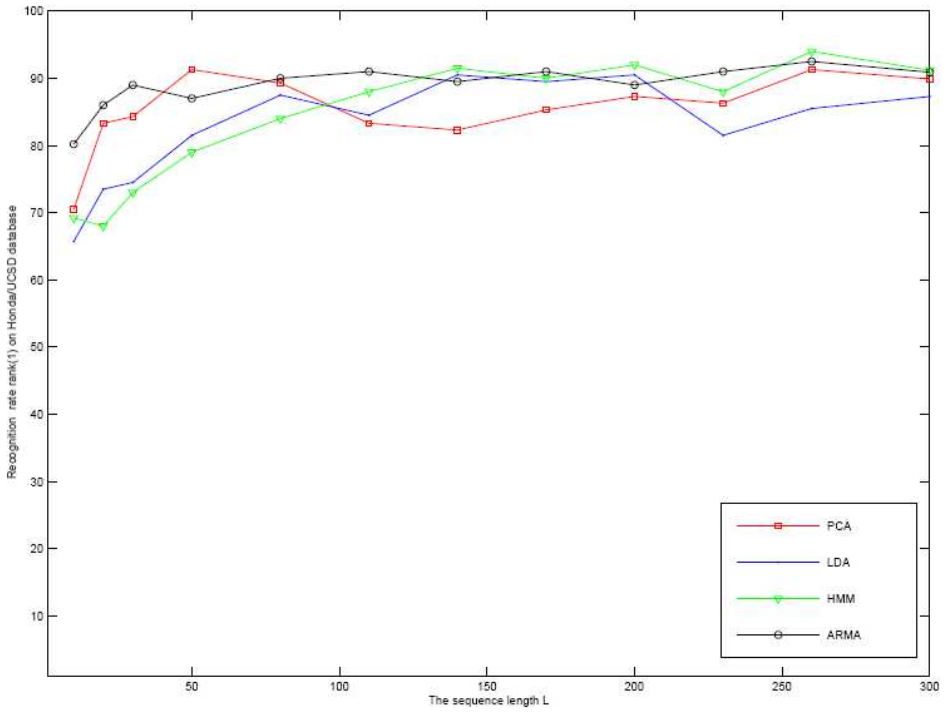

Figure 6.5: Effect of sequence length on recognition rates on Honda-UCSD dataset [36].

correspondence problem. Rather, more often than not they are interested in finding just the image region containing the object (face in this case). Estimation of the 3D orientation of the head is extremely difficult using these approaches. Therefore when such 2D approaches are used as a front-end for tasks such as recognition, multiple view based exemplars [101] are sometimes used in the gallery. While, such a system might improve over the performance of single image based face-recognition systems, such view based exemplars do not capture the structure of the object.

Recently, several methods have been developed for 3D face tracking. [44] uses a closed loop approach that utilizes a structure from motion algorithm to generate

Resolution	40×40	20×20	10×10
PCA	87.1%	81.3%	60.6%
LDA	90.8%	79.5%	56.5%
HMM	92.3%	85.2%	71.2%
ARMA	93.4%	84.1%	74.2%

Table 6.2: Effect of image resolution on recognition rate as reported in [36].

a 3D model of the face. The model is then used to constrain the features in the next frame. The tracking is based on a Kalman filter. In [67], techniques from continuous optimization are applied to a linear combination of 3D face models. They are able to automatically recover the face position and expression for each frame. [55] proposes a hybrid sampling solution using both RANSAC and particle filters to track the pose of a face. Some researchers have proposed using active appearance models for face tracking and/or pose recovery and expression recognition [26][48]. A cylindrical face model for face tracking has been used in [18]. In their formulation, the inter-frame warping function is assumed to be locally linear. In addition, they also assume that the inter-frame pose change occurs only in one of the six degrees of freedom of the rigid cylindrical model. In our approach, we do not have to make any such assumptions. This improves both tracking accuracy and robustness.

6.4.2 The geometric model

The choice of the model to represent the facial structure is very crucial for the problem of face tracking. Several geometric models have been proposed for facial analysis. More often than not, the choice depends on the goal of the analysis. There are several algorithms that do not assume an explicit structural model. They track salient points, features or 2D image patches [63] to recover the 2D or 3D head configuration. On the other extreme, there are algorithms like [44] that use a set of 3D laser-scanned heads represented in a parameterized eigenspace to constrain the structural estimation. A few other focus mainly on 2D tracking (e.g., [101], [9], [23], [30], [37], [93]) which makes a planar model (elliptic, rectangular, etc.) suitable for them.

We would like to restate here that in our work, we aspire to estimate the 3D configuration of the face in each frame. Though a planar model will probably be the simplest one to use, it does not have the capability to handle out-of-plane rotations due to the involved self-occlusions. Moreover the parameters recovered using such a model do not contain information required to estimate the 3D configuration of the face. On the other hand, using a complicated face model (e.g., 3D range data model of an average face), makes the initialization and registration process difficult. In fact, [18] shows experiments where perturbations in the model parameters affect the tracking performance using a complex rigid model (generated by averaging the Cyberware scans of several people), while the simple cylindrical model is robust to such perturbations.

Similar to [18], we use a cylindrical model, though with an elliptical cross-section, to represent a face. The choice of the elliptic cylinder was based on the observation that for most people, the cross section of the head is more elliptic than cylindrical. The choice of the ellipticity does not affect the tracking performance in general but it does make a difference when the face is turned about the vertical axis by a large angle (i.e., high yaw value). Assuming that our cylindrical model reasonably approximates the 3D structure of a face, the problems related to pose and self-occlusion (usually due to pose changes) get automatically taken care of.

From the point of geometrical modeling, the next important issue is the choice of projection. Due to the absence of the calibration parameters, people usually assume orthographic projection. The use of orthographic projection is restrictive and can potentially introduce confusion between scale and pitch. These reasons motivate us to use the perspective projection model. Since we do not know the camera focal length for uncalibrated videos, we show that our approach for pose recovery is fairly robust to the errors in focal length assignment as far as this face tracking is concerned.

Let us assume that the true focal length of the camera imaging a cylinder centered at (X_0, Y_0, Z_0) with height H and radius R be f_0. Let us assume that we erroneously set the focal length to kf_0 (without loss of generality $k \geq 1$). The true projections of feature points on the cylinder are given by

$$x_f = \frac{f_0 X_f}{Z_0 + z_f} \qquad y_f = \frac{f_0 Y_f}{Z_0 + z_f} \qquad \text{where,} \qquad Z_f = Z_0 + z_f \tag{6.17}$$

The projection of feature points of another cylinder with same dimensions but

placed at (X_0, Y_0, kZ_0) and imaged by a camera of focal length kf_0 are

$$\hat{x}_f = \frac{kf_0 X_f}{kZ_0 + z_f} = x_f \left[1 + \frac{(k-1)z_f}{kZ_0 + z_f}\right] = x_f[1 + \delta_f] \tag{6.18}$$

$$\hat{y}_f = \frac{kf_0 Y_f}{kZ_0 + z_f} = y_f \left[1 + \frac{(k-1)z_f}{kZ_0 + z_f}\right] = y_f[1 + \delta_f] \tag{6.19}$$

If $\delta_f \ll 1$, the feature positions for the cylinder at (X_0, Y_0, Z_0) imaged by camera f_0 is equivalent to a cylinder at (X_0, Y_0, kZ_0) imaged by a camera with focal length kf_0. Therefore, when δ_f is small, our estimates of yaw, pitch and roll are reasonably accurate.

If the depth variations in the object (cylinder in our case) are smaller than the distance of the object from the camera center (i.e., $z_f \ll Z_0$) and the field of view is reasonably small, then

$$\delta_f = \frac{(k-1)z_f}{kZ_0 + z_f} \quad < \quad \frac{kz_f}{kZ_0 + z_f} \quad < \quad \frac{\frac{z_f}{Z_0}}{1 + \frac{z_f}{kZ_0}} \quad \ll \quad 1 \tag{6.20}$$

6.4.2.1 Model initialization

The model is initialized using the first frame of the video. Initialization essentially involves finding the parameters for the cylinder (the radius and the height). In the current implementation, we assume that the face is roughly frontal during initialization. We use the optimal edge-based shape detection algorithm [59] to detect the face in the first frame. This algorithm looks for ellipses containing facial features for face detection using the optimal shape operator.

6.4.3 Features

The choice of features is extremely important for the task of 3D pose estimation of a moving face, probably second only to that of the structural model. More than anything else, the features should be easy to detect. In addition, ideally they should be robust to occlusions, and changes in pose, expression and illumination. Humans detect and track faces (known or unknown) effortlessly using features like eyes, nose, mouth, hair etc. For machines, this might not be easy. In a monocular video, only input the machines have is an image which is a 2D projection of the current appearance of the face. The appearance of the features used by humans changes a lot with variations in pose, expression etc. In fact sometimes, few of the features are not even visible in the image. This makes the automatic detection and thereby tracking of these features very difficult.

As stated previously, ours is a hybrid approach which tries to make use of the advantages of a purely geometric approach (useful when partial/complete information about the geometric structure of the object is available) and that of statistical inference. In this work, we stress-test this approach using an extremely simple and easily computable feature. We superimpose a rectangular grid all around the curved surface of our elliptical cylinder. Then mean intensity is computed for each of the visible grids which forms the feature vector. Note that many of the mean values will be undefined which correspond to the the part of the face which is not visible in the frame.

Though quite simple, the feature vector is not all that bad when viewed from

the point of view of our framework. First of all, it is easily computable. Given the current configuration of the face, the grids can be projected onto the image frame and the mean can be computed for each of them. This might seem suspect as the current configuration of the face is not available! Rather that is what we are trying to estimate. Crudely speaking, we first predict the current configuration based on the past configuration and then test its likelihood using the current feature vector. This will become clear once we present the particle-filter framework where each particle represents a configuration of the face. The *mean vector*, by itself, is not invariant to pose but pose is not an issue in our framework as long as the cylindrical assumption is fine. Mean is definitely not invariant to illumination changes. We use robust statistics to make the approach robust to illumination. The fact that the mean is computed for lots of small regions makes it appropriate for robust statistics. The basic idea here is that illumination does not affect the algorithm as long as many of the *means* remain unaffected. The same idea works even for handling partial occlusions and expression changes.

6.4.4 Tracking framework

Once the structural model and the feature vector are fixed, the goal is to estimate the configuration (or pose) of the moving face in each frame of a given video. Breaking this down to each frame, one can see that only information available to perform the desired estimation is the face configurations in the previous frames and the current observation (the current frame). This can be viewed as a dynamic state

estimation problem. Here the state consists of the six configuration parameters: three for the translation and three for the orientation of the face. The Bayesian approach to handle this problem is to gather the available information to come up with the probability density function (pdf) of the state. This estimation can be done recursively for each frame using particle filters.

6.4.4.1 Particle filter

Particle filtering [27][34] is an inference technique for estimating the unknown dynamic state θ of a system from a collection of noisy observations $y_{1:t}$. Quite often, a state space model is used to perform this estimation. The two components of this approach are the state transition model which models the state evolution, and the observation model which specifies the state-observation dependence:

$$\text{State transition model: } \theta_t = f(\theta_{t-1}, u_t), \tag{6.21}$$

$$\text{Observation model: } y_t = g(\theta_t, v_t), \tag{6.22}$$

where u_t is the system noise while v_t is the observation noise. In general, the functions f and g can also be time-dependent. The particle filter approximates the desired posterior pdf $p(\theta_t|y_{1:t})$ by a set of weighted particles $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^N$, where N denotes the number of particles. The state estimate $\hat{\theta}_t$ can be recovered from the pdf as the maximum likelihood (ML) estimate or the minimum mean squared error (MMSE) estimate or any other suitable estimate based on the pdf.

To keep the tracker as generic as possible, we use a simple first order motion model:

$$\theta_t = \theta_{t-1} + u_t, \tag{6.23}$$

where u_t is a Gaussian distribution with zero mean. Based on the domain knowledge, one can come up with a motion model that will be capable of estimating the pdf better with fewer particles. For example, if the task is to track the face of a spectator in a tennis match, a motion model heavily biased towards *yaw* might be a better choice than a generic model.

The observation model involves the feature vector described in the previous section. In our framework, we can rewrite the observation equation as:

$$z_t = \Gamma\{y_t; \theta_t\} = F_t + v_t, \tag{6.24}$$

where y_t is the current frame (the grayscale image), Γ is the mapping that computes the feature vector given an image y_t and a configuration θ_t, z_t is the computed feature vector and F_t is the feature model. The feature model is used to compute the likelihood of the particles (which correspond to different proposed configurations of the face). For each particle the likelihood is computed using the average sum of square differences (SSD) between the feature model and the *mean vector* z_t corresponding to the particle.

On one extreme, the feature model can be a fixed template (say, the feature vector corresponding to the first frame i.e., $F_t = F_0$) while on the other hand one can use a dynamic template e.g, the feature vector belonging to the best particle at the previous frame i.e., $F_t = \hat{z}_{t-1}$. Similar to [45], we refer to the fixed template $F_t = F_0$

as the lost model while the dynamic component $F_t = \hat{z}_{t-1}$ as the wander model. It is worthwhile to note that though the lost component should be credible (assuming initialization is good), quite often it is not capable of handling the appearance changes due to illumination, expression, etc. as the face translates/rotates in the real world. On the other hand, the dynamic nature of the wander component makes it suitable to take care of appearance changes but it is susceptible to drifts. This means that if we have a bad estimation for a frame, it becomes very difficult for the tracker to correct itself in subsequent frames. We use a combination of both which provides resiliency to our tracker. Resiliency is a very important property of any tracker as however good a tracker is, it can always loose track due to an unexpected change in conditions. In the current implementation, the likelihood of a particle is computed as the maximum of the likelihoods using the lost model and the wander model. The prior is biased towards the lost model by 0.52 : 0.48. We take the maximum of the two likelihoods instead of mixing the two to avoid boosting up the probability of a bad particle accidentally. This gives us the capability both to handle appearance changes and to correct the estimation even if the wander model drifts. The number of particles used were typically in the range 200-500.

6.4.4.2 Robust statistics

The performance of the filtering method described is limited by the appropriateness of the likelihood model. If the feature vector or the method of likelihood computation is not good enough to distinguish between different configurations of

the face, tracking can not be expected to be good. Furthermore, we use the mean of grids as the feature vector which by itself is not robust to occlusions, illumination, expression etc. The fact that lots of means are computed over small local regions makes the scenario suitable for the application of robust statistics in the likelihood computation. In the current implementation, we trust only the top half of the means and treat the rest as outliers. The robustified likelihood computation can be represented as:

$$p(y_t|\theta_t^{(j)}) = e^{-\lambda dist} \tag{6.25}$$

$$\text{where,} \quad dist = \frac{\sum_{m,n} \eta(m,n) d(m,n)}{\sum_{m,n} \eta(m,n)} \tag{6.26}$$

where $\eta(m,n)$ is 1 if the $(m,n)^{th}$ grid is visible in both the model and the particle and 0 otherwise, while $d(m,n)$ is computed as:

$$d(m,n) = \begin{cases} (F_t(m,n) - z_t^{(j)}(m,n))^2 & \text{if } d(m,n) < c \\ c & \text{otherwise} \end{cases}$$
$$\text{where,} \quad c = median(\{d(m,n)\}) \tag{6.27}$$

6.4.5 Experiments and results

We conducted three different experiments to show the efficacy of our tracking approach. The experiments are designed to display the ability of the tracker to handle occlusion, expressions and extreme poses. The comparison with the ground truth is also done. In addition, we show how maintaining 3D correspondences help in other problems like recognition.

6.4.5.1 Tracking extreme poses

We conducted tracking experiments on 3 datasets (Honda/UCSD dataset [50], BU dataset [18] and Li dataset [52]). These datasets have numerous sequences in which there are significant illumination changes, expression variation and people are free to move their heads arbitrarily. Figure 6.6 shows few of the frames from several videos with grid points on the estimated cylinder overlaid on the image frame. The first row shows the ability of the tracker to accurately estimate the pose even under extreme poses. Most 2D approaches would not be able to maintain tracks under such severe poses. The second row shows some frames in which the tracker was able to maintain tracks in spite of severe occlusion. The subject waved his hand across his face while simultaneously turning his head. The robust statistics employed in the likelihood computation enables the tracker to maintain track under occlusion. Moderate expressions do not affect our feature since it is the mean intensity within a small surface patch on the face. During certain severe expression changes, robust statistics helps us maintain the track. The third and fourth rows show more tracking results from the BU dataset and the Li dataset, respectively. In the fourth row, the subject is removing his glasses while rotating his head. Our tracker is able to maintain the track in the entire sequence.

6.4.5.2 Ground truth comparison

The BU dataset [18] provides us with ground truth values for the pose of the face in each frame. We conducted tracking experiments on the BU dataset

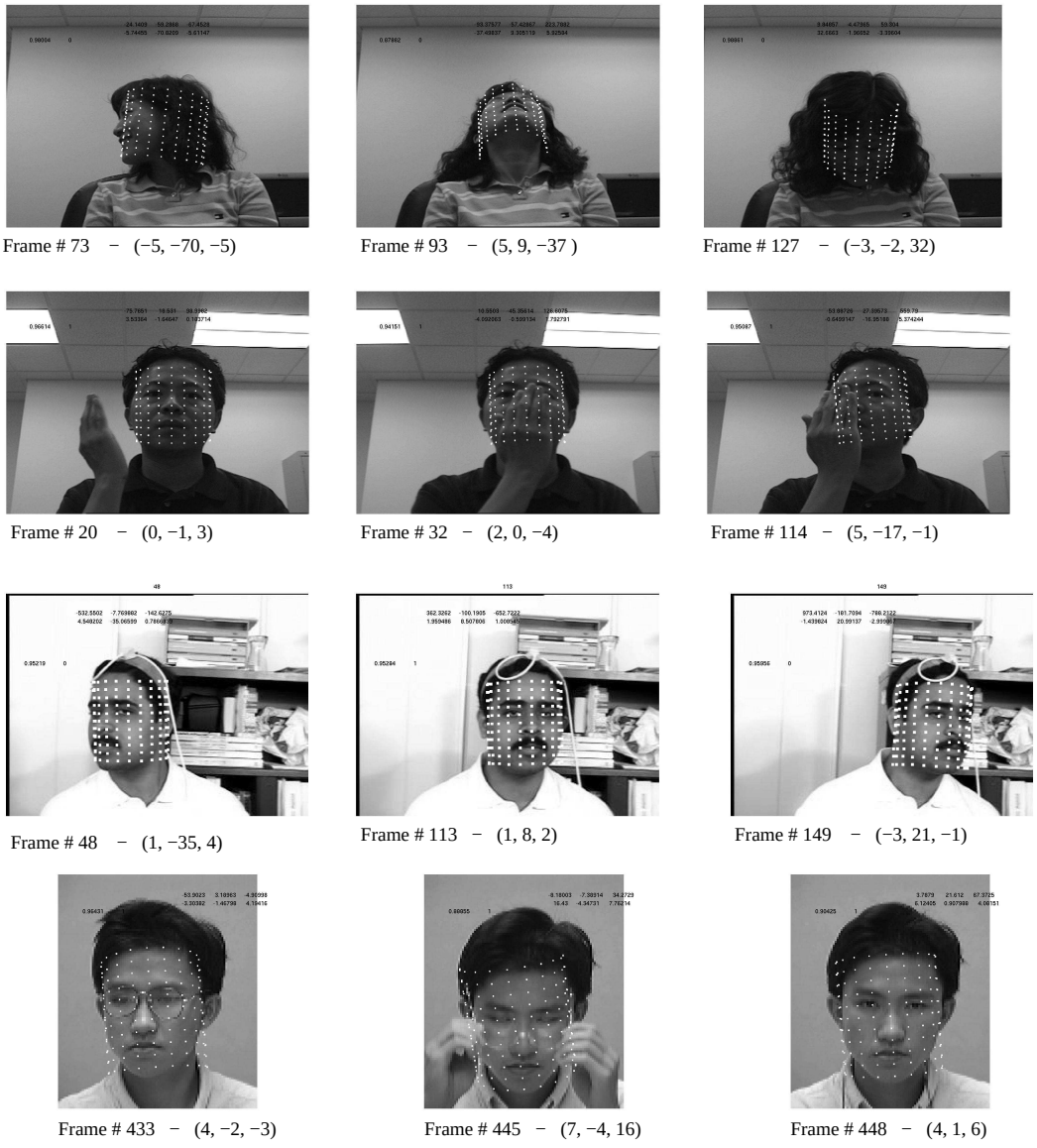

Figure 6.6: Tracking results on different datasets under severe occlusion, extreme poses and different illumination conditions. The cylindrical grid is overlaid on the image plane to display the results. Each frame is labeled with its frame number in the video. The 3-tuple shows the estimated orientation (roll, yaw, pitch) in degrees for each of the frames.

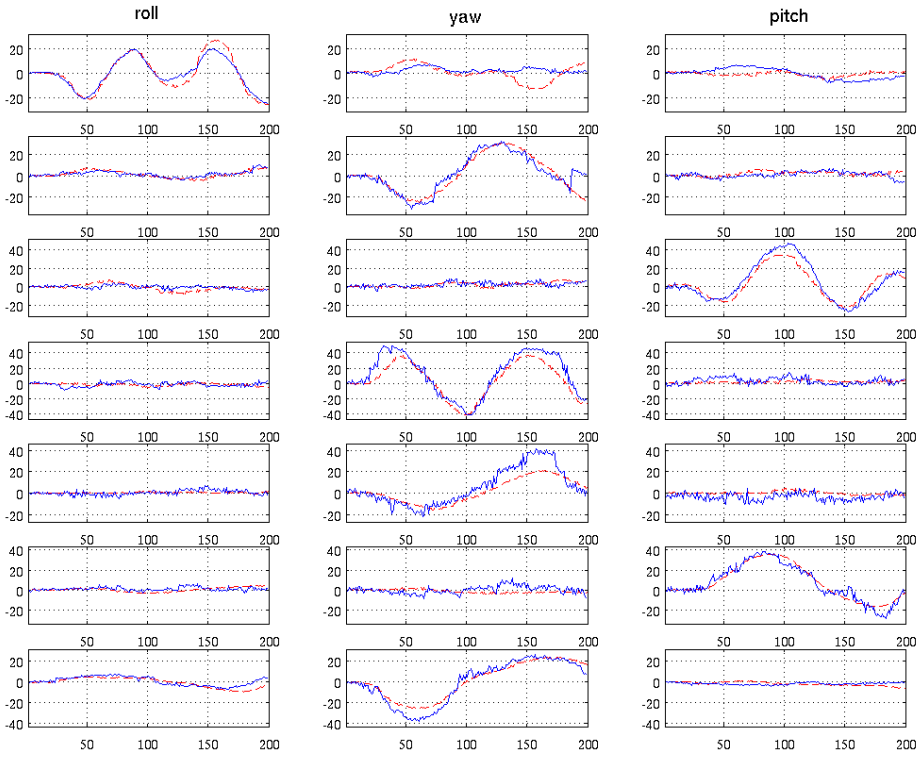

Figure 6.7: Comparison with the ground truth. Each row corresponds to one video displaying the three orientation parameters. The red/dashed curve depicts the ground truth while the blue/solid curve depicts the estimated values.

and compared yaw, pitch and roll estimated by our tracker to the ground truth. Figure 6.7 shows the comparison between the estimated pose of the face and ground truth for six different sequences in the dataset. We see that the tracker accurately estimates the pose of the face in most of these frames.

6.4.5.3 Recognition across non-overlapping poses

Most methods for recognition require that the gallery contains an instance of a face with a pose very similar to the one in the probe. Since our tracking method maintains explicit pose of the face during each frame, we do not need to have the same poses seen in the gallery and the probe. In this experiment we show this by performing recognition on non-overlapping poses. The gallery consists of a video sequence of about 10-15 frames in which the individual turns his head left from about 15 degrees away from frontal to extreme left. The probe consists of a video sequence in which the individual turns his head right from 15 degrees away from frontal all the way to the right. Therefore, there is no pose overlap between the gallery and the probe. In fact, the closet poses in the gallery and the probe differ by at least 30 degrees. We used 10 subjects from the Honda/UCSD dataset [50] for this experiment. For each frame we build a texture mapped cylinder using the tracked pose. We used the minimum sum of squared distance between a gallery model and a probe model as the distance between two videos. This is a very challenging experiment since the poses exhibited by the gallery videos and those exhibited by the probe videos are very different. Therefore, the similarity matrix obtained in this experiment was weakly diagonal. In spite of this, we obtained 100% recognition rate in this experiment, i.e., all the 10 probe videos were recognized correctly. This is very promising and we hope to extend the results to a larger dataset for arbitrary uncontrolled videos of individuals.

6.5 Summary and discussion

In this chapter, we presented two approaches to recognize faces in videos. In particular, we dealt with the problem of tracking and recognizing faces when both gallery and probe consists of face videos.

In the first framework, a moving face is represented as a linear dynamical system (ARMA) whose appearance changes with time. Subspace angles based distance metrics are used to get the measure of similarity between ARMA models representing moving face sequences. The experiments conducted show that the system performs well even in case of extreme 2D and 3D pose variations, expression changes and ordinary illumination conditions.

In the second part of the chapter, we presented a method for tracking facial pose in a video. The tracker is robust to occlusions and illumination changes and maintains track even during extreme poses. We have also shown, how such 3D pose tracking can help in problems like face recognition from videos.

Chapter 7

Summary and Discussion

In this chapter, we summarize the contributions of this dissertation. One of the main objectives of this dissertation is to address the limitations of the existing face recognition algorithms that prevent them from being successful in real systems.

Automatic face recognition is an very unique problem in itself due to the very nature of the problem. Unlike other objects like chair, trees, etc., human face (as a class consisting of all faces in the world) shows relatively small variation across identities. On one hand, this small variation makes it difficult to separate millions of faces from each other, while on the other, it makes it much easier to use or learn generic face-specific properties (geometric, statistical, etc.) to aid in the recognition task. Additionally, in most real scenarios, there are just a few (often just one) samples per identity to perform matching. This makes it quite difficult for systems to recognize faces across variations like illumination, expression, aging, etc.

In this dissertation, we proposed algorithms that take into account these interesting aspects of the problem of automatic face recognition to improve the recognition performance. Specifically, we developed algorithms to perform

- illumination-insensitive matching (Chapter 2 and Chapter 3),
- cohort-based score analysis to improve recognition performance (Chapter 4),
- physics-based revocable face matching (Chapter 5),

- video-based face recognition (Chapter 6), and
- tracking faces in uncalibrated videos (Chapter 6).

Here we provide a brief description of all these algorithms.

In Chapter 2, we modeled face as a linear Lambertian object to perform illumination-insensitive recognition of faces illuminated by single or multiple light sources. Albedo-shape statistics of face as a class are used to address this otherwise ill-posed problem. The low-dimensional linear subspace property of Lambertian reflectance is used to perform this task without any knowledge of number or placement of light sources.

In Chapter 3, we use bilateral symmetry of faces to perform illumination-invariant matching. In particular, we show that illumination-invariant matching is much more tractable for the class of bilaterally symmetric objects. The theoretical analysis leads to an extremely simple illumination-invariant signature of face images used to perform matching. The algorithm is shown to be flawless (modulo a rare condition described in the chapter) if the symmetry and Lambertian assumptions are satisfied.

As mentioned earlier, automatic face recognition presents a very unique challenge from pattern classification point of view due to the presence of just one or a few images per identity. In Chapter 4, we address this issue by making use of the large number of non-match samples to normalize for the differing class distributions thereby improving matching performance

The advancement of biometric in real world applications has led to the concerns

of biometric theft. In Chapter 5, we address these concerns by developing a physics-based approach to generate secure, cancelable and photometrically valid face images to perform matching.

In Chapter 6, we proposed algorithms to perform tracking and recognition of faces in videos. We follow a system identification approach and model moving face using ARMA model. As shown by independent evaluations, the proposed ARMA-based approach compares favorably against other video-based approaches. For tracking, we model face as a cylinder and use particle filter-based inference to recover 3D configuration of face in each frame of the video. The recovered parameters are used to normalize for pose allowing us to recognize faces in video without the need of any pose overlap.

Though the algorithms proposed in the dissertation successfully address various issues, there is still a long way to go before we have commercial face recognition systems which have the generalization capabilities of humans.

Bibliography

[1] G. Aggarwal and R. Chellappa. Face recognition in the presence of multiple illumination sources. In *International Conference on Computer Vision (ICCV)*, pages 1169–1176, 2005.

[2] M. M. Ali. Analysis of stationary spatial-temporal processes: Estimation and prediction. *Biometrika*, 66:513–518, 1979.

[3] H. C. Andrews and B. R. Hunt. *Digital Image Restoration*. Prentice-Hall signal processing series, 1977.

[4] A. Ariyaeeinia and P. Sivakumaran. Analysis and comparison of score normalization methods for text-dependent speaker verification. In *Proceedings of Eurospeech*, pages 1379–1382, 1997.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification system. *Digital Signal Processing*, 10:42–54, 2000.

[6] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381, 2001.

[7] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:218–233, 2003.

[8] 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*, 19:711–720, 1997.

[9] S. Birchfield. An elliptical head tracker. In *Proceedings of the 31st Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California*, pages 1710–1714, November 1997.

[10] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2001.

[11] S. Biswas, G. Aggarwal, and R. Chellappa. Robust estimation of albedo for illumination-invariant matching and shape recovery. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1–8, October 2007.

[12] Z. Biuk and S. Loncaric. Face recognition from multi-pose image sequence. In *Proc. of 2nd International Symposium on Image and Signal Processing and Analysis*, Pula, Croatia, 2001.

[13] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.

[14] R. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. *Guide to Biometrics*. Springer Verlag, 2003.

[15] T. Boult. Robust distance measures for face-recognition supporting revocable biometric tokens. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, volume 3, pages 560–566, 2006.

[16] M. J. Brooks, W. Chojnacki, and R. Kozera. Impossible and ambiguous shading patterns. *International Journal of Computer Vision*, 7(2):119–126, 1992.

[17] H. Bulthoff, S. Edelman, and M. Tarr. How are three-dimensional objects represented in the brain? *MIT AI Memo #1479*.

[18] M. L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(4):322–336, April 2000.

[19] H. Chen, P. Belhumeur, and D. Jacobs. In search of illumination invariants. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–261, 2000.

[20] K. D. Cock and D. B. Moor. Subspace angles and distances between ARMA models. In *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[21] J. Colombi, D. Ruck, T. Anderson, S. Rogers, and M. Oxley. Cohort selection and word grammar effects for speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 85–88, 1996.

[22] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press.

[23] J. L. Crowley and F. Berard. Multi-modal tracking of faces for video communication. In *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR*, 1997.

[24] L.-M. W. H.-J. Deng, Hao-Jiang; Du. Likelihood score normalization and its application in text-independent speaker verification. *Dianzi Yu Xinxi Xuebao (J. Electron. Inf. Technol.)*, 27(7):1025–1029, July 2005.

[25] G. Doretto and S. Soatto. Editable dynamic textures. In *ACM SIGGRAPH Sketches and Applications*, 2002.

[26] F. Dornaika and J. Ahlberg. Fast and reliable active appearance model search for 3D face tracking. *IEEE Transactions on Systems, Man and Cybernetics–Part B: Cybernetics*, 34(4):1838–1853, August 2004.

[27] A. Doucet, N. D. Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice*. Springer-Verlag, New York, 2001.

[28] R. Dovgard and R. Basri. Statistical symmetric shape from shading for 3d structure recovery of faces. In *European Conference on Computer Vision*, 2004.

[29] H. Ekenel and A. Pnevmatikakis. Video-based face recognition evaluation in the chil project - run 1. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 85–90, 2006.

[30] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR*, 1997.

[31] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target dependent score normalization techniques and their application to signature verification. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 35(3):418– 425, August 2005.

[32] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading problem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10:439–451, July 1988.

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

[34] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings on Radar and Signal Processing*, volume 140, pages 107–113, 1993.

[35] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[36] A. Hadid and M. Pietikainen. An experimental investigation about the integration of facial dynamics in video-based face recognition. *Electronic Letters on Computer Vision and Image Analysis*, 5(1):1–13, March 2005.

[37] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, October 1998.

[38] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of Optical Society of America, A*, 11, 1994.

[39] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1(2):89–106, 1991.

[40] B. K. P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3(4):277–299, 1974.

[41] B. K. P. Horn and M. J. Brooks. *Shape from Shading*. MIT Press, Cambridge, Massachusetts, 1989.

[42] D. Jacobs, P. Belhumeur, and R. Basri. Comparing images under variable illumination. In *IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA*, pages 610–617, 1998.

[43] A. K. Jain, U. Uludag, and A. Ross. Biometric template selection: A case study in fingerprints. In *Proceedings of Audio- and Video-Based Biometric Person Authentication*, pages 335–342, 2003.

[44] T. S. Jebara and A. Pentland. Parameterized structure from motion for 3D adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR*, 1997.

[45] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, October 2003.

[46] X. Jiang and W. Ser. Online fingerprint template improvement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1121–1126, August 2002.

[47] A. Lanitis, C. Taylor, and T. Cootes. A unified approach for coding and interpreting face images. In *International Conference on Computer Vision, Cambridge, MA*, pages 368–373, 1995.

[48] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):743–756, July 1997.

[49] K. C. Lee, J. Ho, and D. Kriegman. Nine points of light: acquiring subspaces for face recognition under variable lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–526, December 2001.

[50] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[51] K. C. Lee and B. Moghaddam. A practical face relighting method for directional lighting normalization. In *International Workshop on Analysis and Modeling of Faces and Gestures*, 2005.

[52] B. Li and R. Chellappa. Face verification through tracking facial features. *Journal of the Optical Society of America A*, 18:2969–2981, December 2001.

[53] K. P. Li and J. E. Porter. Normalization and selection of speech segments for speaker recognition scoring. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 595–598, 1988.

[54] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2003.

[55] L. Lu, X. Dai, and G. Hager. A particle filter without dynamics for robust 3D face tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, Washington, D.C.*, 2004.

[56] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. FVC2002: Second fingerprint verification competition. In *Proceedings of the 16th International Conference on Pattern Recognition (3)*, pages 811–814, 2002.

[57] J. Mariethoz and S. Bengio. A unified framework for score normalization techniques applied to text independent speaker verification. *IEEE Signal Processing Letters*, 12(7):532–535, 2005.

[58] R. J. Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48:1164–1170, 2000.

[59] H. Moon, R. Chellappa, and A. Rosenfeld. Optimal edge-based shape detection. *IEEE Trans. on Image Processing*, 11(11):1209–1226, 2002.

[60] Y. Moses, D. Reynard, and A. Blake. Robust real time tracking and classification of facial expressions. In *International Conference on Computer Vision, Cambridge, MA*, pages 296–301, 1995.

[61] J. Navratil, U. V. Chaudhari, and G. N. Ramaswamy. Speaker verification using target and background dependent linear transforms and multi-system fusion. In *Proceedings of EUROSPEECH-01*, September 2001.

[62] J. Navratil and G. N. Ramaswamy. The awe and mystery of t-norm. In *European Conference on Speech Communication and Technology*, pages 2009–2012, 2003.

[63] N. Oliver, A. Pentland, and F. Berard. Lafter: Lips and face real time tracker. In *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR*, 1997.

[64] A. J. O'Toole, A. Roark, and A. H. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, 6:261–266, 2002.

[65] P. V. Overschee and B. D. Moor. Subspace algorithms for the stochastic identification problems. *Automatica*, 29:649–660, 1993.

[66] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. 2005.

[67] F. Pighin, R. Szeliski, and H. Salesin. Resynthesizing facial animation through 3D model-based tracking. In *Seventh International Conference on Computer Vision, Kerkyra, Greece*, pages 143–150, 1999.

[68] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of convex Lambertian object. *Journal of the Optical Society of America A*, pages 2448–2459, October 2001.

[69] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle. Generating cancelable fingerprint templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):561–572, 2007.

[70] N. K. Ratha, K. Karu, S. Chen, and A. K. Jain. A real-time matching system for large fingerprint databases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):799–813, August 1996.

[71] D. Reynolds. Comparison of background normalization methods for text-independent speaker verification. *Eurospeech*, pages 963–966, 1997.

[72] A. Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong. The use of cohort normalized scores for speaker recognition. In *Proceedings of ICSLP*, pages 599–602, September 1992.

[73] C. Ryu, Y. Han, and H. Kim. Super-template generation using successive bayesian estimation for fingerprint enrollment. *AVBPA 2005, Lecture Notes in Computer Science*, 3546:710–719, 2005.

[74] M. Savvides, B. V. K. Vijayakumar, and P. K. Khosla. Cancelable biometrics filters for face recognition. In *Proceedings of International Conference on Pattern Recognition*, volume 3, pages 922–925, 2004.

[75] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. of European Conference on Computer Vision*, Kopenhagen, Denmark, 2002.

[76] A. Shashua. On photometric issues in 3d visual recognition from a single 2d image. *International Journal of Computer Vision*, 21:99–122, 1997.

[77] A. Shashua and T. R. Raviv. The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(2):129–139, February 2001.

[78] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[79] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, December 2003.

[80] D. Simon-Zorita, J. Ortega-Garcia, M. Sanchez-Asenjo, and J. Gonzalez-Rodriguez. Facing position variability in minutiae-based fingerprint verification through multiple references and score normalization techniques. *AVBPA 2003, Lecture Notes in Computer Science*, 2688:214–223, 2003.

[81] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *Proc. of Intl. Conf. on Computer Vision*, 2001.

[82] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20:39–51, 1997.

[83] SVM and K. M. M. Toolbox. http://asi.insa-rouen.fr/ arakotom/toolbox/index.html.

[84] A. B. J. Teoh, A. Goh, and D. C. L. Ngo. Random multispace quantization as an analytic mechanism for biohashing of biometric and random identity inputs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1892–1901, 2006.

[85] S. Tulyakov and V. Govindaraju. Combining matching scores in identification model. In *ICDAR*, pages 1151–1155, 2005.

[86] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:72–86, 1991.

[87] U. Uludag, A. Ross, and A. Jain. Biometric template selection and update: a case study in fingerprints. *Pattern Recognition*, 37(7):1533–1542, 2004.

[88] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:733–742, 1997.

[89] C. Watson and M. Garris. Nist fingerprint image software 2 (nfis2) http://fingerprint.nist.gov/nfis/index.html.

[90] A. Weinstein. Almost invariant submanifolds for compact group actions. *Berkeley CPAM Preprint Series*, 1999.

[91] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2003.

[92] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[93] A. L. Yuille, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *International Conference on Pattern Recognition*, 1994.

[94] A. L. Yuille, D. Snow, R. Epstein, and P. N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *International Journal of Computer Vision*, 35(3):203–222, 1999.

[95] L. Zhang and D. Samaras. Face recognition under variable lighting using harmonic image exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19–25, 2003.

[96] W. Zhao and R. Chellappa. Illumination-insensitive face recognition using symmetric shape-from-shading. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1286–1293, 2000.

[97] W. Zhao and R. Chellappa. Symmetric shape from shading using self-ratio image. *International Journal of Computer Vision*, 45(1):55–75, October 2001.

[98] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[99] S. Zhou, R. Chellappa, and D. Jacobs. Characterization of human faces under illumination variations using rank, integrability, and symmetry constraints. In *European Conference on Computer Vision*, 2004.

[100] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 11:1434–1456, November 2004.

[101] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding (CVIU) (special issue on Face Recognition)*, 91:214–245, 2003.