Alma Mater Studiorum – Università di Bologna
in cotutela con Università di Sassari

DOTTORATO DI RICERCA IN

INGEGNERIA ELETTRONICA, INFORMATICA E DELLE TELECOMUNICAZIONI

Ciclo XXVI

Settore Concorsuale di afferenza: 09/H1

Settore Scientifico disciplinare: ING-INF/05

TITOLO TESI

ADVANCED TECHNIQUES FOR FACE RECOGNITION UNDER CHALLENGING ENVIRONMENTS

Presentata da: **YUNLIAN SUN**

Coordinatore Dottorato

ALESSANDRO VANELLI-CORALLI

Relatore

DAVIDE MALTONI

Relatore

MASSIMO TISTARELLI

Esame finale anno 2014

Advanced Techniques for Face Recognition under Challenging Environments

Yunlian Sun

DISI - Dpt. of Computer Science and Engineering

University of Bologna

A thesis submitted for the degree of

Doctor of Philosophy

2014

I would like to dedicate this thesis to my loving family.

Acknowledgements

My deepest gratitude goes first and foremost to my supervisors Prof. Massimo Tistarelli and Prof. Davide Maltoni for their consistent, illuminating and expert instruction, walking me through all the stages of my PhD study, providing me every chance to attend international conferences and summer schools and creating various opportunities to visit different research groups working in biometrics area.

I would like to thank Prof. Enrico Grosso and Prof. Dario Maio for their invaluable guidance and offering inspiring research environment, resources for undertaking research activities.

High tribute shall be paid to Prof. Yong Xu, Harbin Institute of Technology Shenzhen Graduate School, for offering me valuable suggestions in the academic research and always sharing novel ideas, new research directions with me during my PhD study. He was the supervisor who led me into the world of biometrics and also motivated and helped me to go further in this area.

I am also greatly indebted to Dr. Norman Poh, University of Surrey, for his co-guidance, providing me resources and suggestions for undertaking research activities, guiding me writing research articles, and always sharing with me different research directions and novel ideas.

Many thanks to Dr. Matteo Ferrara and Dr. Annalisa Franco for their co-guidance and helping me for formulating research problems, writing research papers and engaging in discussions.

I deeply appreciate Prof. Rama Chellappa, University of Maryland, College Park and Dr. Arun Ross, Michigan State University, for serving as external reviewers of my PhD thesis and providing me with valuable review comments. Sincere gratitude is also expressed to Prof. Alessandro Vanelli-Coralli, the PhD coordinator, and Ms. Francesca Chiarini, the PhD secretary, for their kindness, patience and so many helps during my PhD study.

I would like to express my gratitude to Prof. Tieniu Tan, Institute of Automation of the Chinese Academy of Sciences, for giving me the chance to visit and work in his group. Sound research environment, timely meetings and enlightening lectures on the related research topics have been proved very beneficial. I would also take the opportunity to thank Dr. Zhenan Sun and other group members for their help and time in listening to me and helping me work out my problems during my stay at Beijing.

I am very grateful to Prof. Chang-Tsun Li, University of Warwick, for giving me the opportunity to visit and work in his group. The period spent in his group has been very productive for my ability to do research. I would also convey my regards to Xingjie Wei, Yu Guan and other group members for their support and suggestions during my stay at Conventry.

Many thanks to Massimo Gessa, Matteo Sanna, Dr. Elif Surer, Alper Kose, Dr. Andrea Lagorio, Dr. Luca Pulina, Laura Pandolfo, Dr. Marinella Cadoni, Dr. Ludovica Lorusso, Dr. Ajita Rattani, Daksha Yadav, Fulvio Stradijot, Fabrizio Chelo and Dr. Stefania Mura for providing timely help, suggestions and discussion during my stay at Sassari. I also owe my sincere gratitude to Dr. Raffaele Cappelli, Dr. Francesco Turroni, Luca Calderoni for their assistance during my stay at Cesena.

Last my thanks would go to my beloved parents, sister and brother for their loving consideration, constant support and great confidence in me all through these years. Whatever progress I have made or I shall make in this life, I dedicate to them.

Abstract

Automatically recognizing faces captured under uncontrolled environments has always been a challenging topic in the past decades. In this work, we investigate cohort score normalization that has been widely used in biometric verification as means to improve the robustness of face recognition under challenging environments. In particular, we introduce cohort score normalization into undersampled face recognition problem. Further, we develop an effective cohort normalization method specifically for the unconstrained face pair matching problem. Extensive experiments conducted on several well known face databases demonstrate the effectiveness of cohort normalization on these challenging scenarios. In addition, to give a proper understanding of cohort behavior, we study the impact of the number and quality of cohort samples on the normalization performance. The experimental results show that bigger cohort set size gives more stable and often better results to a point before the performance saturates. And cohort samples with different quality indeed produce different cohort normalization performance.

Recognizing faces gone after alterations is another challenging problem for current face recognition algorithms. Face image alterations can be roughly classified into two categories: unintentional (e.g., geometrics transformations introduced by the acquisition devide) and intentional alterations (e.g., plastic surgery). We study the impact of these alterations on face recognition accuracy. Our results show that state-of-the-art algorithms are able to overcome limited digital alterations but are sensitive to more relevant modifications. Further, we develop two useful descriptors for detecting those alterations which can significantly affect the recognition performance. In the end, we propose to use the Structural Similarity (SSIM) quality map to detect and model variations due to plastic surgeries. Extensive experiments conducted on a plastic surgery face database demonstrate the potential of SSIM map for matching face images after surgeries.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The access to restricted systems has mostly been controlled by knowledge-based (pin, password) or token-based (ID cards) security. However, these traditional identity management techniques can easily fail when a password is stolen or the card is lost. The technologies of biometric recognition are highly desired to address these problems. Biometrics use the physiological or behavioural characteristics to recognize the identity of an individual. These characteristics are something you possess rather than something you know thus offering a natural, reliable and user-friendly solution to identity management [33, 45]. Face recognition is one of the most widely used biometric systems due to its non-intrusive, natural and easy to use characteristics. Many advanced technologies and commercial systems for face recognition have been developed. These have been summarised in [100].

1.1 Face Recognition Systems

A face recognition system can be either a verification system or an identification system depending on different applications. In the verification system, there is a pre-enrolled template set. Given a query image, the goal is to decide whether the query image comes from the same individual represented by the claimed target template. It performs a one to one comparison to determine whether the person presenting herself/himself to the system is the person she/he claims to be.

Similarly, in an identification system, there is a stored database of faces. Given one probe face image, the identification task aims to decide which person in the stored gallery face set the probe face represents. It involves a one to many search. The system will either make a match and subsequently identify the person or it will fail to make a match.

Both verification and identification consist of two stages: enrollment and matching. The enrollment stage is used to build the user template database. In this stage, one or more than one face photos of a person is captured by a camera to produce an original digital representation. This original representation is then further processed, including detecting the face region, aligning faces and eliminating the impact of different illumination conditions. After these preprocessing procedures, a feature extractor can be employed to generate a set of distinguishable features. These extracted features are then used to build a template or reference model for the person. After storing the template in the template database, we finally register the person into the system. There is one template for each enrolled subject in the template database of the system. An illustration of the enrollment stage is shown in Figure 1.1. The matching stage of a verification system is a one to one match. In this stage, a subject presents itself to the verification system and claims an identity. The system captures facial images of the subject, preprocesses the original digital representation and extracts robust features. Next, the system compares the extracted features with the template of the claimed identity stored in the template database. With this one to one match, the system can finally decide whether the person is a genuine client (i.e., the true identity is the claimed identity) or an impostor client (i.e., the true identity is not the claimed identity). Figure 1.2 shows the matching stage of a verification system. Differing from the verification task, in the identification scenario, the matching is a one to many procedure, which compares the facial features of the client with templates of all the enrolled subjects in the system database. The result is either an enrolled user's identity or a warning message, e.g., "person not identified". An example of the matching stage of an identification system is illustrated in Figure 1.3.

Here, it is necessary to state the terminology used in the two face recognition tasks. For face verification, "query image" is used to represent the test image,

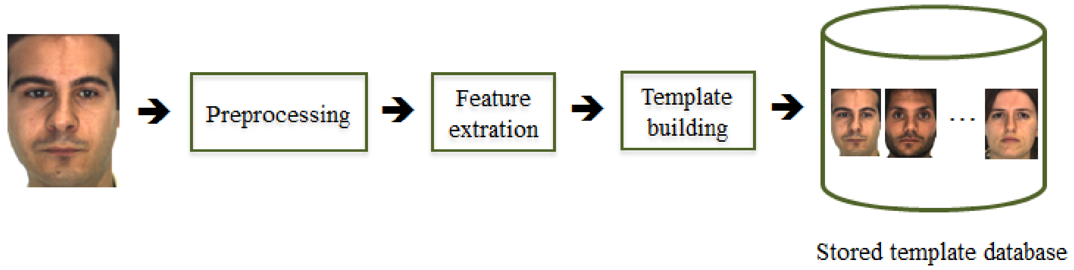


Figure 1.1: The enrollment stage of a face recognition system.

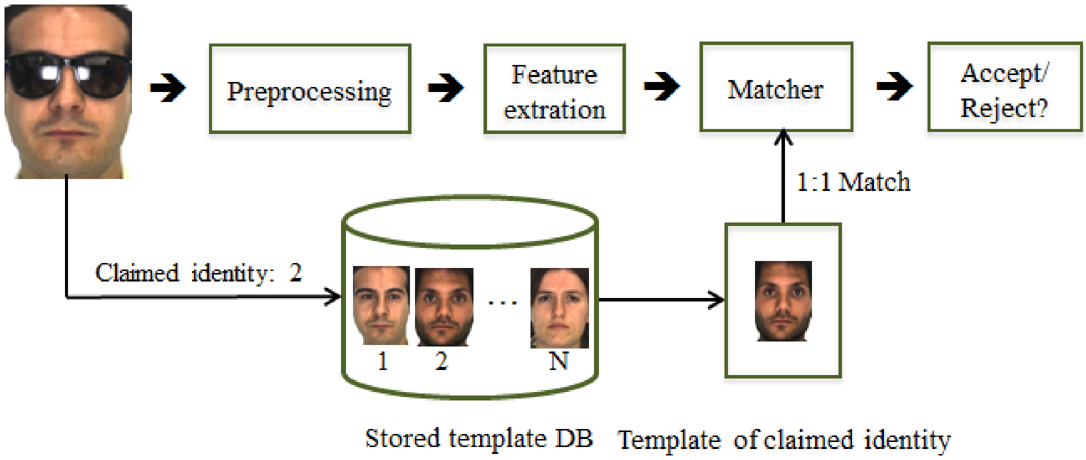


Figure 1.2: The verification stage of a face recognition system.

while in face identification, "probe image" is the more common term. In addition, in an identification system, "gallery set" is the usual term for the enrolled template database.

Very recently, face pair matching, as a very new task in face recognition community, has received a lot of attention [30]. Different from the traditional verification and identification tasks, there is not a pre-enrolled template database, only two face images are given in this task. The aim is to determine whether the two face images represent the same individual, i.e., whether a given image pair is a matching pair or a non-matching pair. Figure 1.4 illustrates this framework. Note that in face pair matching, we do not have any identity information for each face image in the pair. This makes this task a very difficult problem.

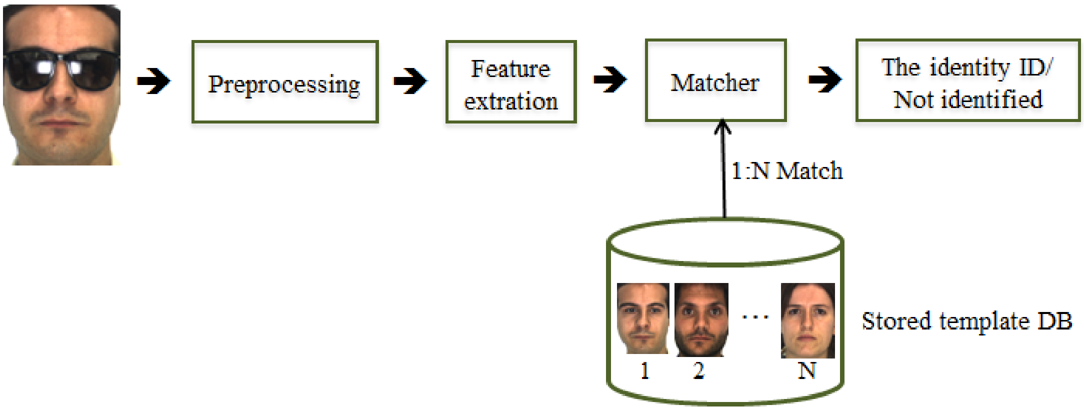


Figure 1.3: The identification stage of a face recognition system.

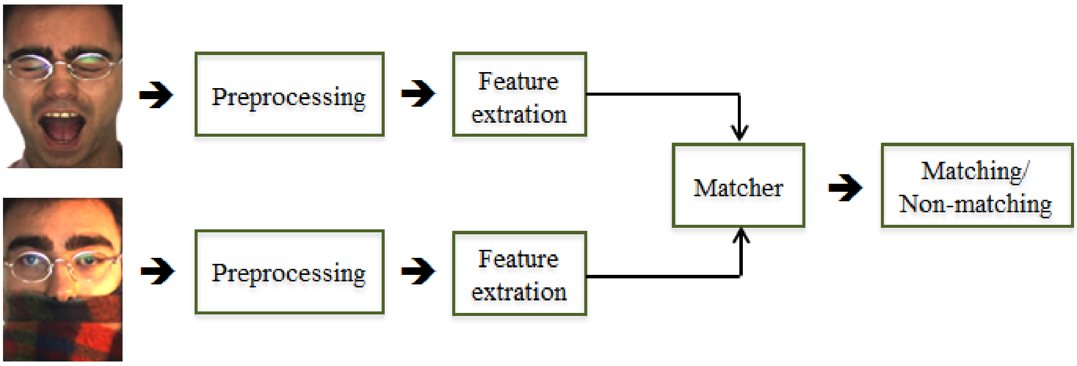


Figure 1.4: The framework of face pair matching.

1.2 Challenging Face Recognition Problems

Face recognition is an important part of the capability of human perception system. Human can identify familiar faces even under severely degraded viewing conditions, such as various illumination conditions, poses, expressions and occlusions. However, a similar automated computer system is not yet able to achieve comparable results due to the large variability presented in the facial appearance of the same person, i.e., intra-personal variations. These variations are easily introduced by the change of the acquisition environment, the manner an individual interacts with the face acquisition device. Automatically recognizing faces becomes more difficult when the intra-personal variations are larger than the variations caused by changes in the face identity, i.e., the inter-personal variations. Zhao et al. [100] provided a critical review of still and video based face recognition systems that have been built over the years and discussed extensively the challenges of face recognition caused by these external factors.

Face recognition across aging poses another challenge for current face recognition techniques [74]. Unlike the above variations caused by the external factors, changes due to this natural phenomenon typically depend on quite a few factors like race, geographical location, eating habits, stress level, etc. Hence, there is no simple geometric/statistical model to analyze appearance changes due to aging, thus making the problem of matching faces across aging extremely difficult.

Plastic surgery becomes worldwide nowadays due to the advanced surgical technologies and the affordable cost. These surgical procedures can modify both the shape and texture of facial features to varying degrees, thus introducing new challenges in designing future face recognition systems. In [78], face recognition across plastic surgery was first introduced to the biometric community. In the related work, various existing algorithms were tested on a database including face images before and after surgeries. The significant performance degradation concluded that the current state-of-the-art face recognition algorithms cannot provide satisfied performance for matching faces gone after plastic surgeries.

Another challenge is the limited number of sample images provided for each person [81]. In many face recognition applications, due to the difficulties of gathering face images and the cost for storing and processing them, only very few or even single training samples are available for building the subject specific template. This problem is also known as undersampled face recognition [23] and directly related to the small sample size problem in statistics and pattern recognition. As is known, many developed face recognition algorithms depend greatly on the size of the training set and thus on the descriptive capability of the training set. In such situations, these face recognition techniques will suffer serious performance drop or even fail to work.

Recognizing faces in the wild is a very new challenge to the current face recognition algorithms [30]. Face images in this scenario are drawn from a very broad distribution, which may present a large range of the variability seen in everyday life. This covers variability in expression, background, pose, ethnicity, age, gender, clothing, camera quality, lighting, makeup and other parameters. Matching face images collected from news articles in the web is such an example, where face images present a large degree of diversity.

1.3 Thesis Contribution

The contributions of this thesis to the methodology of face recognition are summarised as follows:

The thesis investigates cohort score normalization as means to improve the robustness of face recognition under challenging environments. Cohort score normalization, as a way of post-processing the raw matching score using information from a set of cohort samples, i.e., non-matching samples of the subjects being compared, has been widely used in biometric verification systems [8]. In this work, we introduce cohort normalization to sparse presentation based undersampled face identification and unconstrained face pair matching. Experiments performed on the well known face databases show the effectiveness of cohort score normalization on measuring the degradation effect under challenging environments.

Furthermore, we advance the state-of-the-art in cohort score normalization by providing a better understanding of the cohort behavior. In particular, we find that the size of the cohort set indeed has impact on the generalization performance. That is bigger cohort set size usually gives more stable and often better results to a point before the performance saturates. In addition, we conduct an extensive experimental exploration on the FRGC ver2.0 database [70] to give proper answers of two interesting questions about cohort behavior on face recognition. The first one is "How does the quality of cohort samples affect the performance of a face recognition system?", while the second one is "Should we use a cohort set containing most possible subjects(each subject with fewest possible samples as possible) or a cohort set including fewest possible subjects (each subject with most possible samples)?".

Face recognition across alteration is a very challenging problem for current face recognition algorithms. Alterations can be produced both unintentionally (e.g., by the acquisition or printing device) or intentionally (e.g., people modify the digital images or go after plastic surgery to appear more attractive). In this thesis, we study the impact of unintentional alterations and digital beautification of face photos on face recognition accuracy. Furthermore, we develop two facial descriptors specially used for detecting those alterations which can make significant degradation on the face recognition performance. Finally, we present an effective algorithm for face recognition across plastic surgery. The proposed approach uses the Structural Similarity (SSIM) quality map to detect and model variations caused by surgeries [89]. An important feature of the proposed approach, also achieving performance comparable with the current state-of-the-art, is that neither training process is needed nor any background information (i.e., cohort information) from other datasets is required. Extensive experiments conducted on a plastic surgery face database demonstrate the potential of SSIM map for matching face images after surgeries.

1.4 Overview of Thesis

The outline of the thesis is described below.

Overview of Face Recognition. Chapter 2 provides an overview of the structure of a generic face recognition system. Firstly, we present the basic processing stages involved in a face recognition system, including face detection and alignment, photometric normalisation, feature extraction and classification. Secondly, we briefly describe several well known face databases with their common protocols used in our experiments.

Cohort Score Normalization for Face Recognition. In Chapter 3, we first give a brief introduction about cohort score normalization, including the basic concepts used in this technique, the role of cohort normalization in a biometric system and existing approaches for performing cohort normalization. Secondly, the framework of applying cohort normalization to sparse representation based undersampled face recognition is detailed. A powerful cohort normalization method, called picture-specific cohort normalization is also described in this section. Finally, we provide a better understanding of the cohort behavior, including the impact brought about the size of the cohort set, the quality of cohort samples and the structure of the cohort set.

Face Recognition across Alteration. In Chapter 4, we introduce two categories of image alterations: unintentional and intentional alterations. Further, we present a study of the effects of unintentional alterations and digital beautification on face recognition accuracy. Next, we develop two useful solutions for detecting those alterations which can produce a significant performance degradation on face recognition. Finally, we introduce a new method using structural similarity based image quality map for recognizing faces after plastic surgeries.

Conclusions and Future Work. The thesis is drawn to conclusion in Chapter 5 where the directions of future work are also suggested.

Chapter 2

Overview Of Face Recognition

Automated face recognition system is an application of pattern recognition. A block diagram of a generic face recognition system is presented in Figure 2.1. The automated face recognition problem involves four key steps: face acquisition, preprocessing, feature extraction and classification. In the preprocessing stage, the main procedures include face detection, geometric normalization and illumination normalization. In Section 2.1, we provide a highlighted summary of research on these face recognition stages.

Face recognition systems are very difficult to be compared to each other due to the fact that the testing must be performed on a large number of samples in diverse conditions representing realistic scenarios, such as variations in different model database sizes, sensors used, viewing conditions, illumination and background conditions. Therefore, large-scale public databases with well defined protocols are required to achieve these objectives. Section 2.2 introduces the performance measures and several well known face databases for evaluating, characterising and benchmarking the face recognition methods developed and investigated in this thesis.

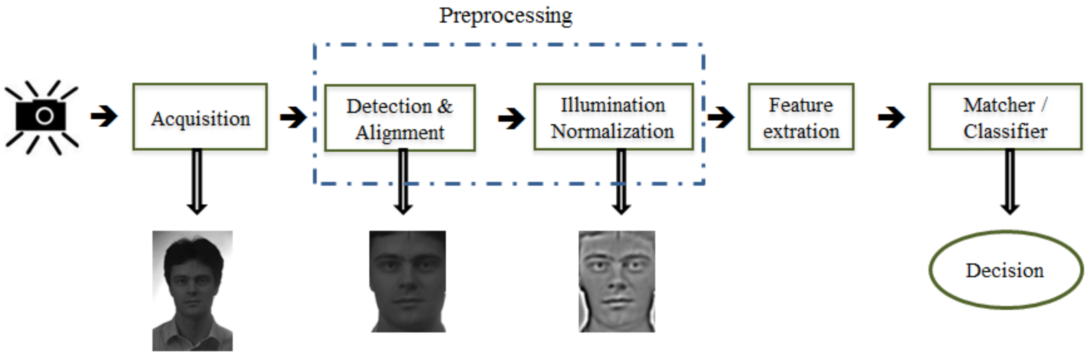


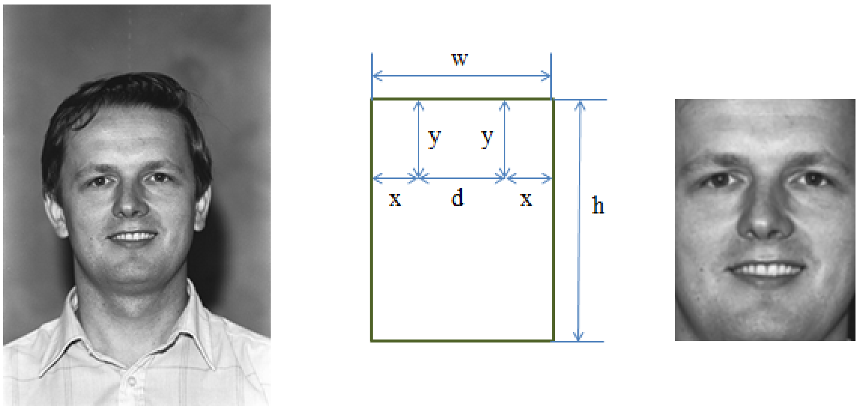
Figure 2.1: The framework of a typical face recognition system.

2.1 Generic Face Recognition

In this section we survey the state-of-the-art of face recognition in the literature. As illustrated in Figure 2.1, face acquisition is the first step for a typical recognition system. Most current face recognition systems are based on face images captured in the visible light spectrum. One problem of these images is that the intra-personal variations in the appearance due to different illumination conditions are larger than the inter-personal changes. This motivated the development of various special devices to obtain different face modalities, such as 3D face shape, near-infrared face images [46], thermal face images, in order to eliminate dependence on illumination conditions. In this work, we focus on recognition from still images in the visible light spectrum only because this type of sensor is widely available. Next, we focus on the following procedures illustrated in Figure 2.1.

2.1.1 Face Detection and Alignment

Face detection and geometric normalization is an essential step for most automatic face recognition systems. Its reliability plays an important role in the performance and usability of a face recognition system [58]. This step aims to provide a face region of interest for the feature extraction stage, in order that the facial features can be extracted from the detected face area. More importantly, this stage is vital



(a) Original image (b) Alignment template (c) Aligned image

Figure 2.2: Face normalization method in our experiments.

to make the extracted facial features comparable with each other. Detecting a face captured in an uncontrolled scene is very difficult. Over the years, significant advances have been made in achieving robust face detection. In [99], the authors provide a comprehensive and critical review of face detection methods. Up to now, perfect face localization is still very difficult to achieve. Therefore, a face recognition method capable of working well in the presence of localization errors is highly desired. In our work, the face detection problem is not considered, only face alignment is performed to normalize the captured face images in terms of geometry and size. Specifically, we assume the coordinates of the eyes in the face images are provided, which can be used as the ground-truth alignment. In our experiments, faces are normalized using the approach illustrated in Figure 2.2. First, we crop out the face region using the template illustrated in Figure 2.2 (b), by placing the two eye centers at fixed locations specified with (x, y) and $(x + d, y)$, respectively. The image is then rescaled to a standard size. The normalized image is shown in Figure 2.2 (c) from its original form as in Figure 2.2 (a). After this geometric normalization, all the face images are of the same size and their eye centers are in the same locations, thus making the comparison between two images meaningful.

2.1.2 Illumination Normalization

Illumination normalization aims to eliminate the illumination effect among different images. The existing approaches for dealing with this issue can be broadly divided into two categories. The first category tries to collect a set of training face images under different lighting conditions and uses them to learn a global model of the possible illumination variations such as a linear subspace [12] and an illumination cone [10], which are then generalized to the variations seen in new images. One disadvantage of this category is a large number of training images are required for the global model. In comparison with these approaches, there is not a training procedure in normalization based approaches. This category tries to employ some conventional image processing transformations to remove the influence of illumination variations from face images. These normalization based methods include the simple Histogram Equalisation [27], Multiscale Retinex method [36] and approaches based on Self Quotient Image [11, 21]. In this work, we do not employ any these techniques for illumination normalization. As we shall see in the following sections, our empirical results clearly show that the techniques developed in this work can get stable performance in the presence of different lighting conditions.

2.1.3 Feature Extraction

Effectively extracting useful features from face images is of great importance to a face recognition system. Over the years, many feature extraction algorithms have been developed. According to the type of features used by various methods, the existing feature extraction methods can be broadly classified into two groups: holistic methods and local methods.

In holistic methods, each face image is represented as a single high-dimensional vector by concatenating the gray values of all pixels in the face. These appearance based methods can implicitly preserves all the detailed shape and texture information which is useful for identifying faces. More importantly, they can capture more global information of the face than local methods. However, one disadvantage of this group is many training face images per person are needed to accurately estimate the intra-personal variations. In the literature, this group includes those widely used subspace methods [87, 88]. The eigenface methods (principal component analysis, PCA)uses the Karhunen-Loeve Transform to produce the most expressive subspace for face representation and recognition [83]. The fisherface (linear discriminant analysis, LDA) is an example of the most discriminating subspace methods. It tries to find a set of features which can best separate different face classes from each other[12]. In [9], the independent component analysis (ICA) is introduced as a powerful feature extraction tool for face recognition. ICA can be considered as a generalization of the PCA. However, the transformation vectors found by ICA are not necessarily orthogonal. Furthermore, it captures discriminative information by not only considering the covariance matrix, but also utilizing the high-order statistics.

Local methods which use local facial features for face recognition are relatively mature approaches in the field with a long history [18, 42, 54, 90]. In comparison with holistic methods, local methods provide additional flexibility to recognize a face based on its parts, thus the common and class-specific features can be easily identified. Furthermore, different facial features can increase the diversity of the classifiers[40], which is helpful for face recognition. The local methods can be further divided into two categories. The first category is the local feature-based methods, which usually locate local fiducial points first and then extract features on the detected feature points. Most early face recognition methods belong to this category. In these methods, a set of geometrical measures, such as the width of the head and the distance between eyes, are extracted and stored in the database as templates for the later matching usage. One representative method in this category is Elastic Bunch Graph Matching (EBGM) [90], which is known robust against illumination change, distortion and scaling [43].

The second category of local methods is based on the local appearance of faces. These approaches simply divide the face image into sub-regions. With the defined local regions, the next step is to decide how to represent the information of them. This is of great importance for the performance of a recognition system. The commonly used features include intensity features [58, 80] and a variety of derived features, such as Gabor wavelet [41, 49], local binary pattern (LBP) [6, 65], the scale-invariant feature transform (SIFT) features [15, 51]. In general, gray-value feature is the simplest feature without loss of texture information, while Gabor features and other derived features are more robust against illumination changes and some geometrical translations.

The reader can be referred to [100] for a comprehensive and critical survey of face extraction methods. In this section, we present only those feature extraction methods which are employed by our work.

Gabor. The Gabor filters (kernels) with orientation μ and scale υ are defined as [49]:

$$\psi_{\mu,\upsilon}(z) = \frac{\|k_{\mu,\upsilon}\|^2}{\sigma^2} e^{\frac{-\|k_{\mu,\upsilon}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{\mu,\upsilon}z} - e^{\frac{-\sigma^2}{2}}] \tag{2.1}$$

where $z = (x, y)$ denotes the pixel, and the wave vector $k_{\mu,\upsilon}$ is defined as $k_{\mu,\upsilon} = k_\upsilon e^{i\phi_\mu}$ with $k_\upsilon = \frac{k_{\max}}{f^\upsilon}$ and $\phi_\mu = \frac{\pi\mu}{8}$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain. Besides, σ determines the ratio of the Gaussian window width to the wavelength. The convolution of an image I with a Gabor kernel $\psi_{\mu,\upsilon}$ outputs $G_{\mu,\upsilon}(z) = I(z) * \psi_{\mu,\upsilon}$, where "$*$" denotes the convolution operator. The Gabor filtering coefficient $G_{\mu,\upsilon}(z)$ is a complex number, which can be rewritten as $G_{\mu,\upsilon}(z) = M_{\mu,\upsilon}(z) \cdot e^{i\theta_{\mu,\upsilon}(z)}$ with $M_{\mu,\upsilon}(z)$ being the magnitude and $\theta_{\mu,\upsilon}(z)$ being the phase. It is known that magnitude information contains the variation of local energy in the image. In [49], with five scales and eight orientations of the Gabor filters, the augmented Gabor feature vector F is defined via uniform down-sampling, normalization and concatenation of the Gabor filtering coefficients as follows:

$$F = \left(a_{0,0}^{(\rho)^T} a_{0,1}^{(\rho)^T} \cdots a_{4,7}^{(\rho)^T} \right)^T \tag{2.2}$$

where $a_{\mu,\upsilon}^{(\rho)}$ is the concatenated column vector from down-sampled magnitude matrix $M_{\mu,\upsilon}^{(\rho)}$ by a factor of ρ, and T is the transpose operator.

LBP. The LBP operator is a good means of summarizing the local gray-level structure. It takes a local neighborhood around each pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary-valued image patch as a local image descriptor [6, 65]. With a neighborhood of size 3 × 3 pixels, the mathematical equations for computing the LBP value at a pixel p are given by

$$LBP(p) = \sum_{n=0}^{7} 2^n s(I_n - I_p) \tag{2.3}$$

where I_n represents the intensity value of a neighborhood pixel n, I_p is the intensity value of the central pixel p, and

$$s(u) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2.4}$$

The LBP values of all the pixels in an image are then quantized into a histogram for a texture descriptor. Generally, before extracting LBP features, the face image is first divided into several regions. For each region, we can compute a histogram. The final LBP feature vector is calculated by concatenating histograms corresponding to all the regions. In addition, for face recognition, uniform local binary patterns are more widely used. An LBP is “uniform” if it contains at most one 0-1 and one 1-0 transition when viewed as a circular bit string. Uniformity is important because it characterizes the patches that contain primitive structural information such as edges and corners.

SIFT. The scale invariant feature transform, called SIFT descriptor, is proposed in [51] and proved to be invariant to image rotation, scaling, translation, partly illumination changes, and projective transform. Given an image $I(x, y)$, the corresponding scale space image $L(x, y, \sigma)$, at a scale σ, is obtained as

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.5}$$

where $G(x, y, \sigma)$ is a variable-scale Gaussian function and the symbol $*$ represents the convolution operation. A set of Difference of Gaussian (DOG) images, between scales separated by a multiplicative factor k, are obtained by

$$DOG = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{2.6}$$

With the computed DOG images, next the scale space extrema is detected by choosing the local maxima or minima among eight neighbors of a pixel in the current DOG image, and nine neighbors each in the scales above and below the current DOG image. The extrema points are further processed to select those keypoints with high stability. A 36 bin orientation histogram covering the 360 degree range of orientations is then generated by computing the gradient magnitudes $m(x, y)$ and orientations $\theta(x, y)$ of sample points within a region around the keypoint. The gradient magnitude and the orientation are computed as follows

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{2.7}$$

$$\theta(x, y) = tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \tag{2.8}$$

The orientation of the keypoint is computed as the highest peak in the orientation histogram associated with it. The feature descriptor is obtained by sampling the gradient magnitude and orientations within a descriptor window around a keypoint. The final keypoint descriptor of dimension $4 \times 4 \times 8 = 128$ is generated by computing an 8 bin orientation histogram over 4×4 sample regions within the descriptor window.

2.1.4 Classification

The goal of a classifier is to compare the features of a test face image with those of the template and give the final decision in terms of some similarity measures. For face identification, the most widely used classifier is the Nearest Neighbor (NN) classifier [25]. An important issue for designing the NN classifier is how to measure the similarity. A direct way is to measure how similar the two compared images are. Another possible way is to compute the distance between the two image features. These two measures are the inverse of each other. There are many possible similarity and distance measures such as Cosine similarity, Euclidean distance, Mahalanobis distance. Let x_1 and x_2 represent the feature vectors of two face images being compared, we present those similarity and distance measures that will be used in this work below:

$$\textit{Cosine similarity}: \qquad d = \frac{{x_1}^T x_2}{\|x_1\|\|x_2\|} \tag{2.9}$$

$$\textit{Euclidean distance}: \qquad d = \|x_1 - x_2\| \tag{2.10}$$

$$\textit{Hellinger distance}: \qquad d = \|\sqrt{x_1} - \sqrt{x_2}\| \tag{2.11}$$

Support Vector Machine (SVM) classifier [19, 79] and boosting classifier [85, 97] are also applied to face identification. Both classifiers are originally proposed for two-class discriminant problems, i.e., verification problems. As is known, face identification is a multi-class problem. To convert this multi-class problem into a binary problem, one direct approach is to evaluate the difference between two feature vectors as a basis for determining whether the images are of the same person. Another approach is to establish a set of classifiers each of which separates a single class from all the remaining classes.

In comparison with the NN classifier, those binary classifiers are more commonly used in face verification systems. In addition, a training stage is required to estimate the values of the classifier parameters. Generally, binary classifiers are score-based classifiers. That is, the input of these classifiers is the matching score (Euclidean distance or cosine similarity) of the two compared feature vectors or a set of component matching scores (in score-level fusion applications). Typical binary classifiers include the above mentioned SVM, boosting classifiers and logistic regression (LR) classifier [37]. In this work, for face verification and face pair matching, we apply the LR classifier for the decision making. Next, we give a brief introduction of this classifier.

Suppose $Y = 1$ represent two face images from the same person, while $Y = 0$ denotes a pair of images from two different persons. After obtaining the matching score or the matching score vector (in the fusion context) X, the logistic regression uses a parametric form for the $P(Y = 1 \mid X)$ distribution as follows

$$P(Y = 1 \mid X) = \frac{1}{1 + exp\left(\omega_0 + \sum_{i=1}^{n} \omega_i X_i\right)} \tag{2.12}$$

where $P(Y = 1 \mid X)$ is the probability of the given instance X belonging to the positive class, and $\frac{1}{1+exp\left(\omega_0+\sum_{i=1}^{n} \omega_i X_i\right)}$ is a sigmoid function. Next, a logit transformation is performed on the probability to transform it into a quantity which is linearly related to the explanatory variables (items in the matching score vector). The logit transformation is given by

$$logit(P(Y = 1 \mid X)) = log\left(\frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)}\right) \tag{2.13}$$

where $P(Y = 0 \mid X)$ is the probability of the given instance X coming from the negative class. With the sigmoid function in Eq 2.12, the following equation can be obtained:

$$logit(P(Y = 1 \mid X)) = -\left(\omega_0 + \sum_{i=1}^{n} \omega_i X_i\right) \tag{2.14}$$

Next, the classification can be performed by the following linear classification rule:

$$\begin{cases} Y = 1 & \left(\omega_0 + \sum_{i=1}^{n} \omega_i X_i\right) < 0 \\ Y = 0 & otherwise \end{cases} \tag{2.15}$$

A common approach used for learning the parameters ω_i is to choose parameter values that maximize the conditional data likelihood, which is the probability of the observed Y values in the training data, conditioned on their corresponding explanatory variable values.

2.2 Performance Evaluation

Comparing different face recognition algorithms is not an easy task due to the fact that evaluation must be performed on a large number of samples acquired in various conditions representing realistic scenarios. Hence, effective performance measures and large-scale public databases are significant for this task. In this section, we introduce the performance indicators together with face databases used in this work.

2.2.1 Performance Measures

Finally, to well evaluate different face algorithms, we give some evaluation measures that will be used in this work. As presented in Section 1.1, there are generally three face recognition tasks. Since each task has its own set of performance measures, we separately present those measures used in each task.

Face identification. When a probe face of an individual is presented to an identification system, the system first compares the input face to each gallery template stored in the database. The output of the system is a set of similarity scores. It is expected that the correct match will have the top similarity score. the identification rate at Rank-1 or Rank-1 accuracy is the probability that the system returns a correct result with the top match. If the correct result is in either the top or second ranked score, we call it Rank-2 accuracy. Note that we do not necessarily care if it is in the top or second rank specifically, just that it is in one of these two positions. Similarly, the probability of correct identification at Rank-N means the probability that the correct match is somewhere in the first N highest similarity scores. A Cumulative Match Characteristic (CMC) curve shows the probability of identification for numerous ranks [33]. In the following work, we use both Rank-1 accuracy and CMC curves (from Rank-1 to Rank-10) for evaluating identification algorithms.

Face verification. In the verification task, a user must first make a claim as to his/her identity and the verification system then compares the user's query face to the claimed template, which outputs a similarity score. This similarity score is then compared to the system-specific threshold for the final decision. For a genuine claim (i.e., the user's claim is true), if its similarity score is lower than the threshold, the system will incorrectly determine that the user is not from its claimed identity. This is called a false rejection or false non-match. The false rejection rate (FRR) or false non-match rate (FNMR) is the probability that the system fails to verify a genuine claim, which measures the percent of valid inputs which are incorrectly rejected [33]. When dealing with an impostor user (i.e., the user's claim is not true), if its similarity score is higher than the threshold, the system will incorrectly verify that the user is from the claimed identity. This error is called a false acceptance or false match. The false acceptance rate (FAR) or false match rate (FMR) is the probability that the system incorrectly matches an impostor user to a non-matching template in the database [33]. It measures the percent of invalid inputs which are incorrectly accepted.

There is a tradeoff between FAR and FRR in every verification system, as both FAR and FRR are a function of the threshold. If the threshold is raised, there will be fewer false accepts and more false non-matches. While a lower threshold can lead to a lower FRR and higher FAR. For a given value of the threshold, there is a corresponding pair of FAR and FRR. They can be plotted against each other as a curve known as Receiver Operating Characteristic (ROC) to express the behavior of FAR and FRR. Detection error trade-off (DET) is another visual characterization of the trade-off between FAR and FRR [53]. DET is obtained using normal deviate scales on both FAR and FRR axes. This more linear graph illuminates the differences for higher performances. In many applications, the threshold is usually chosen at the unique operating point where FAR is equal to FRR, called Equal Error Rate (EER) [33]. Lower EER leas to better performance. In addition, FNMR at a FMR of 1% (FMR100) and FNMR at a FMR of 1‰(FMR1000) are also used as performance indicators in some sections [53].

Face pair matching. Since face pair matching is also 1:1 matching, we use the same performance indicators as those used in face verification.

2.2.2 The AR Database

The AR database was collected in a lab environment, all the pictures were taken under strict controlled conditions [57]. It consists of over 4,000 color images from 126 persons. All the images are captured under frontal views but with different expressions, different illumination conditions and with different occlusions. For each individual, 26 pictures were taken in two separate sessions, each with 13 pictures. The two sessions are separated by two weeks (14 days) time. Figure 2.3 and Figure 2.4 show the 13 face crops of an individual captured in each of

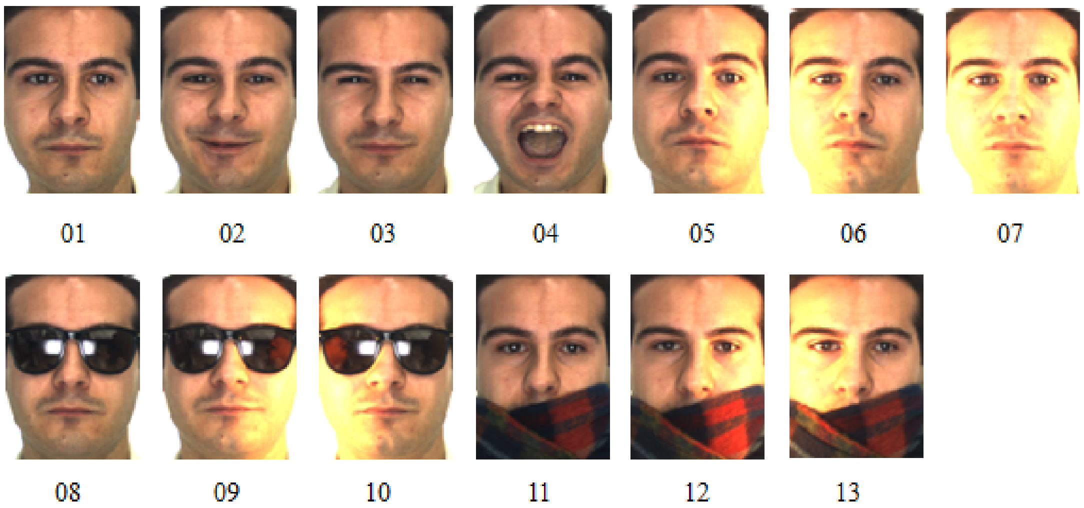


Figure 2.3: The 13 images of an individual from the first session of the AR database.

the two sessions [59].

2.2.3 The FERET Database

The FERET database was collected at George Mason University and the US Army Research Laboratory facilities [69]. This database consists of more than 13,000 facial images corresponding to more than 1,500 individuals. The diversity of the database is across age, gender and ethnicity. The images were captured in grey scale at resolution 256 by 384. The standard evaluation protocol on this database is designed for face identification, which includes a gallery partition and four probe partitions as listed in Table 2.1. Figure 2.5 illustrates some example images from each of these partitions on this database.

2.2.4 The FRGC ver2.0 Database

The Face Recognition Grand Challenge Version 2.0 (FRGC ver2.0) is a large database along with a data corpus of 50,000 images which consists of high resolution still images, 3D scans and multi-images of each subject taken under con-

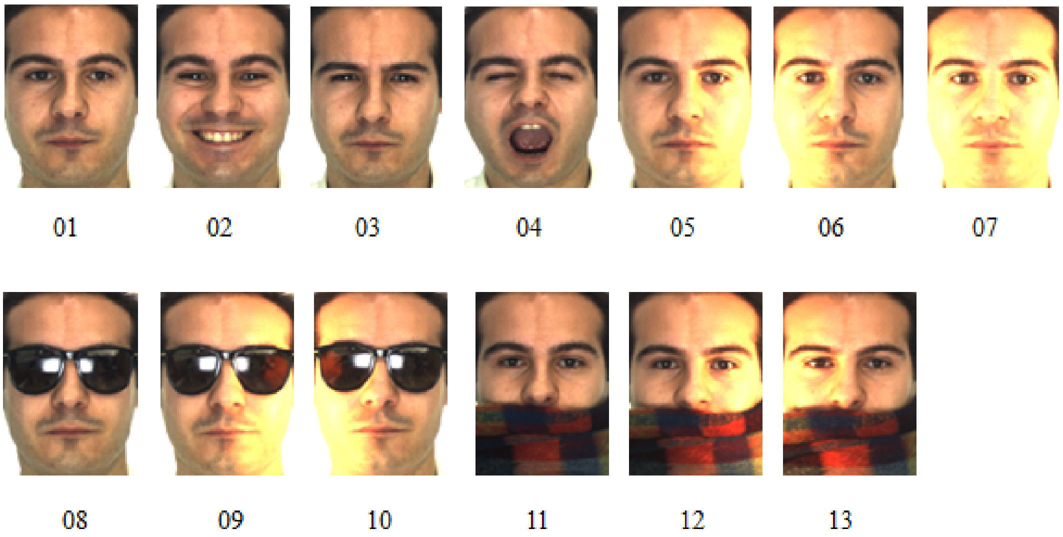
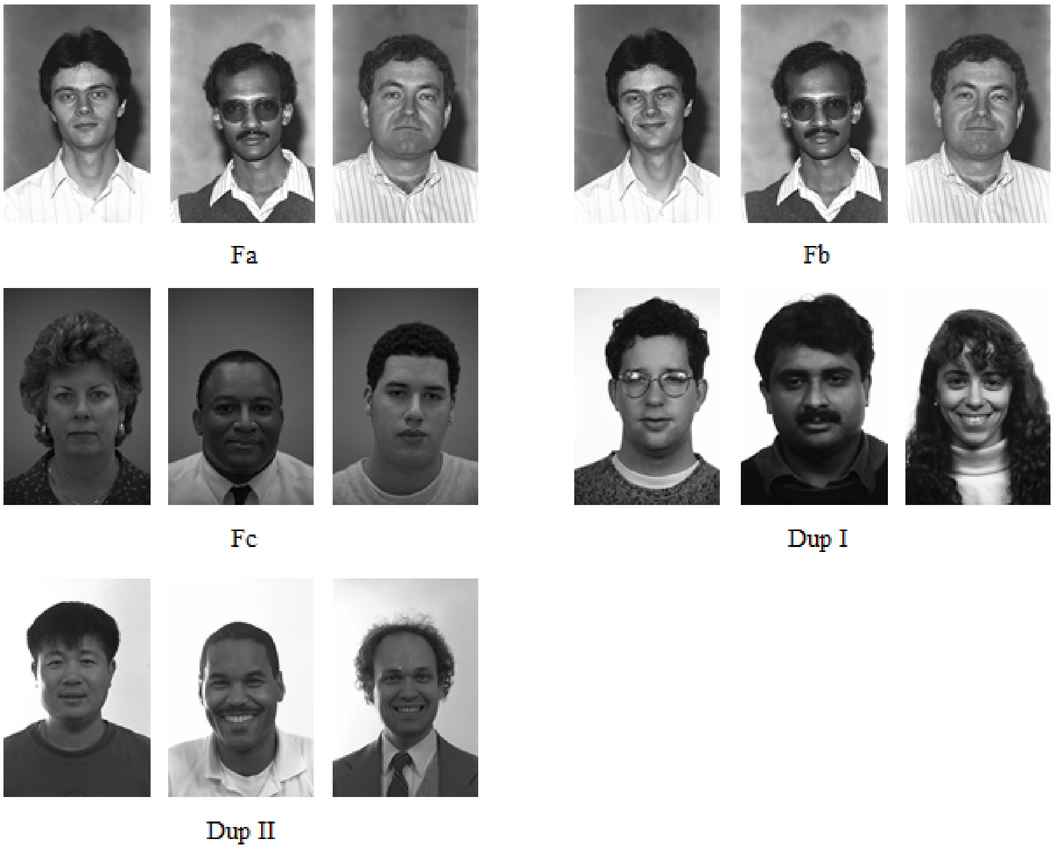


Figure 2.4: The 13 images of an individual from the session second of the AR database.

Table 2.1: Description of the subsets of the FERET database.

	# of images	Description
Fa gallery set	1,196	images with two different facial expressions: neutral and smile
Fb probe set	1,195	images with facial expressions other than neutral and smile
Fc probe set	194	images with different lighting conditions
Dup I probe set	722	images taken in a different time (between a minute and 1031 days after their gallery entries)
Dup II probe set	234	images taken at least a year after their gallery entries



Fa

Fb

Fc

Dup I

Dup II

Figure 2.5: Sample images from each partition on the FERET database.

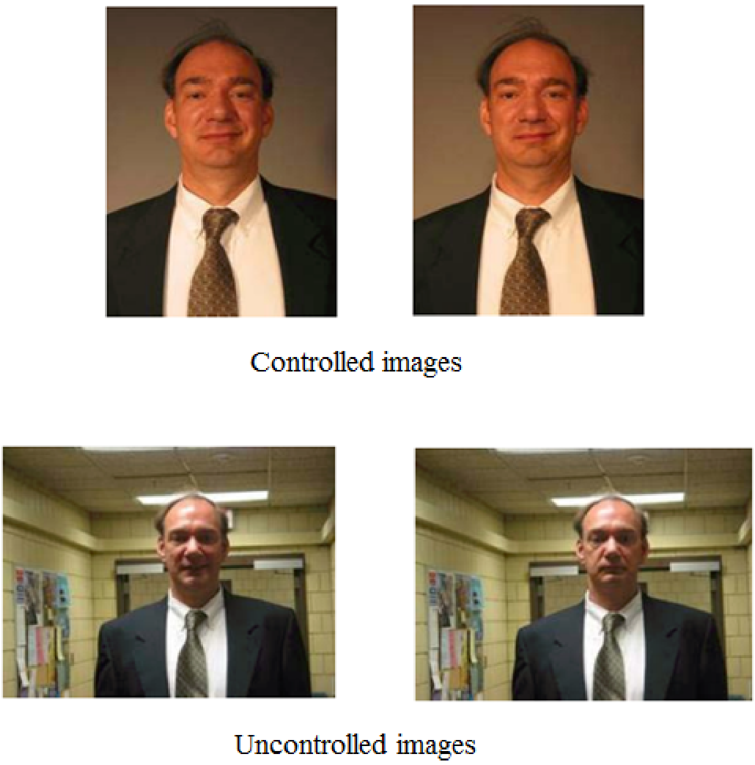
trolled and uncontrolled environments [70]. Data in this database was collected at the University of Notre Dame, and was divided into training and test partitions. The data used for training was acquired during the 2002-2003 academic year, including 12,776 images from 222 persons where 6,388 were captured under controlled conditions (i.e., in a lab setting under two lighting conditions, images were taken with frontal pose and two facial expressions) and the others are taken in uncontrolled environments (i.e. in an outdoor environment, images were taken with two facial expressions). Images in the test partition were collected in the 2003-2004 academic year consisting of 4,007 subject sessions from 466 subjects.

The experimental protocol on this database consists of 6 experiments. In this work, we use images only from Experiment 4. This experiment is designed to measure the recognition performance on controlled versus uncontrolled frontal face still images. 16,028 controlled images and 8,014 uncontrolled images are included in this database. We show some examples of these controlled and uncontrolled images in Figure 2.6.

2.2.5 The LFW Database

Typically, face databases are collected in laboratories. These databases try to manipulate a large number of variables in a lab environment such as different poses, lighting conditions, expressions and occlusions. However, face images captured in a lab environment cannot well measure the "natural" distribution of faces. In order to well study face recognition problems in the real word, it is better to train and test face recognition algorithms on highly diverse sets of faces. The labeled faces in the wild (LFW) is a database containing as many as possible the nature variations [30]. In this dataset, more than 13,000 face images were collected from news articles in the web, which present a large degree of diversity. These images are from 5,749 different subjects. Of these, 1,680 subjects have two or more images and the remaining 4,069 people have just a single image in the database.

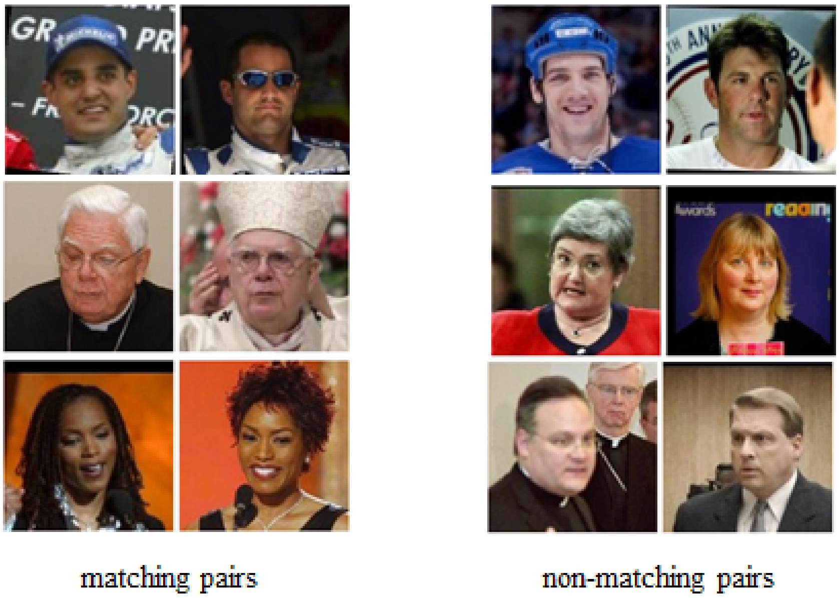
This database is designed for evaluating algorithms for face pair matching. The authors organized the data into two "Views". View 1 is for algorithm development, and View 2 is used to finally evaluate different algorithms. In View 1,



Controlled images

Uncontrolled images

Figure 2.6: Sample images from Experiment 4 on the FRGC ver2.0 database.



matching pairs

non-matching pairs

Figure 2.7: Examples of matching and non-matching pairs on the LFW database.

there are a training set and a test set. The training set includes a total of 2,200 pairs of images, half matching and half non-matching. The test set consists of 500 matching pairs and 500 non-matching pairs. View 2 is a 10-fold cross validation experiment, which consists of 10 subsets of the database. In addition, the authors provided researchers two training settings: image-restricted and unrestricted. Under the image-restricted setting, only the information about whether a pair of images is matched or mismatched are given to the experimenters. In contrast, the individuals' names associated with a pair of training images are given under the unrestricted training paradigm. That is, the experimenters are allowed to form as many matched and mismatched pairs as desired from a set of images labeled with individuals'names. One valued aspect of the LFW database is that the individuals used for training and those for test are mutually exclusive. In this work, we evaluate our developed algorithm on View 2 under the restricted training paradigm. Figure 2.7 shows some matching and non-matching pairs used in our work.

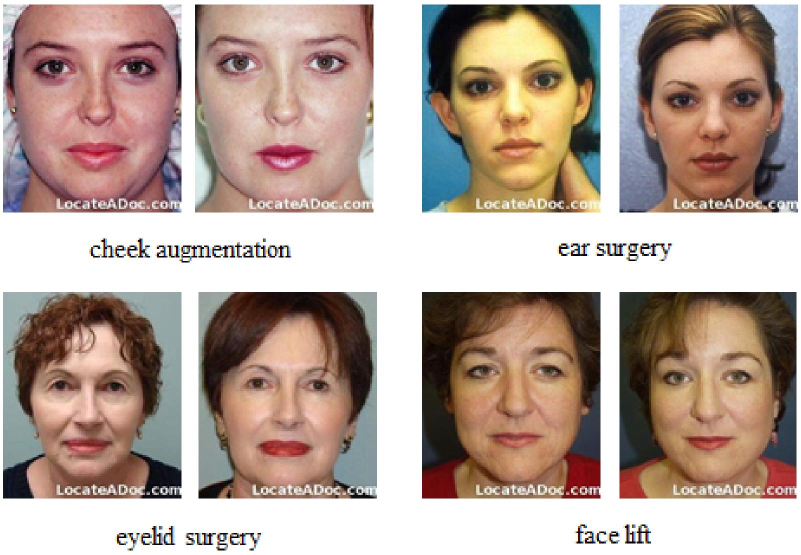


Figure 2.8: Some pre-surgery (left) and post-surgery (right) pairs on the plastic surgery database.

2.2.6 The Plastic Surgery Database

The plastic surgery database was introduced for evaluating current face algorithms on recognizing face images across plastic surgery [78]. This database contains images of individuals before and after facial plastic surgeries. The images were downloaded mainly from two websites [7, 26]. A total of 1,800 fully frontal face images from 900 subjects are included in this database. For each subject, there is 1 pre-surgery facial image and 1 pos-surgery facial image. The database contains a wide variety of local and global surgeries. A few pre-surgery and post-surgery pairs are shown in Figure 2.8. Table 2.2 lists the details of images in the plastic surgery database covering different types of surgeries.

2.3 Summary

In this chapter, we gave an overview of current face recognition techniques. An automated face recognition system usually includes the following important stages: face acquisition, face detection and alignment, illumination normalization, fea-

Table 2.2: Details of the plastic surgery database.

Type	Surgery procedure	# of subjects	Effects on face appearance
Local	Dermabrasion	32	local skin texture
	Blepharoplaty	105	eyelid
	Rhinoplasty	192	nose
	Brow lift	60	forehead
	Otoplasty	74	ear
	Others	56	local skin details or face components
Global	Skin peeling	73	global skin texture
	Rhytidetomy	308	global skin texture or face structures

ture extraction and classification. Face detection and geometric normalization is the essential step for a typical automatic face recognition system. Its reliability plays an important role in the performance and usability of a face recognition system. Illumination normalization aims to eliminate the illumination effect among different images. For feature extraction, many effective algorithms have been developed over the years. According to the type of features used by various methods, the existing feature extraction methods can be broadly classified into holistic methods and local methods. The goal of a classifier is to compare the features of a test face image with those of the template and give the final decision in terms of some similarity measures. For each of these stages, we presented the related state-of-the-art techniques used in this work. Finally, we introduced the performance measures and several well-known face databases used in this work for evaluating our developed algorithms.

Chapter 3

Cohort Score Normalization for Face Recognition

Cohort based score normalization is a procedure, which aims to post-process the matching score in a biometric verification system, using information from a set of cohort samples, i.e., non-matching samples / impostors of the claimed identity. In recent years, many cohort based score normalization methods have been proposed to improve the performance of biometric verification systems. Cohort models have been proposed to model language processing and lexical retrieval. For biometric applications, this technique was initially proposed for speaker recognition. The performance improvement achieved by making use of cohort information has also been seen in fingerprint verification systems and multi-biometrics framework. In this section, we introduce cohort normalization into face recognition systems to handle the challenging undersampled face recognition problem. Further, we develop an effective cohort normalization method specifically for the unconstrained face pair matching problem. In the end, we do an exhaustive experimental exploration to give a better understanding of the cohort behavior, such as how much useful information is contained in the cohort samples? What impact will be produced if cohort samples of different qualities are employed?

3.1 Introduction

Generally, a biometric verification expert makes the decisions based directly on the matching score between the query sample and the claimed template. However, it is vulnerable to a variety of degradation factors where score distribution variations between the enrollment and query sessions might be introduced. Cohort based score normalization is an effective tool to measure these various forms of degradation effects in relation to a set of cohort samples, which are non-matching samples of the claimed template model. Many approaches have been proposed for cohort based score normalization in the literature including employing moments of cohort scores for distribution scaling and exploiting useful information from sorted cohort score patterns. Cohort normalization has achieved promising results in various biometric systems.

3.1.1 Cohort Normalization for Biometric Verification

Biometric verification is a process that uses a person's physical or behavioural characteristics to validate the claimed identity of a query sample [31]. In a typical biometric verification system, a target model or template is first built for each subject using a few samples of the subject in the enrollment stage. In the test phase, a query sample will be compared against the claimed target model using a classifier or matching algorithm, of which the output is a matching score. Most biometric matching algorithms make the verification decision based purely on this matching score between the query sample and the claimed template. However, there are many degrading factors due to various forms of noises presented on the sensed data, which can make the straightforward usage of the raw matching score unreliable. For instance, when dealing with a noisy signal, the similarity of even a genuine query with the claimed identity gets very low. Generally, it is difficult to make a biometric expert adequately cope with all the degradation factors, thus post-processing the raw matching score, i.e., score normalization [32, 73], has been identified as an important stage. Score normalization aims to map the original matching score to a domain where the degradation effects on the matching score distribution is neutralized.

For score normalization, it will be of great use if the knowledge of class distributions is available. However, in many practical systems, only a few or even single samples are provided for each enrolled identity, making it difficult to learn these distributions. For example, if the score distribution of one subject follows Gaussian distribution, for deriving the specific distribution, both the mean and the standard deviation need to be well estimated. However, for those systems with very few samples for each subject, it is hard to get a promising estimation for both parameters. It is in these situations that one can make use of non-matching biometric samples already present in the database. The motivation for this is that all template models will be subject to the same degradation. It is therefore sensible to normalize the original matching score by measuring the corrupting effects in relation to a pool of non-matching models. These non-matching models are cohort models [76], which can be selected either from other reference models in the same database or from reference models of other databases.

Cohort models have been proposed to model language processing and lexical retrieval [56]. For biometric applications, cohort based score normalization was initially proposed for speaker recognition [75, 76]. In some literature, the term "background model" is used to indicate the same concept [75]. The performance improvement achieved by cohort score normalization has also been seen in fingerprint verification [3] and multi-biometrics framework [4].

Figure 3.1 shows a conventional fingerprint verification system. In this system, a query fingerprint is compared to its claimed template. The obtained matching score y is then used for the final classification. To illustrate the role of cohort based score normalization, an improved system with cohort normalization is shown in Figure 3.2. In this system, a set of H cohort samples are provided for cohort normalization. Through comparing the query fingerprint with the cohort samples ($1 : H$ match), or comparing the template fingerprint with the cohort samples ($1 : H$ match), or comparing both fingerprints with all the cohort samples, a set of cohort scores can be obtained. Cohort normalization is then employed to map the raw matching score y to a normalized value y^n by either estimating the score distribution parameters from the calculated cohort scores or extracting auxiliary information from sorted cohort score patterns. As shown in Figure 3.2, cohort score normalization is a procedure performed in the score level.

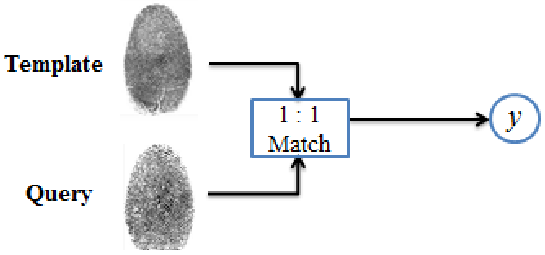


Figure 3.1: A fingerprint verification system without cohort score normalization.

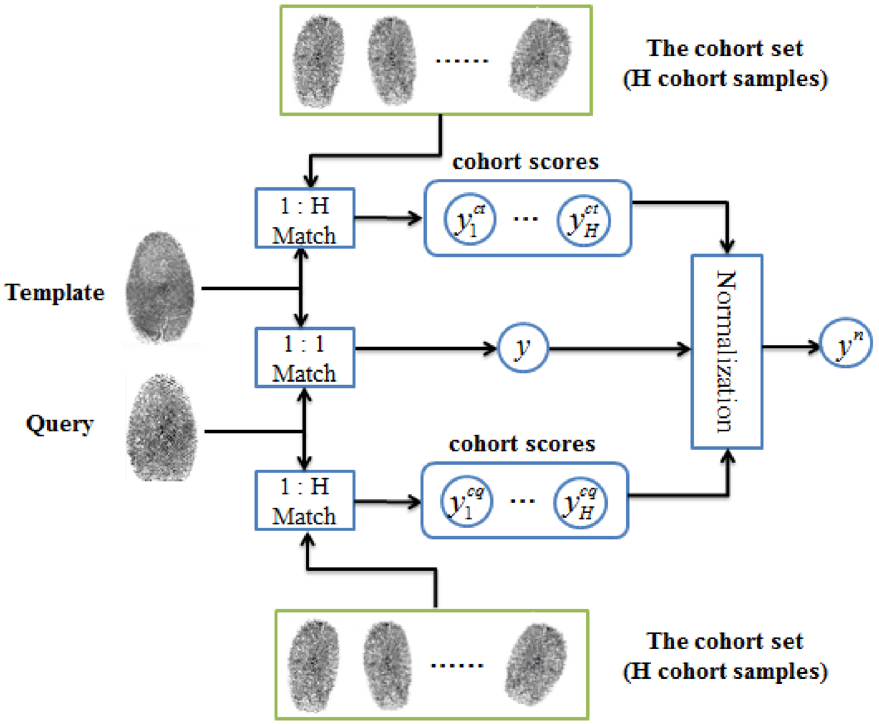


Figure 3.2: A fingerprint verification system with cohort score normalization.

3.1.2 Cohort Normalization Methods

In the literature, many cohort based score normalization approaches have been proposed. These approaches can be roughly categorized into two classes. Methods in the first class try to estimate score distribution parameters from cohort scores and then use these estimated parameters to perform distribution scaling, whereas approaches in the second class try to extract some useful information from sorted cohort score patterns.

Cohort normalization by distribution scaling. Among those approaches in the first category, zero-normalization (Z-norm) [8] and test-norm (T-norm) [8] are two most widely used algorithms in practical biometric systems. Z-norm is a method performed in the enrollment stage. In this method, the cohort scores used for normalization are matching scores between the template model and all the cohort samples. Under the assumption that score distribution of each subject class follows Gaussian distribution, Z-norm uses the mean and variance estimation for distribution scaling. During the enrollment stage, for each enrolled subject, the corresponding template model is matched against all the cohort samples and the matching scores (i.e., cohort scores) are then used to estimate the mean and standard deviation specific to each subject for the impostor distribution. Hence, Z-norm is impostor centric and leads to user-specific normalization parameters [73]. The advantage is that the estimation of the normalization parameters can be performed off-line during the enrollment. Suppose there are a total of H cohort samples, the cohort scores between the template model and these cohort samples can be represented as $Y^{ct} = \{y_1^{ct}, ..., y_H^{ct}\}$. Let y be the original matching score between the query sample and the claimed template model, then Z-norm can be performed as follows:

$$y^z = \frac{y - \mu^{ct}}{\sigma^{ct}} \tag{3.1}$$

where μ^{ct} is the expected value of Y^{ct} and σ^{ct} is the standard deviation. y^z is the Z-normalized score.

Among those widely used cohort normalization methods, T-norm is one of the most well-established ones. What is different from Z-norm is the cohort score set. In T-norm, cohort scores are computed between the query sample (not the template model) and all the cohort samples. Hence, this normalization can only be performed on-line during the test stage. What is similar to Z-norm is the assumption on the score distribution. That is, T-norm also assumes that the score distribution of each subject class follows Gaussian distribution, and thus uses the mean and variance estimation for distribution scaling. It is expected that the normalized score obtained by non-matching samples would have zero mean and unit variance. Denote the cohort scores obtained by comparing a query sample with the H cohort samples by $Y^{cq} = \{y_1^{cq}, ..., y_H^{cq}\}$. The normalized score y^t by T-norm can be given by:

$$y^t = \frac{y - \mu^{cq}}{\sigma^{cq}} \tag{3.2}$$

where μ^{cq} is the expected value of Y^{cq} and σ^{cq} is the standard deviation.

Cohort normalization by sorted cohort scores. While T-norm has been a dominant approach since its proposal, attempts have been made to exploit the patterns of sorted cohort scores. By combining the original matching score with the maximum of cohort scores using a SVM classifier, Tulyakov et al. proposed a very different cohort normalization method from those based on distribution scaling [82]. In this method, if the best score is the original matching score, the maximum of cohort scores can be viewed as “the second best score”. Note that, the cohort scores used in this method are similar to those used in T-norm, i.e., matching scores between the query sample and all the cohort samples. Therefore, the cohort sample corresponding to the maximum cohort score is the most similar one to the query sample.

Another similar approach is to use a few top cohort scores (not only the maximum one) together with the raw matching score as input features of a SVM classifier [4]. Note that, in this method, the cohort scores between the query and several most similar cohort samples are used for normalization. This approach has been shown its effectiveness in a multi-biometric framework. It is worth nothing that in this approach, the cohort scores need to be sorted in order that the most similar cohort samples (corresponding to the most largest cohort scores) can be determined.

For exploiting sorted cohort score patterns, the above two methods use only the most similar cohort samples to the query sample. Polynomial regression based cohort score normalization [61], however, tries to extract some discriminative information from both the most similar and the most dissimilar cohort samples. In this approach, all the cohort samples are first sorted with respect to their closeness to the claimed template model (not the closeness to the query sample). After this ordering process, each rank corresponds to its own cohort sample. That is, Rank 1 corresponds to the most similar cohort sample to the claimed template, whereas Rank H corresponds to the most dissimilar one, if a set of H cohort samples are provided. Next, the distribution of cohort scores for each rank can be obtained by matching a number of impostor and genuine query samples with the cohort sample corresponding to that particular rank. It has been shown that such distribution of cohort scores produced some discriminative patterns between genuine and impostor query samples. Figure 3.3 illustrates the discriminative patterns between genuine queries/match queries and impostor queries/non-match queries implied in the sorted cohort scores. The data used for drawing the cohort score distribution is the face modality of the Biosecure database [67]. The cohort set contains a total of 325 cohort face images. The cohort scores are produced by comparing 54,108 impostor and 334 genuine accesses with the cohort samples. The mean and variance of cohort scores for each type of queries (i.e., genuine and impostor) are used to represent the distribution of cohort scores of each particular rank. As shown in Figure 3.3, the distribution of cohort scores for genuine claims follows a decreasing profile versus the rank order, while for impostor claims it follows a relatively constant profile. Another interesting discovery is that the discriminative patterns are presented not only on the first few ranks (corresponding to the most similar cohort samples to the claimed template) but also on the last few ranks (corresponding to the most dissimilar cohort samples). Hence, polynomial regression based cohort normalization extracts these discriminative patterns from both the most similar and the most dissimilar cohort samples. In comparison with the several above presented cohort normalization methods, polynomial regression performs not only the matching between the query sample and the cohort samples during the test stage, but also the matching between each enrolled template and the cohort samples during the enrollment stage (off-line) for producing the user-specific rank orders.

Let $C = \{c_1, ..., c_h, ..., c_H\}$ be the H sorted cohort samples with respect to

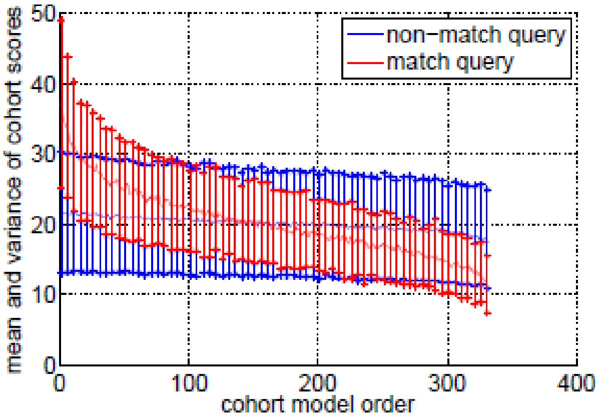


Figure 3.3: The mean and variance of scores produced by ordered cohort samples versus rank order for genuine and impostor queries of the face modality of Biosecure database. Rank 1 and 325 correspond to the most similar and the most dissimilar cohort samples to the claimed template [61].

their closeness or similarity to the claimed template model, in which c_1 is the most similar cohort sample to the claimed target model, whereas c_H is the most dissimilar one. Denote by $Y^{cq} = \{y_1^{cq}, ..., y_h^{cq}, ..., y_H^{cq}\}$ the H cohort scores between the query sample and each corresponding cohort samples in C, i.e., y_h^{cq} is the matching score between the query and c_h. The following is the methodology for polynomial regression based cohort normalization.

The H cohort scores in Y^{cq} are first considered as discrete points on a function of rank orders as follows

$$y_h^{cq} = f(h) \tag{3.3}$$

This function is then approximated using polynomial regression as follows

$$f(h) \approx w_n h^n + w_{n-1} h^{n-1} + ... + w_1 h + w_0 \tag{3.4}$$

where $W = [w_0, w_1, ..., w_n]$ is the polynomial coefficient vector and n is the degree of the polynomial. Further, the H cohort scores can be approximated by these $n+1$ coefficients. These coefficients can be subsequently combined with the original matching score y to improve the recognition performance using a classifier, e.g., SVM or LR classifier.

3.1.3 Understanding of Cohort Behavior

Cohort score normalization has received much attention in various biometric systems. However, there still lacks the understanding of cohort normalization. How different choices of cohort samples affect the generalization performance? What result will be yielded if different numbers of cohort samples are used? How much useful information is contained in the cohort samples? What impact will be produced if cohort samples of different qualities are employed? For the selection of cohort samples, should we use as many subjects as possible (each containing as few samples as possible) or should we choose as few subjects as possible (each consisting of as many samples as possible)? All these cohort behaviors need to be given a better understanding. However, in the literature, there is no work focusing on this. Hence, more efforts are desired to be made in this direction.

3.2 Cohort Normalization based Sparse Representation for Undersampled Face Recognition

Sparse representation based classification (SRC) has received much attention in face recognition with multiple training samples of each subject in recent years. However, it cannot be easily applied to a recognition task with insufficient training samples under uncontrolled environments. On the other hand, cohort normalization, as a way of measuring the degradation effect under challenging environments in relation to a pool of cohort samples, has been widely used in the area of biometric authentication. In this paper, we introduce cohort normalization to SRC-based face recognition with insufficient training samples. Specifically, a user-specific cohort set is selected to normalize the raw residual, which is obtained from comparing the test sample with its sparse representations corresponding to the gallery subject, using polynomial regression. Experimental results on AR and FERET databases show that cohort normalization can bring SRC much robustness against various forms of degradation factors for undersampled face recognition.

3.2.1 Overview

Robust face recognition via sparse representation is a technique developed for a face recognition system, where a rich set of carefully controlled training face images are provided [94]. Under the assumption that all the training samples from a single class lie in a low-dimensional subspace [87, 88] of a high-dimensional space, this technique tries to code the given test face image as a sparse linear combination of all the training images themselves, i.e., uses the fewest possible training samples to interpret the test sample. Sparse representation-based classification (SRC) measures the sparsity of the coding vector by l_0-norm, which counts the number of nonzero entries. Since the l_0-minimization problem is NP-hard, the l_1-minimization [95], as the closest convex function to l_0-minimization, is employed to find the sparsest coding vector. By optimizing the sparsity of such an over-complete linear representation, the dominant nonzero entries in the coding vector can reliably indicate the identity of the test sample. Finally, SRC performs the classification by checking which class yields the minimum representation error. This technique can effectively handle errors due to occlusion and corruption uniformly by exploiting the sparsity on the location of the distorted pixels in the face image.

Based on the findings in human perception that, accurate low-level and mid-level visual perceptions are a result of sparse representation of visual patterns using highly redundant visual neurons [66], sparse coding represents a signal using very few atoms chosen out of an over-complete dictionary. This methodology originally aims to represent and compress the input signals rather than perform classification. However, the sparsest representation is naturally discriminative: it selects the most compact representation and rejects all other possible but less compact representations. In [94], the discriminative nature of sparse representation was exploited to perform the task of face recognition. In the related work, a sparse representation-based classification (SRC) framework was proposed and demonstrated to have high performance on the face identification task.

As a holistic representation method, SRC is not as robust as local methods to facial variations. Yang et al. adopted local Gabor features into SRC to make it less sensitive to facial variations and reported impressive accuracy [96]. As presented above, in case of multiple well-aligned face training images per subject, SRC can extract stable and discriminative facial representations and thus achieve good performance. However, when moderate or large variations in pose and displacement are presented in the test image, SRC degrades sharply. There is some literature working on improving the original SRC for robust alignment, e.g., a series of linear programs were adopted to iteratively minimize the sparsity of the registration error in [86]. Furthermore, many real applications can only offer few, or even single, facial images per individual. Thus, SRC cannot be easily applied to undersampled and one sample per person face recognition [81].Some work for dealing with this problem includes Extended SRC [23] and Superposed SRC [22]. In Extended SRC, an auxiliary intra-class variant dictionary was introduced to extend SRC to undersampled face recognition and achieved promising results. While the Superposed SRC is based on a "prototype plus variation" representation model, in which the dictionary is assembled by the class centroids and the sample-to-centroid differences. This simple variant of SRC leads to an enormous improvement under uncontrolled conditions and only a single sample per class is available.

As described in Section 3.1, using cohort information to improve the recognition performance of a biometric expert has become a popular way. Many cohort-based score normalization approaches, which aim to post-processing the raw matching score, have been proposed to improve the performance of biometric verification systems. Though a lot of work has been done in exploiting cohort information for biometric authentication, to the best of our knowledge, little has been done to employ cohort normalization for undersampled face recognition. In this paper, for the first time, cohort normalization was introduced to improve the performance of undersampled face recognition using SRC. Specifically, a user-specific [61] cohort set is selected to perform score normalization using polynomial regression. In the context of face recognition using SRC, the raw score is the residual obtained from comparing the test sample with its sparse representation corresponding to the gallery model.

After sparsely coding the test sample with a gallery dictionary, we compute the raw residual between the test sample and each gallery subject. Then we employ the cohort set as a dictionary to sparsely code the test sample and get a set of cohort residuals between the test sample and each cohort sample. Before normalizing the raw residual using these cohort residuals, all the cohort samples are firstly sorted with respect to their closeness or similarity to the gallery subject, note that this is an offline procedure. These cohort residuals are then arranged with respect to rank orders of their corresponding cohort samples and considered as discrete points on a function of rank orders. Next polynomial regression is employed to approximate this function. Finally, the approximated polynomial coefficients are combined with the raw residual to improve the recognition performance.

3.2.2 Related Work

Denote the given k_m training samples from the m^{th} subject class by columns of a matrix A_m and the number of subjects in the training set as M. Assume that enough well-aligned training images of each of the M subjects are provided, these images are stacked as the columns of matrix $A = [A_1, ..., A_m, ..., A_M]$, i.e., the dictionary used to represent a test image. Given a new test image y, also well aligned, but possibly subject to illumination variation, the sparse representation framework tries to represent y as a sparse linear combination of the entire training dictionary A through solving the following optimization problem:

$$(l_0): \qquad \hat{x}_0 = \arg min \|x\|_0 \quad subject\, to \quad Ax = y \tag{3.5}$$

where $x = [x_1, ..., x_m, ..., x_M]$ is a coefficient vector (or coding/representing vector) and $x_m \in R^{k_m}$. Assume $K = \sum_{m=1}^{M} k_m$, i.e., we have a total of K training samples, we can easily get $x \in R^K$. $\|\cdot\|_0$ denotes the l_0-norm, which counts the number of nonzero entries. Suppose that y is from the m^{th} class, then $y \approx A_m x_m$ holds well, a naturally good solution to x will be that most coefficients in x are nearly zeros and only the coefficients associated with the m^{th} class have significant values. That is, the sparse non-zero entries in x can well encode the identity of the test sample y. SRC measures the sparsity of the coefficient vector by l_0-

norm. However, the combinatorial l_0-norm minimization is a NP-hard problem. To address this problem, the l_1-minimization, as the closest convex function to l_0-minimization, is employed in sparse coding as follows:

$$(l_1): \qquad \hat{x}_1 = \arg min\|x\|_1 \quad subject\ to \quad Ax = y \tag{3.6}$$

where $\hat{x}_1$ is the sparsest solution solving by the l_1-minimization and $\|\cdot\|_1$ denotes the l_1-norm. It has been shown that l_0-minimization and l_1-minimization are equivalent if the solution is sufficiently sparse [24]. The l_1-minimization can be solved in polynomial time by standard linear programming approaches.

Once the l_1-minimization problem has been solved, i.e., the estimate of the sparse coefficient vector $\hat{x}_1$, the identification of a given test sample can be processed by assigning the test sample y to an object class of which the training samples are associated with most significant coefficient values in the coding vector. To better harness the subspace structure associated with images in face recognition, SRC classifies y based on how well the coefficients associated with all training samples of each subject reproduce y. For each subject, let $\delta_m : R^K \rightarrow R^K$ be a characteristic function which selects the coefficients associated with the m^{th} subject. For $x \in R^K$, $\delta_m(x) \in R^K$ is a new vector whose only nonzero entries are the entries in x that are associated with the m^{th} subject. SRC performs the final classification as follows:

$$identity(y) = \arg \underset{m}{min}\{r_m(y)\}, \quad where \quad r_m(y) \doteq \|y - A\delta_m(\hat{x}_1)\|_2 \tag{3.7}$$

where $r_m(y)$ is the residual between y and its sparse representation using training samples of the m^{th} object class.

3.2.3 Cohort Normalization based Sparse Representation

The main idea of our proposal is to apply polynomial regression based cohort normalization to SRC-based undersampled face identification. Figure 3.4 illustrated our approach. After using the gallery dictionary A to sparsely represent the test sample y, we denote by r_m the raw residual between y and the m^{th} subject, where

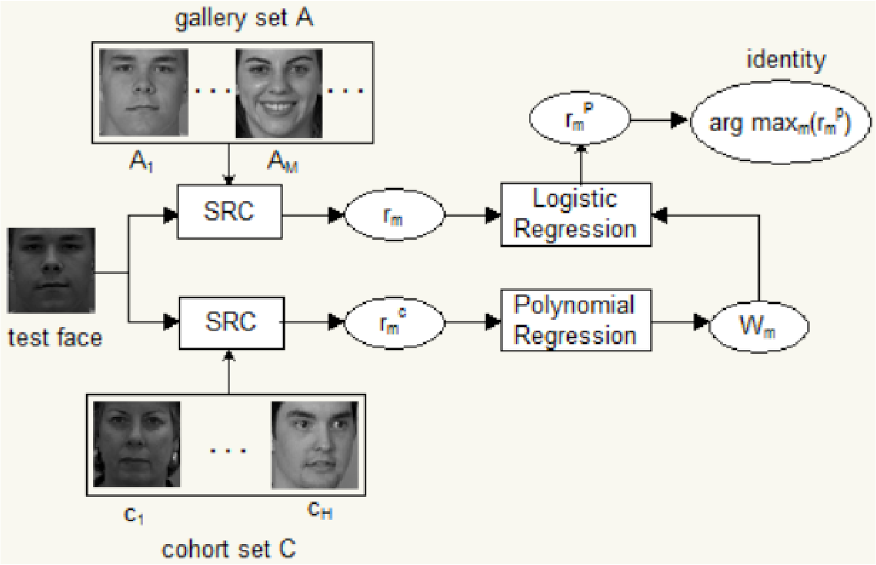


Figure 3.4: The framework of undersampled face recognition with sparse representation and cohort normalization.

$m = 1, 2, ..., M$. Next we focus on how to normalize the raw residual r_m using cohort information.

Suppose $C = \{c_1, ..., c_h, ..., c_H\}$ be the cohort set which contains H face images of subjects different from those M gallery subjects. The cohort set C is then employed as a dictionary to sparsely code the test sample y, which can produce H cohort residuals between y and each cohort sample. Before performing score normalization using these H cohort residuals, all the cohort samples in C are firstly sorted with respect to their closeness or similarity to the m^{th} gallery model. We represent by $C_m = \{c_{m1}, ..., c_{mh}, ..., c_{mH}\}$ the sorted cohort set of the m^{th} subject, in which c_{m1} is the most similar cohort sample to the m^{th} gallery model, while c_{mH} is the most dissimilar one. Note that sorting cohort models is an offline procedure. Next we arrange the H cohort residuals with respect to rank orders of their corresponding cohort samples in C_m. Let $r_m^c = \{r_{m1}^c, ..., r_{mh}^c, ..., r_{mH}^c\}$ be the H cohort residuals corresponding to the H cohort samples in C_m. In other words, r_{mh}^c is the residual between y and c_{mh}.

In [61], the authors illustrated that arranged cohort scores of genuine and impostor claims exhibit a discriminative pattern. Here, we also illustrate the distribution of cohort residuals generated by ordered cohort samples for genuine and impostor claims on the AR database. Figure 3.5 shows the distributions using Gabor and intensity features, respectively. For both figures, the distribution of cohort residuals for genuine claims follows an increasing profile versus the rank order, while for impostor claims it follows a relatively constant profile. Hence, we can obviously get that arranged cohort residuals of matching pairs (y and A_m are from the same person) and non-matching pairs (y and A_m are of different subjects) show some discriminative information. Next, we consider these cohort residuals as discrete points on a function of rank orders as follows

$$r^c_{mh} = f_m(h) \tag{3.8}$$

where $h = 1, 2, ..., H$. Next polynomial regression is employed to approximate this function as follows.

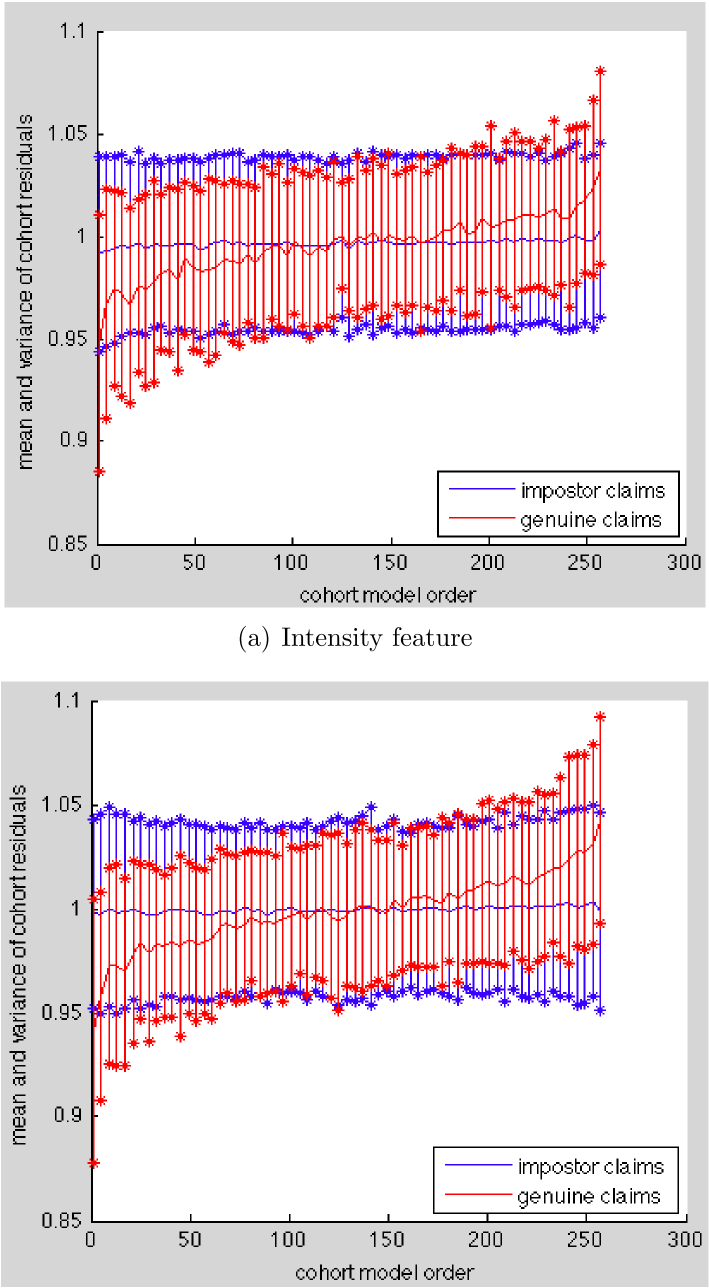
$$f_m(h) \approx w_{mn}h^n + w_{m,n-1}h^{n-1} + ... + w_{m1}h + w_{m0} \tag{3.9}$$

where $W_m = [w_{m0}, w_{m1}, ..., w_{mn}]$ is the polynomial coefficient vector. Then, the arranged cohort residuals obtained from the sorted cohort models of the m^{th} subject can be approximated by these $n + 1$ coefficients. These parameters can be combined with the raw residual r_m to improve the recognition performance. Here, we use a logistic regression classifier to perform the final classification as follows:

$$r^p_m = P(C \mid r_m, W_m) \tag{3.10}$$

where $P(C \mid r_m, W_m)$ represents the posterior probability of being a matching pair.

Finally, we get a total of M posterior probabilities $\{r^p_1, ..., r^p_m, ..., r^p_M\}$ corresponding to M gallery subjects. As described above, r^p_m denotes the posterior probability of being a matching pair. That is the larger r^p_m is, the more probable the test sample y is from the m^{th} subject. Hence, we can consider r^p_m as a kind of similarity between y and the m^{th} gallery model, i.e. the normalized score between y and the m^{th} gallery model. Consequently, we classify y based on



(a) Intensity feature

(b) Gabor feature

Figure 3.5: The distribution of cohort residuals generated by ordered cohort samples for genuine and impostor claims on the AR database. (a) Results of Intensity feature; (b) Results of Gabor feature.

$\{r_1^p, ..., r_m^p, ..., r_M^p\}$ by assigning it to the subject class that maximizes the posterior probability:

$$identity(y) = \arg \underset{m}{max}\{r_m^p\} \tag{3.11}$$

In summary, the main steps of SRC with cohort normalization are as follows:

1) Sort cohort samples with respect to their closeness to the gallery model.

2) Sparsely code y over the gallery dictionary and compute the raw residuals.

3) Sparsely code y over the cohort dictionary and compute the cohort residuals.

4) Perform polynomial regression and get the polynomial coefficient vector.

5) Employ a logistic regression classifier to approximate the posterior probability.

6) Perform classification via Eq 3.11.

As is observed, Eqs 3.3, 3.4 and Eqs 3.8, 3.9 look similar. Hence, it is significant to highlight the differences between them. In the former case, y_h^{cq} is an output of a biometric classifier or matcher, and it represents a cohort score between the query sample and a cohort sample in a biometric verification task. r_{mh}^c stands for a residual between the test sample and its sparse representation corresponding to a cohort sample, and this is more commonly used in the context of identification.

As is known, SRC does not perform well in a face recognition task with insufficient training samples under uncontrolled environments. On the other hand, we know that in a biometric verification system, the biometric samples used for building a template for each individual are usually limited, sometimes even a single sample. Hence, it is novel and reasonable to consider undersampled face recognition in a situation similar to biometric verification. Due to the success of applying cohort normalization to biometric verification, it is expected that cohort normalization can bring SRC good generalization ability for undersampled face recognition.

3.2.4 Experiments and Results

In this subsection, we present experiments on several representative face databases to assess the merit of our proposal. For fair comparisons, we choose $l1\_ls$ sparse reconstruction solver for l_1-minimization in both the original SRC and SRC with cohort normalization with identical parameters, so that the performance difference will be solely induced by the adoption of cohort normalization. For feature extraction, intensity and Gabor features [49] are respectively extracted to represent face images. It is not necessary to use all the cohort residuals to perform polynomial regression, as concluded in [61], a number of most similar and most dissimilar ones can provide much discriminative information. Hence, in all the following experiments, we select s most similar and s most dissimilar cohort residuals for polynomial regression. That is we choose only $2 * s \prec H$ cohort residuals, $\{r^c_{m1}, ..., r^c_{ms}, ..., r^c_{m,H-s+1}, ..., r^c_{mH}\}$, to perform polynomial regression. Note that the degree of polynomial has little impact on the generalization performance, hence we employ a linear function to fit the cohort residual function for simplicity, i.e. $n = 1$. For logistic regression classifier, we use l2-penalized logistic regression which leads to maximum likelihood estimate.

For each experiment in this subsection, we divide the total database into three partitions, where the identities are disjoint from one another. One of them is for the cohort set, while the remaining two are used to form the development and evaluation set. These partitions are referred to as C, D and E, respectively. C is the cohort set, D is used as the development set to train the polynomial coefficients, and E is the evaluation set. About the performance evaluation, we compute the Rank-1 accuracy for all the experiments [33].

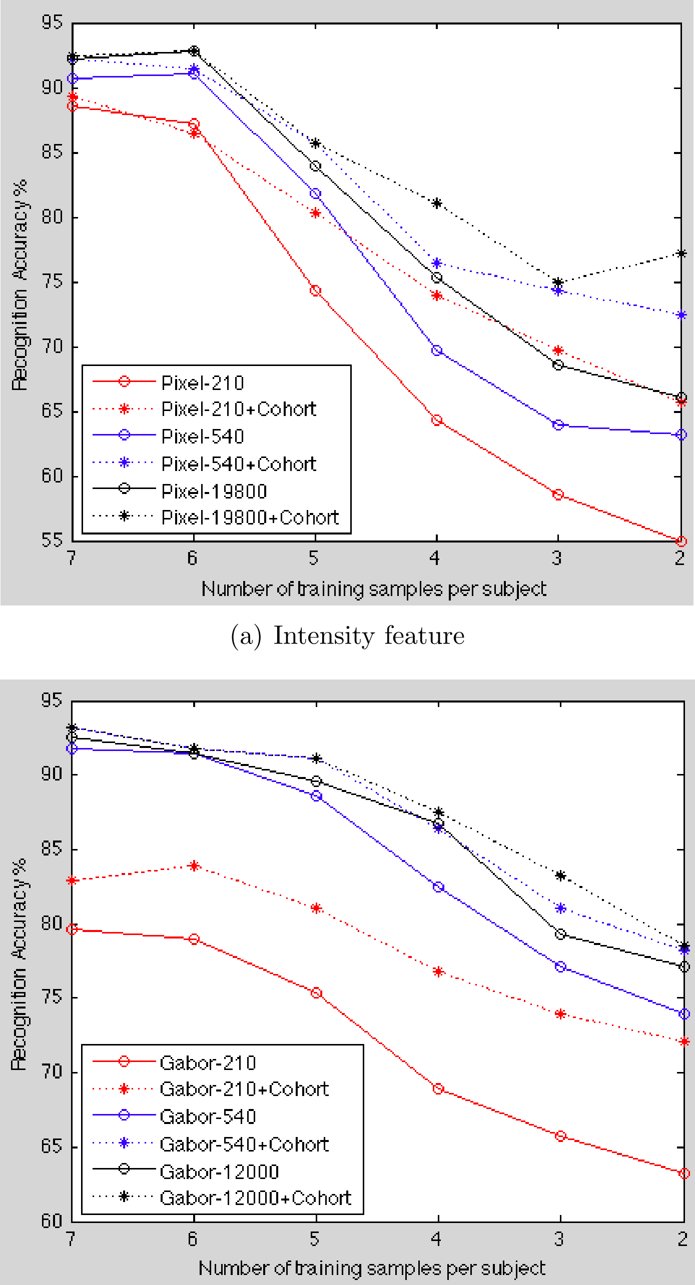
Undersampled face recognition. We perform this experiment to demonstrate the efficacy of our proposed approach for face recognition with insufficient training images. Specifically, we use the AR database, which consists of over 4,000 frontal images from 126 subjects [57]. As in [94], we select a subset (only with illumination changes and expression variation) consisting of 50 male subjects and 50 female subjects. For each subject, there are totally 14 images, seven images from Session 1 and the other seven from Session 2. All the images are cropped to 165×120 dimensions. We choose 40 subjects for both D and E. And for each subject, the seven images from Session 1 are used for training and the other seven from Session 2 for test. The remaining 20 subjects are considered as cohort models, each with 14 images, i.e., the size of the cohort set is $H = 280$. Here we select only $2 * s = 140$ cohort residuals for polynomial regression, i.e., 70 most similar ones and 70 most dissimilar ones. The intensity feature is a $165 \times 120 = 19,800$ dimensional vector. About Gabor features, we choose a family of Gabor kernels with five scales and eight orientations. The Gabor filter responses are then downsampled by a 116 uniform lattice based on the approach in [49]. Finally, we get a 12,000 dimensional feature vector for each face image. For both intensity and Gabor features, we further select dimensions of 210 and 540 randomfaces [94]. To test the undersampled effect, we reduce the number of training samples per individual from 7 to 2 in a step of 1.

Figure 3.6 shows the comparative performance of SRC with and without cohort normalization on the evaluation set E. As is expected, SRC degrades sharply as the number of training images decreases. Figure 3.6 (a) shows the results of using intensity features. It can be seen that in most of the total 18 cases (3 features with 6 samples sizes), SRC with cohort normalization performs better than the original SRC. Only in one case, no performance improvement is shown in SRC with cohort normalization. That is when there are 6 training samples per class with 210-dimensional pixel based randomfaces, SRC achieves 87.14%, while our approach achieves 86.43%. For intensity features, when the number of training samples per class is large, the performance shows little difference between SRC with and without cohort normalization. However, the superiority of SRC with cohort normalization gets more and more significant as the sample size decreases. For example, when the number of training samples is two, the accuracy increases 10.71%, 9.29% and 11.07% for the three kinds of features of dimensions 210, 540 and 19,800 (total feature vector), respectively. The results of Gabor features are shown in Figure 3.6 (b), where we can see that in all the total 18 cases, performance improvement is achieved in SRC with cohort normalization. Similarly for Gabor features, when there are few training samples, there is a significant improvement achieved by cohort normalization. For Gabor features of dimension 210, 540 and 12,000 (total feature vector), the most significant increases are respectively 8.93% with 2 training samples, 4.28% with 2 training images and 3.92% with 3 training images. In comparison, cohort normalization brings more performance improvement on intensity features than Gabor features on undersampled face recognition.

One sample per person face recognition. In this section, several experiments are performed to evaluate the performance of the proposed algorithm against various forms of degradation factors using a single training sample per subject. The first experiment is still conducted on the AR database. We still use the subset consisting of 50 male subjects and 50 female subjects. However, for each subject, we choose only 13 images from Session 1. As is known, there are not only illumination and expression variation but also facial disguises with sunglasses and scarves among the 13 images. Figure 3.7 shows the 13 images of one subject in this experiment. The images are cropped to 165 × 120 dimensions. 40 subjects are chosen for the development set D and another 40 subjects are chosen for the evaluation set E. For each subject, the single image with natural expression and illumination are used for training and the other 12 images for test. The remaining 20 subjects are considered as cohort models, i.e., the size of the cohort set is $H = 260$. Similarly to the experimental settings in undersampled face recognition, we select only $2 * s = 140$ cohort residuals for polynomial regression. About Gabor features, we still extract a 12,000 dimensional feature vector for each face image. We then select dimensions of 210 and 540 for both intensity and Gabor based randomfaces.

Table 3.1 lists the recognition accuracy for this experiment. Obviously, the accuracy increased by normalizing the raw residual with cohort information is considerable for both intensity and Gabor features. With intensity features of 3 dimensions 210, 540 and 19,800, cohort normalization makes the accuracy increased 20.41%, 19.16% and 14.79%, respectively. In comparison with intensity features, the improvement on Gabor features is much smaller. As shown in Table 3.1, the accuracy increased for the two Gabor randomfaces are 6.66% and 7.08%. Furthermore, when using the total Gabor feature (12,000-dimensional), the accuracy improves fractionally (0.63%). With single training sample under illumination and expression changes and occlusions, our proposal achieves 79.37% with intensity features and 87.08% with Gabor features.

To validate the robustness of our approach to various variations, we illustrate



(a) Intensity feature

(b) Gabor feature

Figure 3.6: The comparative recognition accuracy on the AR database as the number of training images per subject decreases. (a) Results of Intensity feature; (b) Results of Gabor feature.

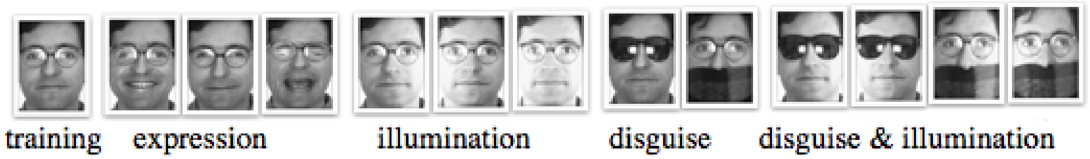


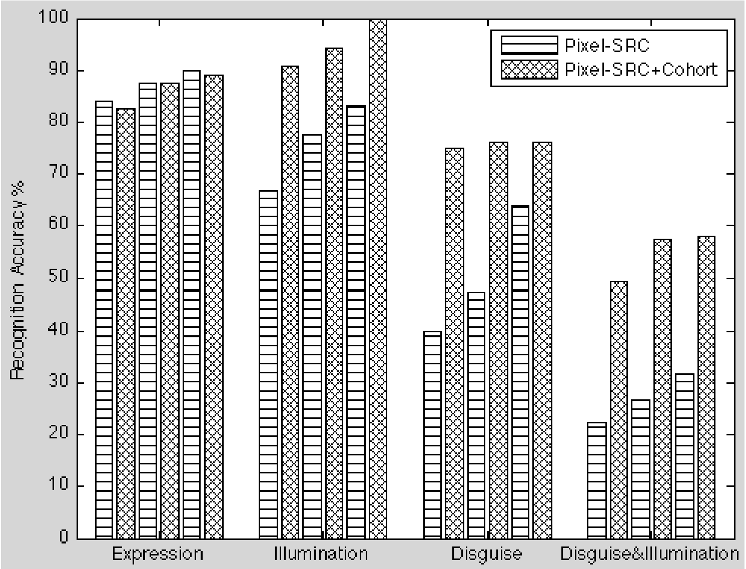
Figure 3.7: The 13 images of an individual on the AR database. The first image is used for training, while the rest 12 images with severe variation is for test.

Table 3.1: Comparative accuracy of SRC with and without cohort normalization on the AR database using single training sample per individual.

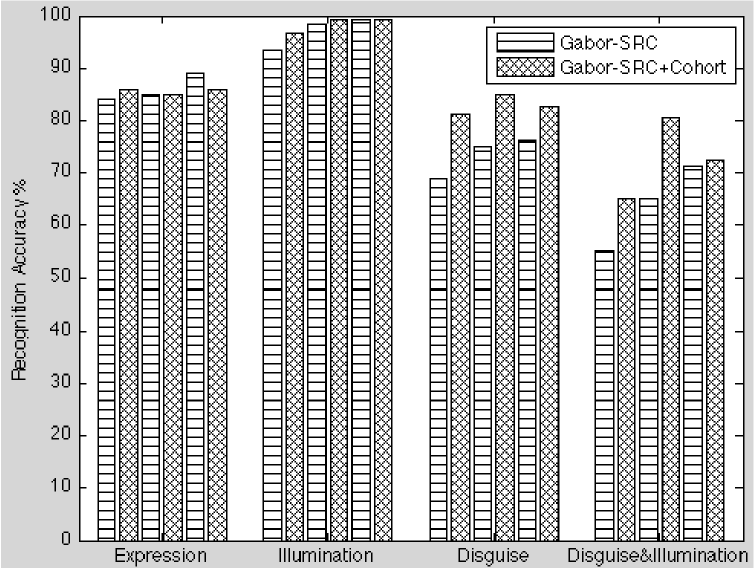
Feature	Accuracy-Intensity(%)		Accuracy-Gabor(%)	
	SRC	SRC+Cohort	SRC	SRC+Cohort
Randomfaces-210	51.88	72.29	74.17	80.83
Randomfaces-540	58.13	77.29	80.00	87.08
Total feature vector	64.58	79.37	83.54	84.17

the accuracy of test images with different kinds of variations. First, the 12 test images are separated into 4 groups with different facial variations as shown in Figure 3.7. The results are illustrated in Figure 3.8. For each of the 4 groups, there are totally 6 bars, which correspond to features of three dimensions with and without cohort normalization. For example in the "expression" group, the first bar illustrates the accuracy of 210-dimensional features without cohort normalization, while the second bar represents the accuracy of 210-dimensional features with cohort normalization. From this figure, we can see that although cohort normalization does not achieve significant improvement on expression variation, it does bring much robustness against illuminations changes and occlusions in SRC-based face recognition with single training sample, especially when using intensity features as facial representations. For example in "disguise" group, for the 3 intensity features, the accuracy increases 35%, 28.75% and 12.5%, respectively, while for the 3 Gabor features, the accuracy increases 12.5%, 10% and 6.25%.

The second experiment is performed on the FERET database [69]. We choose a subset from the total database, which includes 1,400 images from 200 subjects,



(a) Intensity feature



(b) Gabor feature

Figure 3.8: The accuracy of test samples with different variations on the AR database. (a) Results of Intensity feature; (b) Results of Gabor feature.

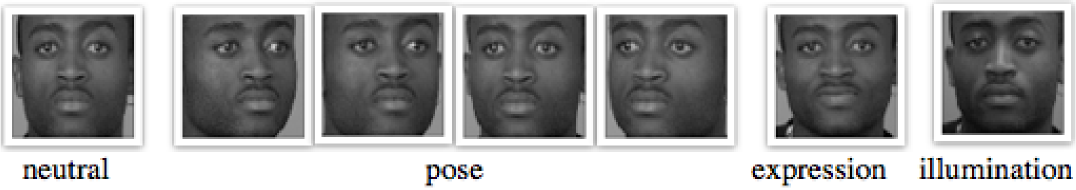


Figure 3.9: The seven images of one person on the FERET database.

each with 7 images. Each subject is composed of 7 images with pose, expression and illumination variations. In this experiment, each image has the size of 80×80. The seven images of one person are shown in Figure 3.9. We randomly select 85 subjects for both D and E. For each subject, the neutral image is used as the only training image, while the other six images are used for test. The remaining 30 subjects are used to build the cohort set. Hence, the size of the cohort set is $H = 210$. And only $2 * s = 70$ cohort residuals are finally used for polynomial regression. About Gabor features, we choose a family of Gabor kernels with five scales and eight orientations. The total Gabor feature of each image is a 4,000 dimensional vector generated by down-sampling the Gabor filter responses with a 8×8 uniform lattice. And dimensions of 210 and 540 for both intensity and Gabor based randomfaces are then calculated.

The recognition accuracy for this experiment is shown in Table 3.2. For all the six features, our approach performs better than the original SRC. When using intensity features, SRC with cohort normalization improves 10.2%, 11.18% and 8.82%, respectively. For the two Gabor based randomfaces, cohort normalization achieves significant improvement with 7.25% and 6.27%. However, the accuracy improves little when using the total Gabor feature, with only 3.33%. Note that there are pose variations in this database from -25 to +25. Next we illustrate the accuracy of test images with different variations. The 6 test images are first separated into 3 groups with different facial variations as shown in Figure 3.9. The results are illustrated in Figure 3.10. Similarly to the results on the AR database, not much improvement is seen in SRC with cohort normalization to expression variation. However, when there are pose and illumination variations, cohort normalization can really achieve some performance improvement. As shown in the

Table 3.2: Comparative accuracy of SRC with and without cohort normalization on the FERET database using single training sample per individual.

Feature	Accuracy-Intensity(%)		Accuracy-Gabor(%)	
	SRC	SRC+Cohort	SRC	SRC+Cohort
Randomfaces-210	30.98	41.18	47.65	54.90
Randomfaces-540	40.00	51.18	52.55	58.82
Total feature vector	46.67	55.49	58.24	61.57

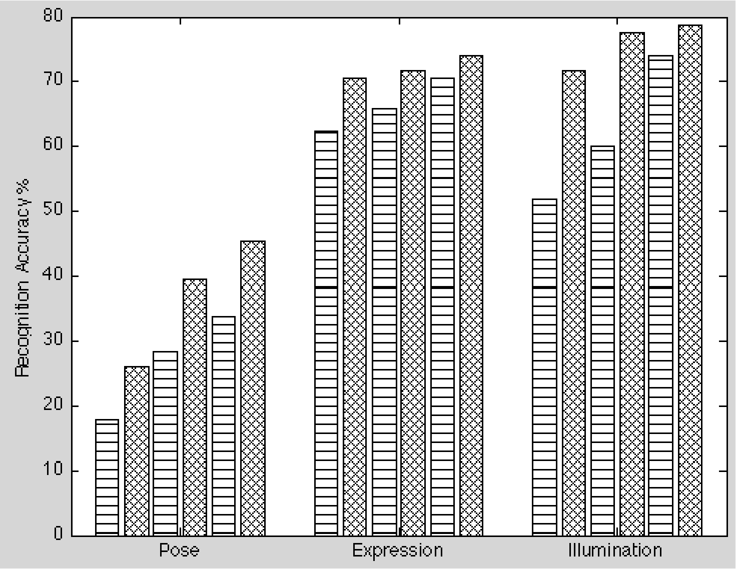
"pose" group, for the 3 intensity features, the accuracy increases 8.24%, 10.88% and 11.47% respectively, while for the 3 Gabor features, the accuracy increases 7.35%, 7.35% and 4.7%.

3.2.5 Conclusion

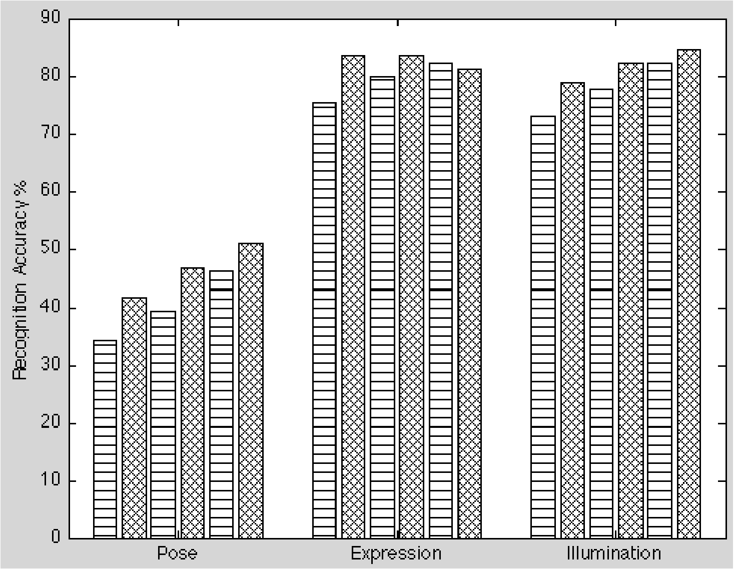
Summing up, in this work, cohort normalization was introduced to SRC-based undersampled face recognition to measure the degradation effect caused by facial variations, such as illumination variation, expression changes and various poses. Experiments performed on the AR and FERET databased clearly demonstrated that cohort normalization can bring SRC good generalization ability for undersampled face recognition, especially when there are illumination and pose variations and occlusions. Furthermore, it seems that SRC with cohort normalization performs much better than the original SRC as the number of training samples decreases.

3.3 Picture-Specific Cohort Score Normalization for Face Pair Matching

Face pair matching is the task of deciding whether or not two face images belong to the same person. This has been a very active and challenging topic recently due to the presence of various sources of variation in facial images, especially under



(a) Intensity feature



(b) Gabor feature

Figure 3.10: The accuracy of test samples with different variations on the FERET database. (a) Results of Intensity feature; (b) Results of Gabor feature.

unconstrained environment. We investigate cohort normalization that has been widely used in biometric verification as means to improve the robustness of face recognition under challenging environments to the face pair matching problem. Specifically, given a pair of images and an additional fixed cohort set, two picture-specific cohort score lists are computed and the correspondent score profiles of which are modeled by polynomial regression. The extracted regression coefficients are subsequently classified using a classifier. We advance the state-of-the-art in cohort normalization by providing a better understanding of the cohort behavior. In particular, we found that the choice of the cohort set had little impact on the generalization performance. Furthermore, the larger the size of the cohort set, the more stable the system performance becomes. Experiments performed on the LFW benchmark show that our system achieves performance that is comparable to the state-of-the-art methods.

3.3.1 Overview

Face pair matching is a relatively new task in face recognition. Given two pictures, the goal of this task is to determine whether the two face images represent the same individual. Note that in face pair matching, we do not have any identity information for each picture in the pair. This makes this task a very difficult problem. The LFW database is a specific benchmark for evaluating algorithms for face pair matching [30]. This database is designed for face recognition under uncontrolled environments. Faces in this database are collected from Yahoo News and have enormous variability. There has been significant work on the LFW in recent years. Among these approaches, many organize an additional background dataset and exploit either identity or discriminative information as prior knowledge from these background samples. Typically, the identities of the face images in the background set and those in the test sets are disjoint. By using a set of background samples, Wolf et al. [91, 92] proposed the one-shot similarity (OSS) and two-shot similarity (TSS) kernel to learn discriminative models exclusive to the vectors being compared. Kumar et al. [38] also employed an extra identity dataset for either building a set of attribute classifiers or simile classifiers. Note that the term “reference set” was used to represent the background set in this work. The main idea of "associate-predict" (AP) model [98] is to handle intra-personal variation using extra generic identities as a bridge. Taking advantage of a reference set of faces, which has no identity overlap with the test faces, Berg et al. learned a large set of Tom-vs-Pete classifiers and significantly improved on the published state-of-the-art [13].

As described in Section 3.1, using cohort information to improve the recognition performance of a biometric expert has become a popular way. The performance improvement achieved by making use of cohort information has been seen in fingerprint authentication [3] and multi-biometrics framework [4]. Though much work has been done in exploiting cohort information for biometric verification, little is known about employing cohort normalization for matching a pair of face images captured from news articles on the web, which can vary in all possible ways due to illumination, pose and misalignment. In this work, we are interested to find out whether or not merely post-processing raw matching scores using cohort normalization can achieve performance comparable to those reported by state-of-the-art approaches on face pair matching. In this section, we introduce cohort normalization to face pair matching in unconstrained environments.

Our proposed method can be summarised as follows. Given a pair of images (I^A, I^B) let us compute the similarity score between the image pair as sc^O. In order to exploit the cohort information, we will use an additional fixed cohort set, C. Each and every image in C is compared with each of the image in the pair. This procedure produces two picture-specific lists of cohort for I^A and I^B, respectively. We denote the two ordered list of cohorts as C^A and C^B. Both C^A and C^B are sorted variants of C. Cohort samples in C^A are sorted with respect to their closeness to I^B, whereas all the cohort samples in C^B are ranked with respect to their similarity to I^A. Next we compute two lists of cohort scores, sc^A and sc^B. sc^A contains scores between I^A and cohort samples in C^A, while scores in sc^B are those between I^B and C^B. Scores in each of sc^A and sc^B are then considered as discrete points on a function of rank orders. Next, polynomial regression is employed to approximate these two functions. Finally, we combine the polynomial coefficients extracted from both lists of cohort scores sc^A and sc^B, along with the original matching score, sc^O, in a classification framework, in order to improve the recognition performance.

Our contributions are as follows:

1) Proposal of picture-specific cohort normalization for face pair matching. For each picture in the pair of images being compared, we compute an exclusive cohort score list. Then we extract the discriminative patterns from the two cohort score lists to perform score normalization.

2) Better understanding of the behavior of cohort normalization. In particular, we address two important questions.

   - How much discriminative information is contained in the cohort samples alone? We will quantify this discriminative information empirically in terms of EER.
   - How do the choice and the size of cohort samples affect the performance of a cohort-based normalization procedure?

3) Comparison with the state-of-the-art methods. We will benchmark our proposed system against the state-of-the-art methods using the LFW database.

3.3.2 Picture-Specific Cohort Normalization

The general idea is to apply polynomial regression based cohort normalization to face pair matching under uncontrolled environments so that it can well normalize against score distribution variation between the two pictures presented in a given pair of images. Figure 3.11 illustrates our approach.

Picture-specific cohort selection. Let (I^A, I^B) denote the pair of images to be compared and sc^O be the raw matching score between picture I^A and I^B. Given an additional fixed cohort set $C = \{c_1, ..., c_h, ..., c_H\}$, in which there are totally H cohort samples, we need to select an exclusive cohort list for each of I^A and I^B. We denote the two picture-specific cohort lists as C^A and C^B respectively. Both C^A and C^B are sorted variants of C, the only difference among the three sets lies in the rank order of cohort samples.

Before the selection of C^A and C^B, a set of cohort scores between each picture of the pair and all the cohort samples in C are first computed. Cohort samples sorted with respect to their closeness to I^B constitute C^A. Let $\{c^A_1, ..., c^A_h, ..., c^A_H\}$

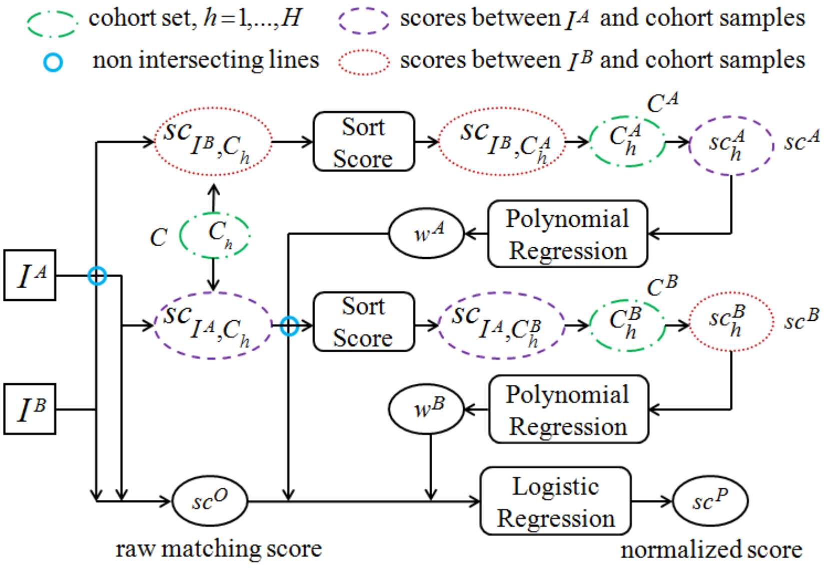


Figure 3.11: The framework of picture-specific cohort score normalization for face pair matching.

be the H sorted cohort samples in C^A, where, c_1^A is the most similar cohort sample to I^B, while c_H^A is the most dissimilar one. In the same way we can determine the cohort list for picture I^B, that is $\{c_1^B, ..., c_h^B, ..., c_H^B\}$, in which c_1^B is the most similar cohort sample to I^A. Next we can get two picture-specific cohort score lists $sc^A = \{sc_1^A, ..., sc_h^A, ..., sc_H^A\}$ and $sc^B = \{sc_1^B, ..., sc_h^B, ..., sc_H^B\}$. The H scores in sc^A are matching scores between I^A and each cohort sample in C^A, hence sc_1^A is the matching score between I^A and c_1^A. With the two cohort score lists sc^A and sc^B, next we focus on how to extract the discriminative patterns involved in them using polynomial regression.

Extraction of discriminative patterns using polynomial regression. Followed with the conclusion in [61] that sorted cohort scores of matching pairs (I^A and I^B are two pictures from the same person) and non-matching pairs (I^A and I^B are of different subjects) exhibit a discriminative pattern, these cohort scores are then considered as discrete points on a function of rank orders. With the two picture-specific cohort score lists sc^A and sc^B obtained above, we can get the following two functions of rank orders.

$$sc_h^A = f^A\left(h\right) \tag{3.12}$$

$$sc_h^B = f^B\left(h\right) \tag{3.13}$$

where $h = 1, 2, ..., H$. Next we employ polynomial regression to approximate the two functions as follows.

$$f^A\left(h\right) \approx w_n^A h^n + w_{n-1}^A h^{n-1} + ... + w_1^A h + w_0^A \tag{3.14}$$

$$f^B\left(h\right) \approx w_n^B h^n + w_{n-1}^B h^{n-1} + ... + w_1^B h + w_0^B \tag{3.15}$$

where $w^A = [w_0^A, w_1^A, ..., w_n^A]$ and $w^B = [w_0^B, w_1^B, ..., w_n^B]$ are the two approximated polynomial coefficient vectors. Further, the cohort scores in sc^A can be approximated by the $n+1$ coefficients in w^A, and sc^B can be roughly represented by w^B. Now we can use w^A and w^B to roughly represent the discriminative patterns included in sorted cohort scores.

Score normalization using logistic regression. Here we show how to normalize the original matching score sc^O using the discriminative patterns w^A and w^B extracted from sorted cohort scores. Each of the three components $\{sc^O, w^A, w^B\}$ contains different discriminative power and should be aggregated in a reasonable way. We can fuse these parameters by training a linear SVM or logistic regression classifier to obtain more discriminative weights on each parameter using a development dataset. As observed in [61], logistic regression classifier shows the superiority of fusion performance over SVM on face modality. Therefore, here we use the logistic regression classifier to approximate the final normalized score as follows:

$$sc^P = P\left(M \mid sc^O, w^A, w^B\right) \tag{3.16}$$

where $P\left(M \mid sc^O, w^A, w^B\right)$ represents the posterior probability of being a matching pair. To put it differently, the larger sc^P is, the more probable I^A and I^B

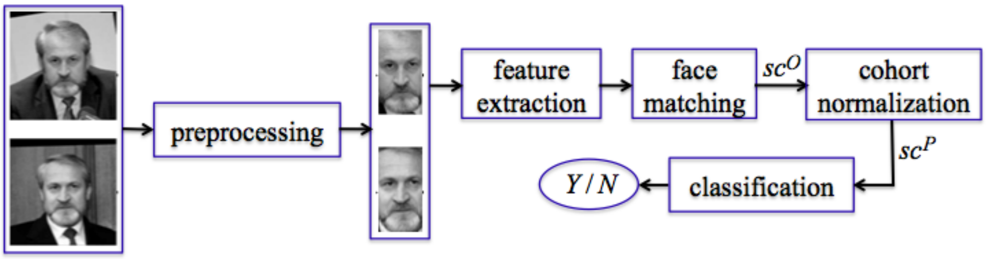


Figure 3.12: Overview of face pair matching process with cohort score normalization.

come from the same subject.

3.3.3 Application to Face Pair Matching

In this subsection, we show how the proposed picture-specific cohort score normalization can be applied to face pair matching on the LFW database in detail. As presented in Section 2.2.4, there are two evaluation settings provided by the authors of the LFW: image-restricted and unrestricted. For our experiments, we evaluate performance on View 2 under the restricted setting, which contains 6,000 image pairs. The image pairs are divided into ten splits in such a way that the proportion of matching and non-matching images are balanced (1:1 ratio). Therefore, each split has 600 image pairs. The performance of an algorithm is measured by a ten-fold cross-validation procedure. For more details, refer to [30]. There are three versions of the LFW available: original, funneled and aligned [91]. Here we use the aligned version in all experiments.

The framework of our approach is presented in Figure 3.12. There are mainly four steps involved: preprocessing, feature extraction, cohort normalization and classification.

Preprocessing. In the aligned version, all the images are of the same size 250 × 250 pixels. At the preprocessing step, we simply crop the image to remove the background, leaving only a face area of 150 × 80 pixels. Note that we do not apply any photometric approaches at this step.

Feature Extraction. For the feature extraction step, we carry out experiments on four facial descriptors: Intensity, Gabor wavelets [49], LBP [6] and SIFT [15].

The intensity feature vector of length $12,000$ is formed by concatenating all the pixels. For extracting LBP features, we divide each image into non-overlapping blocks of 10×10 size and extract 59-bin uniform LBP histograms for each block. And the length of the final LBP feature vector is $7,080$.

To compute the Gabor feature, we adopt five scales and eight orientations of the Gabor filters. The final Gabor feature vector is obtained by concatenating the responses at different pixels selected uniformly with a 10×10 down-sampling rate. The length of the Gabor feature vector is $4,800$. For the SIFT descriptor, we also divide the image into non-overlapping blocks of 16×16 size, and compute a 128D SIFT descriptor for each block. All descriptors are then concatenated to a single vector of length $5,760$.

Cohort Normalization. The methodology of cohort score normalization has been detailed in Section 3.3.2. Here we present only some points which are related to the experimental design. For the matching score, we compute both the Euclidian distance and the Hellinger distance between two descriptor vectors. As concluded in [61], the degree of polynomial has little impact on the generalization performance, for simplicity, we employ a linear function to fit the two cohort score functions $f^A(h)$ and $f^B(h)$, i.e., $n = 1$.

For logistic regression classifier, we use l_2-penalized logistic regression which leads to maximum likelihood estimate. To perform the final matching using the logistic regression classifier, a thresholds need to be determined. As is known, the output of a logistic regression classifier is the posterior probability of being a matching pair. Hence generally, 0.5 is set as the threshold. In this section, for most of the experiments, we use 0.5 as the threshold. Only in those experiments where EER is used for the performance evaluation measure, the threshold is the unique operating point where the FAR is the same as the FRR.

Classification. After cohort normalization, we can directly use the matching score for the final classification. As described above, the matching score is the posterior probability of being a matching pair. Thereby, two images with a matching score higher than the threshold are classified into the matching pair, i.e., both are from the same individual, and vice versa for those with matching score lower than the threshold.

3.3.4 Experiments and Results

Here, we show the experimental results on the LFW databse.

Results from different individual descriptors. Our first set of experiments is designed to test the classification accuracy increased by cohort score normalization with individual feature descriptors. For each of the ten folds of View 2 of the LFW dataset, we reserve one out of the nine splits as the cohort split, another split as the validation set, and the remaining eight splits as the training set. Note that in each of the ten experiments, the cohort splits are different from each other.

Recall that each split has 600 image pairs, or a total of 1,200 face images. In order to speed up the computation, we use only 600 randomly selected images from the cohort split (out of the $1,200$ available) to constitute the final cohort set for any of ten-fold experiments. It is worth nothing that, we use the term "cohort split" to represent the split where we choose cohort samples, while "cohort set" represents the final fixed cohort set for score normalization, i.e., the before mentioned C. For C^A and C^B, we call them "cohort lists", in which cohort samples are ordered.

The results are described in Table 3.3. Each type of features is tested with the original feature vector (Euclidean distance) and the square root of the feature vector (Hellinger distance). As shown in Table 3.3, cohort normalization improves about 6% over the Euclidean distance. With LBP descriptor, we can get an increased accuracy of almost 9.5%. Using the Hellinger distance improves the accuracy by about 5%. The highest accuracy we can get from a single type of feature is 77.38% using cohort normalization with the Hellinger distance of the SIFT descriptor.

Table 3.3 shows the absolute improvement introduced by cohort normalization. However, it would be more interesting if the relative improvement of a given system is compared to its respective baseline that is one without using cohort normalization. Since there are 8 independent experiments (due to 4 face descriptors and 2 distances), we shall summarize the results using the relative change of EER [61, 72] with respect to the performance of the baseline system. Here we employ EER as the performance measure due to its sensitivity to minute changes induced

Table 3.3: Comparative classification accuracy of different descriptors and distances with and without cohort normalization.

	Intensity	Gabor	LBP	SIFT
Euclidean (no cohort)	0.6502	0.6985	0.6500	0.7140
Euclidean (with cohort)	**0.6830**	**0.7560**	**0.7443**	**0.7703**
Hellinger (no cohort)	0.6497	0.7100	0.7132	0.7183
Hellinger (with cohort)	**0.6913**	**0.7680**	**0.7707**	**0.7738**

by cohort score normalization. The relative change of EER is given as follows:

$$\text{rel. change of EER} = \frac{\text{EER}_{cohort} - \text{EER}_{baseline}}{\text{EER}_{baseline}} \tag{3.17}$$

where EER_{cohort} is the EER of a given system with cohort normalization whereas $\text{EER}_{baseline}$ is the EER of the baseline (without cohort normalization). A negative change of EER implies an improvement over the baseline system. This statistic has the advantage that one can establish confidence intervals of the relative merit of our approach with respect to the baseline system. These confidence intervals can be conveniently visualized using a boxplot, where the median, the first and third quarter as well as the fifth and 95-th percentiles of the data can be visualized.

The relative changes of EERs for the above 8 individual experiments are illustrated in Figure 3.13. As can be observed, in all of the 8 experiments, our cohort normalization mechanism improves the baseline system significantly.

How much can cohort discriminative patterns alone help matching faces? In [61], the authors illustrated that the cohort score profiles of genuine and impostor claims exhibit a discriminative pattern. Therefore, as a sanity check, we also verify this behavior by using matching and non-matching image pairs on the LFW database. The experimental settings are similar to those in our first set of experiments (i.e., results from different individual descriptors). However, here we compute cohort scores for only the eight development splits not for

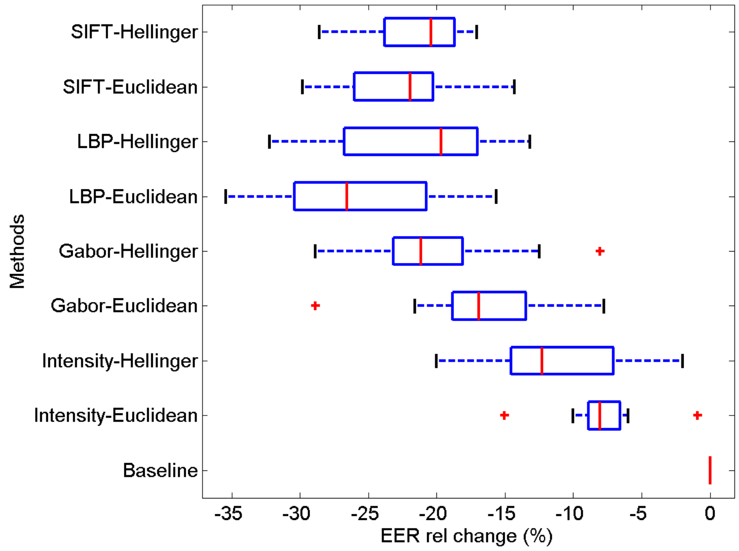
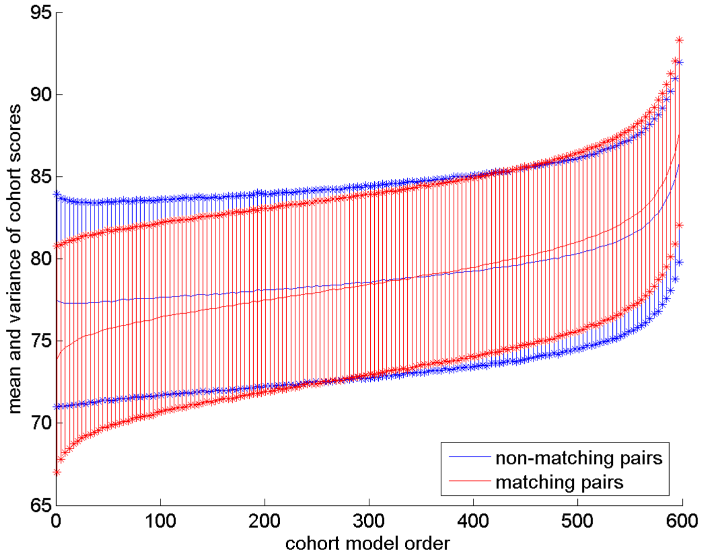


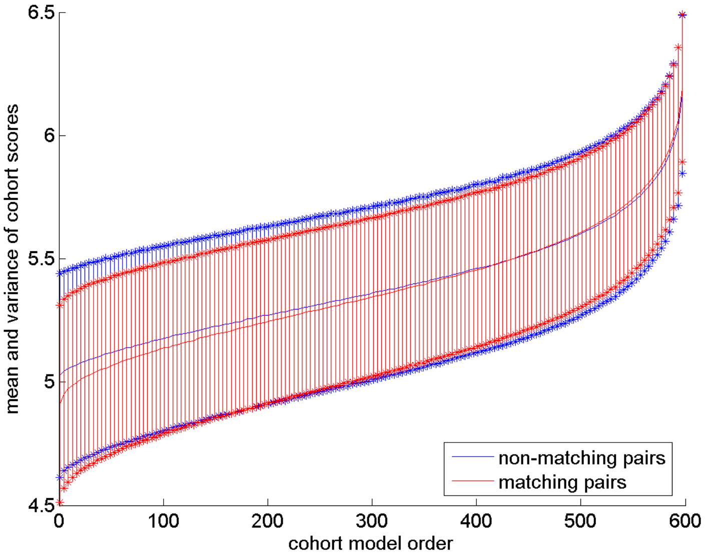
Figure 3.13: The boxplot of the relative change of EER with different individual descriptors.

the single evaluation split due to its limited number of pairs. For each pair, we can get two picture-specific cohort score profiles $f^A(h)$ and $f^B(h)$, each of which is a single vector of 600. Recall that the ordering of the cohort score profile for I^A is determined by I^B; and that of I^B is determined by I^A. Finally, we can get a total of $48,000$ $(= 2 \times 300 \times 8 \times 10)$ matching cohort score profiles and $48,000$ non-matching cohort score profiles. Next we respectively compute the mean and variance of matching and non-matching cohort score profiles. Figures 3.14 (a) and (b) show the distributions using Gabor and LBP features with Euclidean distance. We observe that the cohort score profiles of matched image pairs tend to have a higher gradient, i.e., the scores tend to increase with the rank order, whereas for non-matched image pairs, the profiles are generally flatter.

Therefore, we have verified that the cohort score profiles sorted by the reciprocal image in the image pair contain some discriminative information. Next, we quantify this discriminative information in order to find out how much the discriminative patterns from the cohort score profiles alone can help the image pair matching. We use the same experimental settings as those in our first set



(a) Gabor feature



(b) LBP feature

Figure 3.14: The distribution of cohort scores generated by ordered cohort samples for matching and non-matching pairs. (a) Results of Gabor feature; (b) Results of LBP feature.

Table 3.4: Comparative EERs of using the raw score and using cohort discriminative patterns alone (without the raw score) with four different feature descriptors and two distance metrics.

	Intensity	Gabor	LBP	SIFT
sc^O (Euclidean)	**0.3453**	0.3047	0.3477	0.2980
$w^A + w^B$ (Euclidean)	**0.3603**	0.3717	0.3793	0.3557
sc^O (Hellinger)	**0.3480**	0.3000	0.2963	0.2927
$w^A + w^B$ (Hellinger)	**0.3417**	0.3667	0.3587	0.3580

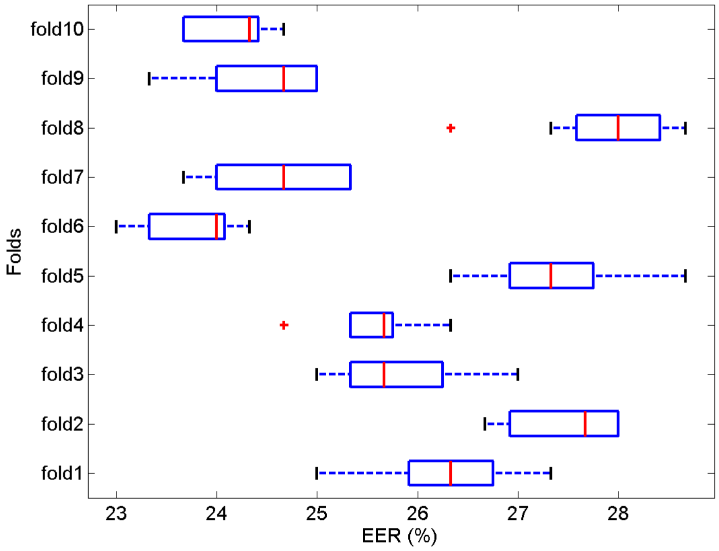
of experiments. The comparative EERs of using original matching scores and solely cohort discriminative patterns are shown in Table 3.4. "sc^O" rows are EERs obtained by using only the raw matching score, while "$w^A + w^B$" rows list EERs which are obtained by using only the discriminative patterns extracted from sorted cohort scores. To realize this, logistic regression is trained using only w^A and w^B without the raw score. With the Euclidean distance, using the discriminative patterns leads to EERs 4% higher than those obtained by using the raw matching score; while this value is 5% with the Hellinger distance. However, for the intensity feature, both the baseline and cohort scheme have very comparable EER. Indeed, the cohort scheme is slightly better in performance, having attained 34.17% compared to its baseline counter that has an EER of 34.80%.

How the choice and size of the cohort set affect the performance? In all of the above experiments, for each of the ten experiment folds, we have selected the cohort set randomly from a split. However, we do not know how the choice of any cohort split, as well as the size of cohort set, may have impact on the proposed cohort-based normalization procedure. For this reason, next we perform a set of experiments to evaluate the impact of these two design parameters.

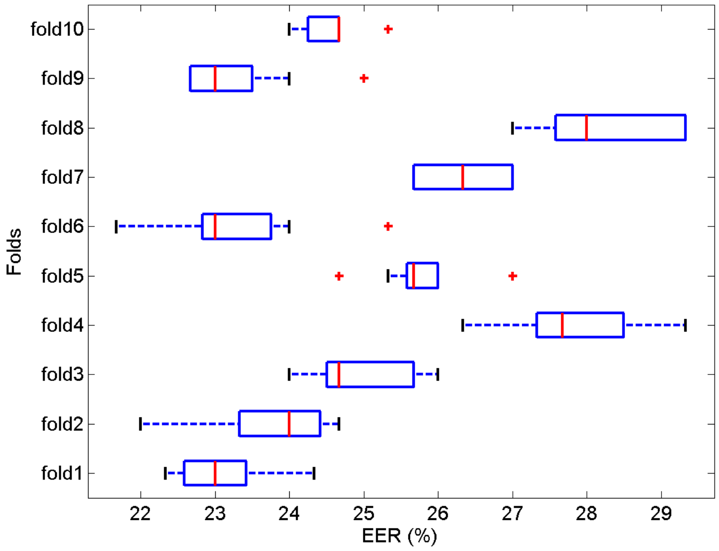
First, we use different splits as the cohort split but the size of the cohort set remains the same. Note that for each experiment fold, we can select one split from the nine training splits for the cohort split. In other words, we can perform each experiment fold nine times, each time using a different cohort split. With the cohort split ($1,200$ images), we still select only half images for the final cohort set (600 images). We use the boxplot of EERs to illustrate the impact caused by using different cohort sets for the ten folds. The results of Gabor and LBP descriptors using the Euclidean distance are shown in Figures 3.15. It can be observed that the choice of cohort set introduces an acceptable variation of approximately 1% EER on the system performance (as measured on the test split).

Next, we change the size of the cohort set. We do not consider all the ten experiment folds this time but keep to only one fold due to the high computational cost, that is, we test only on one split. Denote the number of images in the cohort split as M (here $M = 1,200$). We select only m images for the cohort set. Let the value of m vary from 100 to 900. This is a "m choose M" problem and is a combinatory problem. There are many choices of "m choose M". For each value of m, we run 100 random samplings, then compute the mean and variance of the total 100 EERs. Figure 3.16 shows the results of Gabor and LBP descriptors using the Euclidean distance. The solid lines represent the mean of EERs, while the dashed lines represent its variance. As shown in this figure, the larger the size of the cohort set, the lower the EER. At the same time, the variance of the EER will also decrease. In other words, with more cohort samples, one can get a more stable result.

Comparison with the state-of-the-art methods. The high variability of face images in the LFW database makes identity verification a very challenging task. Many interesting approaches have been proposed to improve the performance of face pair matching. In [91], OSS, TSS and the ranking descriptor were proposed to facilitate better recognition performance. Guillaumin et al. developed a logistic discriminant approach which learns the metric from a set of labelled image pairs (LDML) [28]. In this section, we compare our cohort normalized score with these state-of-art approaches. Due to diverse information used by different algorithms (fusion of different descriptors as well as different metrics), it is more reasonable to make comparison between the same individual descriptors and also the same metric. Here, we choose OSS (the best result of [91]) and LDML as the control algorithms.



(a) Gabor feature



(b) LBP feature

Figure 3.15: The boxplot of the relative change of EER with different choices of the cohort set. (a) Results of Gabor feature; (b) Results of LBP feature.

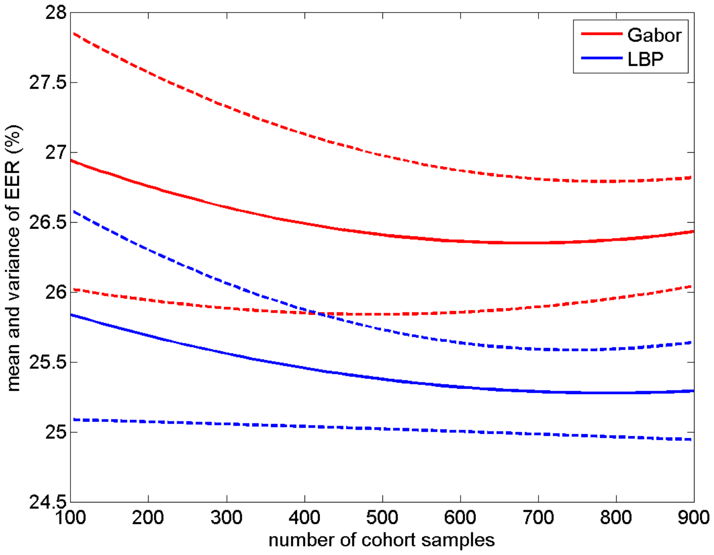


Figure 3.16: The mean and variance of EERs as the number of cohort samples increased.

Table 3.5 shows the comparative results on the image-restricted training benchmark ("View 2'). Note that here we show only several descriptors since only the result of these descriptors are provided in the original literature [28, 91]. In comparison with OSS, cohort score normalization with Gabor and SIFT features outperforms OSS when using either Euclidean or Hellinger as the distance measure. With LBP feature, our approach is slightly worse than OSS. When comparing with LDML using SIFT feature, the accuracy obtained by LDML with PCA of dimension 35 is comparable with cohort score normalization. However, when the dimension of PCA turns 55, the performance of LDML decreases notably, and is much worse than our results.

3.3.5 Conclusion

We introduced picture-specific cohort score normalization to pair matching of faces captured from news articles on the web. With the proposed framework, we showed that the cohort information alone has a certain discrimination power

Table 3.5: Comparative classification accuracy on the image-restricted training benchmark ("View 2").

Algorithms	Euclidean	Hellinger
Gabor(C1)-OSS	0.7396	0.7437
LBP-OSS	**0.7663**	**0.7820**
SIFT-OSS	0.7576	0.7597
SIFT-LDML-PCA(35)	0.7660	**0.7750**
SIFT-LDML-PCA(55)	0.7280	0.7280
Gabor-Cohort	**0.7560**	**0.7680**
LBP-Cohort	0.7443	0.7707
SIFT-Cohort	**0.7703**	**0.7738**

which is just marginally worse than the raw matching score. When this information is properly extracted, i.e., using polynomial regression in our case; and is appropriately combined with the raw matching scores, i.e., using logistic regression, we nearly always observe an improvement in the system performance over the baseline system that relies on the raw matching scores.

In addition to the above finding, we have also improved our understanding on the behavior of cohort normalization. In particular, we found that the choice of the cohort set had little impact on the generalization performance. The larger the size of the cohort set, however, the more stable the result. We validated our approach on the LFW dataset and achieved competitive matching performance.

3.4 On the Understanding of Discriminative Cohort Behavior for Face Recognition

Cohort score normalization has received much attention in various biometric systems. Our above presented work has successfully applied this technique to a face biometric expert. However, there still lacks the understanding of cohort normalization. How does the quality of cohort samples affect the performance of a face recognition system? Should we use a cohort set containing most possible subjects (each subject with fewest possible samples) or a cohort set including fewest possible subjects (each subject with most possible samples)? Which one is better for a face recognition system? All these cohort behaviors need to be given a better understanding. Although, in Section 3.3.4, we have presented an initial work on this, more efforts are desired to be made in this direction. To give proper answers to the above mentioned questions, in this section, we conduct an extensive experimental exploration on the FRGC ver2.0 database in a face verification scenario. We employ polynomial regression based cohort normalization method to explore the discriminative cohort behavior involved in cohort samples. From the experimental results, we found that cohort samples with different quality indeed produce different cohort normalization performance. Furthermore, the results have shown that the selection of a cohort set, which includes only cohort samples with good quality, between as many subjects as possible and as few subjects as possible has little impact on the generalization performance.

3.4.1 Overview

Numerous methods have been developed for face recognition with impressive performance. In recent years, we have witnessed more and more research efforts on face recognition under uncontrolled settings [39, 63, 71, 93]. Among the various algorithms, organizing an additional face database and then trying to extract some useful information from this extra database has become more and more popular as a way of making the current face recognition systems more robust to various challenging problems.

To address the problem of recognizing a face from its partial image, Liao et al. [47] proposed an alignment-free sparse representation approach which represents each face image with a set of keypoint descriptors and constructs a large dictionary from all the gallery descriptors. In this way, any test face image, holistic or partial, can be sparsely represented by the dictionary. The gallery descriptors used in this approach are extracted from a set of background face images together with one of the two images being compared. Note that the set of background face images does not contain the same subject as either of the two images being compared. In [44], an extra training set was organized to build a Gaussian Mixture Model (GMM) [29] on the spatial-appearance features. Each Gaussian component builds correspondence of a pair of features to be matched between two face images being compared. The proposed probabilistic elastic matching algorithm achieved superior performance over state-of-the-art methods for pose variant face verification. In defense of sparsity based face recognition [22], the authors tried to construct an intra-class variation matrix from an extra training set (containing images from the FRGC ver2.0 database [70]) for a "prototype plus variation" representation model for sparsity based face recognition. In the proposed Superposed SRC, the dictionary was assembled by the class centroids (i.e., training samples of the gallery subjects) and the sample-to-centroid differences (i.e., the constructed intra-class variation matrix). The Superposed SRC led to a substantial improvement on SRC when dealing with problems like face recognition from a single image per person. Aggarwal et al. [5] presented a part-wise facial characterization based approach using sparse representation technique to address the challenges involved in automatic matching of faces across plastic surgery. The presented work utilized face images from sequestered non-gallery subjects with similar local facial characteristics to fulfill the requirement of multiple training images per gallery subject in SRC.

As presented in Section 3.3.1, there are also many approaches, which are developed for recognizing faces collected from news articles in the web (e.g., the LFW database [30]), organizing an additional background dataset and exploit either identity or discriminative information as prior knowledge from these background samples. For comparing two faces under significantly different settings, Yin et al. proposed to "associate" one input face with alike identities from an extra generic identity dataset. With the associated faces, the authors can then "predict" the appearance of one input face under the setting of another input face, or discriminatively "predict" the likelihood whether the two input faces are from the same person or not. The proposed "Associate-Predict" model can well handle face recognition with large intra-personal variation. With a set of background samples, Wolf. et al defined several similarity functions to learn discriminative models exclusive to the two faces being compared: One-Shot similarity (by answering the question "whether the first face looks more similar to the second face than to any face from the background set?"), Two-Shot similarity (by answering the question "are both faces well separated from the background faces?") and ranking similarity ("do these two faces have similar sets of neighboring samples in the background set?").

Though much work has been done in exploiting useful information from an extra background dataset to improve the face recognition performance, little is known about the understanding of the background behavior. For example, in most literature, the authors randomly select a set of face images from one/more known face databases to form the background set. To our best knowledge, there is no literature working on how to organize a good background set for face recognition. Are there any suggestions we can follow when we organize a background set? What impact will be brought about if background samples of different qualities are employed? Should we use a cohort set containing most possible subjects (each subject with fewest possible samples) or a cohort set including fewest possible subjects (each subject with most possible samples)? Which one is better for a face recognition system? All these background behaviors need to be given a better understanding. Although, in Section 3.3.4, we have presented some initial work on this, more efforts are desired to be made in this direction.

As is known, using cohort information to improve the recognition performance of a biometric expert has become a popular way. The performance improvement achieved by making use of cohort information has been seen in fingerprint authentication [3] and multi-biometrics framework [4]. Our work in Sections 3.2 and 3.3 has successfully applied cohort score normalization to undersampled face recogntion and unconstrained face pair matching. In this section, we focus on face verification and aim to achieve a proper understanding of cohort behavior on face recognition. For the cohort normalization method, we use polynomial regression based cohort normalization to extract discriminative information from the cohort set, thus exploring the discriminative cohort behavior involved in cohort samples. Specifically, we are interested to answer the following two questions. The first one is "How does the quality of cohort samples affect the performance of a face recognition system?", while the second one is "Should we use a cohort set containing as many subjects as possible (each subject with as few samples as possible) or a cohort set including as few subjects as possible (each subject with as many samples as possible)?".

In our face verification task, we assume all the template models are acquired under controlled conditions, thus having a good quality. While for the query samples, we consider those acquired under both controlled and uncontrolled conditions. In other words, there are two types of query samples: those with good quality and those with bad quality. To get proper answers to the above proposed two questions, we conduct an extensive experimental exploration on the FRGC ver2.0 database [70]. From the experimental results, we found that cohort samples with different quality indeed produce different cohort normalization performance. Furthermore, the results have shown that the selection of a cohort set, which includes only cohort samples with good quality, between as many subjects as possible and as few subjects as possible has little impact on the generalization performance.

3.4.2 Experimental Settings

Experimental design. To find out the effect of cohort's quality on the generalization performance, we organize three sets of cohort samples: good, bad and both cohort sets. In the good cohort set, all the cohort face images are captured in well controlled conditions. That is all the cohort samples are of good quality. While in the bad cohort set, cohort samples are of bad quality and acquired under uncontrolled environments. For the cohort samples in the both cohort set, we include cohort samples with both good and bad quality. One important thing for organizing the three cohort sets is to control the number of cohort samples in the three sets. As concluded in Section 3.3, bigger cohort set size gives more stable and often better results to a point before the performance saturates. Hence, to make a fair comparison among the three cohort sets, we need to have the same number of cohort samples in them.

As is known, in a typical face verification system, there are usually two stages involved: the enrollment stage (building a template model for each user) and the test stage (validating the authenticity of a query sample to its claimed identity). During the test stage, a query sample is compared to its claimed template model. Generally, when building a template model for each user in the enrollment stage, it is difficult to get a template model which captures all possible intra-personal variations. Therefore, in our following experiments, we assume all the template

Table 3.6: The 8 combinations of target, query and cohort samples.

Cohort condition	Good query	Bad query
Without cohort	TQgood	TQbad
Good cohort	TQgoodCgood	TQbadCgood
Bad cohort	TQgoodCbad	TQbadCbad
Both cohort	TQgoodCboth	TQbadCboth

models are acquired in well controlled conditions, i.e., with good quality. For the query sample, it is unknown whether it is obtained in controlled conditions or not. Hence, we consider two kinds of query samples: good quality and bad quality.

If we use "T" to represent template models, "Q" to represent queries and "C" to denote cohort samples, we have a total of 8 combinations among template models, queries and cohort samples as shown in Table 3.6. "TQgood" is the direct comparison between the target and the query with good quality (without cohort score normalization). "TQgoodCgood" represents using cohort samples of good quality to normalize the raw matching score between the target and the good query.

Face database. We use the FRGC ver2.0 database [70] for our experimental evaluation. In this database, the authors provided the researchers with 6 experiments. For our experiments, we choose face images from the Experiment 4. This experiment is designed to measure progress on recognition from uncontrolled frontal still images. The target set consists of single controlled still images, and the query set consists of single uncontrolled still images. As presented in Section 2.2.3, there are a total of 16,028 controlled images and 8,014 uncontrolled images included in Experiment 4. These images are from 466 subjects. For our experiments, we choose 465 subjects to perform a 5-fold experiments, thus each fold contains $465 \div 5 = 93$ different subjects. Finally, we have a total of 15,988 controlled images and 7,994 uncontrolled images. For each fold, we list the number of controlled images together with that of uncontrolled images in Table 3.7. For each of the 5-fold experiments, we choose one fold for the final evaluation, one fold for selecting cohort samples, and the other three folds for training classifiers.

Table 3.7: The numbers of controlled and uncontrolled images in the 5 folds.

Fold No	1	2	3	4	5
# controlled images	2,780	3,424	3,264	2,928	3,592
# uncontrolled images	1,390	1,712	1,632	1,464	1,796

In this way, the identities in the evaluation, development and cohort sets are disjoint from each other. Furthermore, in each of the 5 experiments, the cohort folds are different from one another.

When a fold is chosen for evaluation or development use, we need to determine which images are used for the target models, which are for the good queries and which are used for the bad queries. As listed in Table 3.7, for each fold, the number of controlled images is twice that of uncontrolled images. Actually, for each subject in the total 93 subjects of one fold, the number of controlled images is twice that of uncontrolled ones. Hence, to group all the images in one fold into the target, good query and bad query sets, we simply use all the uncontrolled images for the bad query set. For the controlled images of each of the 93 subjects, we choose half of them for the target set, while the other half for the good query set. In this way, the numbers of images in the target, good query and bad query sets are all the same as listed in Table 3.8. We follow the protocol developed by authors of the FRGC ver2.0 database for our experiments. That is, each query sample is compared with all those target models in the target set. We list the total number of matching scores in each fold together with numbers of genuine scores and impostor scores in Table 3.8. It is worth nothing that, for each fold, we have exactly the same genuine matches and impostor matches between "TQgood" and "TQbad". The only difference between "TQgood" and "TQbad" in our experiments lies in the quality of query samples, thus providing a fair comparison between the two situations. Similarly, "TQgoodCgood" / "TQgoodCbad" / "TQgoodCboth" and "TQbadCgood" / "TQbadCbad" / "TQbadCboth" include the same genuine matches and impostor matches. Finally, the large numbers of both genuine and impostor comparisons guarantee the reliability of our experimental results.

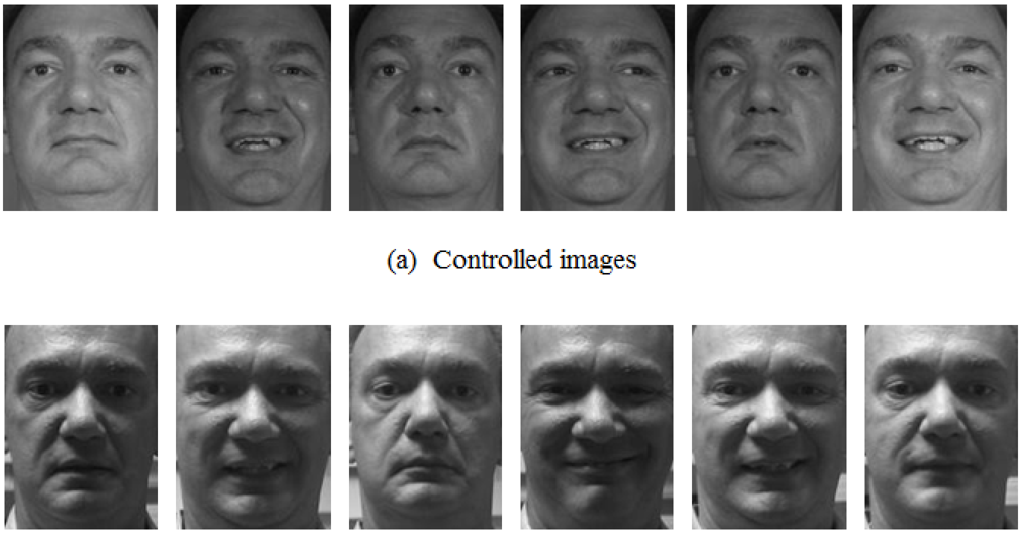
If a fold is chosen for the cohort use, we need to select images from this fold

Table 3.8: The verification configuration of the 5 folds.

Fold No	1	2	3	4	5
# target models	1,390	1,712	1,632	1,464	1,796
# good queries	1,390	1,712	1,632	1,464	1,796
# bad queries	1,390	1,712	1,632	1,464	1,796
# total matches	1,932,100	2,930,944	2,663,424	2,143,296	3,225,616
# genuine matches	32,092	44,608	41,048	36,464	49,064
# impostor matches	1,900,008	2,886,336	2,622,376	2,106,832	3,176,552

for the good, bad and both cohort sets. As concluded in Section 3.3, the size of the cohort set has an impact on the cohort normalization performance, i.e., bigger cohort set size gives more stable and often better results to a point before the performance saturates. Hence, in our experiments, we choose the same numbers of cohort samples for the three cohort sets. First, we separate the whole fold into three partitions as the above target, good query and bad query sets. Next, we randomly select 700 images from the good query set to construct the good cohort set. Similarly, we select the same 700 images from the bad query set for the bad cohort set. Note that the difference between the 700 images in the good and bad cohort sets lies in only the quality. In other words, 700 good cohort samples and 700 bad cohort samples are from the same subjects, and each subject contains the same number of good and bad cohort samples. Finally, we choose half images from the good cohort set and half images from the bad cohort set to construct the both cohort set. In this way, all the good, bad and both cohort sets include 700 images.

Feature extraction and cohort normalization. Before performing feature extraction, all the images are geometrically normalized to a fixed size. With the provided coordinates of the four eye corners, we compute the coordinates of the two eye centers. After the location of eye centers, the distance between the eye centers is set as 60 pixels. Finally, all the images are normalized to the size of 110 × 80. Several normalized examples are shown in Figure 3.17. For feature extraction, we work on Gabor [49]and LBP [6] features. For calculating the Gabor feature, we adopt five scales and eight orientations of the Gabor filters.



(a) Controlled images

(b) Uncontrolled images

Figure 3.17: Some aligned images from the FRGC ver2.0 database. (a) images with good quality; (b) images with bad quality.

The final Gabor feature vector is obtained by concatenating the responses at different pixels selected uniformly with a 10×10 down-sampling rate. To extract LBP features, we divide each image into non-overlapping blocks of 10×10 size and extract 59-bin uniform LBP histograms for each block. For the matching score, we compute the cosine similarity between two descriptor vectors.

For cohort normalization, we employ polynomial regression based cohort normalization to extract discriminative information from cohort samples [61]. Specifically, we employ a linear function to fit the cohort function $f(h)$ (i.e., Eq 3.4). For logistic regression classifier, we use l_2-penalized logistic regression which leads to maximum likelihood estimate. To perform the final matching using the logistic regression classifier, a threshold needs to be determined. Here, we use the unique operating point where the FAR is equal to the FRR as the threshold.

Table 3.9: The mean EERs of the 5-fold experiment on face verification with three different cohort conditions.

Feature	Query	Czero	Cgood	Cbad	Cboth
Gabor	Qgood	11.23%	5.86%	8.53%	7.00%
	Qbad	28.67%	22.45%	26.58%	21.22%
LBP	Qgood	7.46%	4.61%	5.68%	4.97%
	Qbad	31.85%	23.30%	28.50%	22.80%

3.4.3 Experimental Results

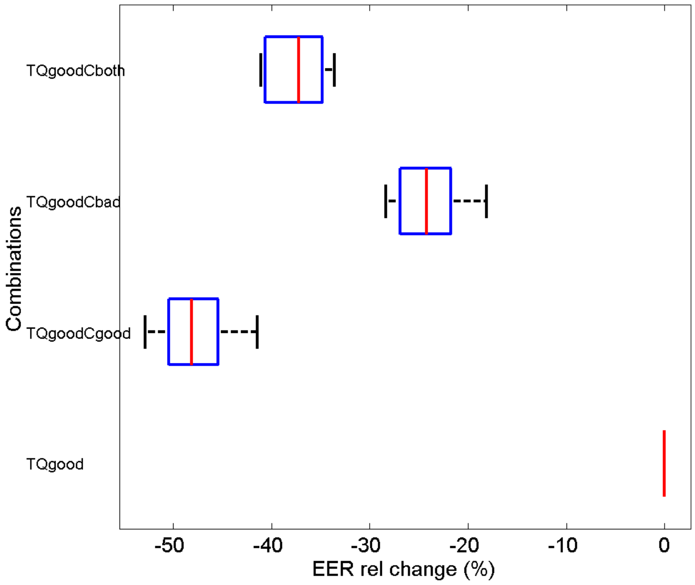
Next, we show the results of the experiments performed on the FRGC ver2.0 database, thus giving suitable answers to the two questions proposed in Section 3.4.1.

Question 1. As described above, our experiment is a 5-fold experiment. We summarize the mean EERs of the 5 experiments in Table 3.9. We use "Czero" to represent the baseline system without cohort score normalization, i.e., the systems "TQgood" and "TQbad" in Table 3.6. Obviously, systems with cohort score normalization, no matter using what kind of cohort samples, outperform the baseline systems. Interestingly, we found that, for cohort score normalization, the quality of cohort samples indeed matters. For recognizing query samples of good quality, using cohort samples of good quality achieves the best performance. For matching bad queries, using "Cgood" and "Cboth" obtain the similar cohort normalization performance. With bad cohort samples, cohort normalization achieves much worse performance than with good cohort samples. For example, for Gabor and LBP features, using "Cbad" leads to 4.13% and 5.20% higher EERs than using "Cgood". As shown in Table 3.9, with good cohort samples, "TQgoodCgood" gets 5.37% and 2.85% lower EERs than the baseline system "TQgood" for Gabor and LBP features. The figures for the comparison between "TQbadCgood" and "TQbad" are 6.22% and 8.55%.

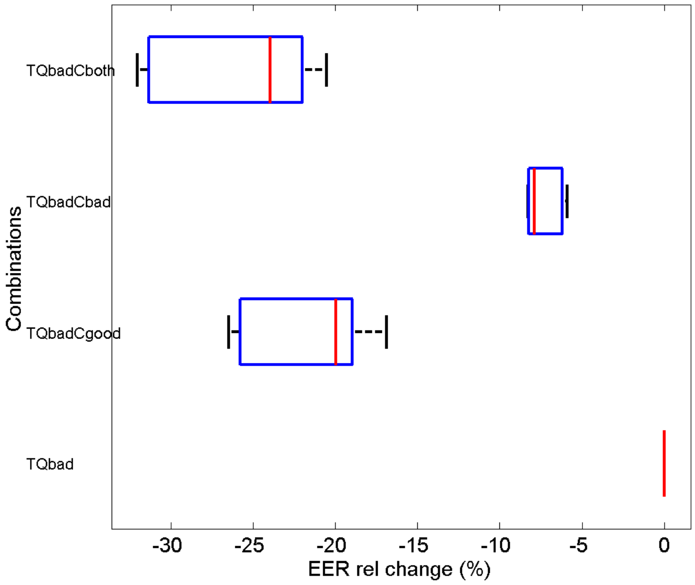
Table 3.9 shows the absolute improvement introduced by cohort normalization. We also illustrate the relative improvement of a given system in comparison with its respective baseline that is one without using cohort normalization. The relative change of EER is shown in Figure 3.18 for Gabor features and Figure

3.19 for LBP features. As can be observed, for matching a good query using either Gabor or LBP, cohort normalization with good cohort samples leads to the best performance. While when recognizing queries of bad quality, the system "TQgoodCboth" can give the best results. However, the system "TQgoodCgood" achieves comparable performance to "TQgoodCboth". Finally, we illustrate the cohort score profiles of genuine and impostor claims generated by ordered cohort samples. Here, we compute the cohort score profiles for only Fold 5 due to its maximum number of matchings. Note that each cohort score profile is a single vector of 700. For genuine cohort score profiles, we compute for all the 49,064 genuine matchings. While for impostor, we randomly select only 49,634 from the total 3,176,552 impostor matchings. In this way, we can get a total of 49,064 cohort score profiles for genuine claims and 49,634 for impostor claims. Figure 3.20 shows the distributions of "TQgoodCgood" and "TQbadCgood" using LBP as features. The results of using "Cbad" and "Cboth" with LBP features are illustrated in Figure 3.21 and Figure 3.22.

Question 2. As shown in the above experiments, cohort normalization with good cohort samples can lead to an obvious improvement to the baseline system. Hence, in the following experiments, we use only good cohort samples for polynomial regression based cohort normalization. To answer the second question proposed in Section 3.4.1, we organize two different cohort sets. The first cohort set contains as many subjects as possible, each subject with fewest possible samples. While there are fewest possible subjects, each subject with as many samples as possible, in the second cohort set. We use the same 5 folds as those used in the former experiments. If a fold is chosen for the cohort use, we select only cohort samples captured under controlled environments for constituting our two cohort sets. The 6 combinations of target, query and cohort samples are listed in Table 3.10. Here, we use "Cgood1" to represent the first cohort set and "Cgood2" for the second cohort set. In comparison with the systems in Table 3.6, only the cohort conditions are changed. The other experimental settings remain fixed. Thus, the verification configuration of the 5 folds remain the same to that listed in Table 3.8. We describe the cohort configuration of the 5 folds in Table 3.11. For each fold, the numbers of cohort samples of both "Cgood1" and "Cgood2" are the same. In this table, we list not only the number of subjects in each cohort set,

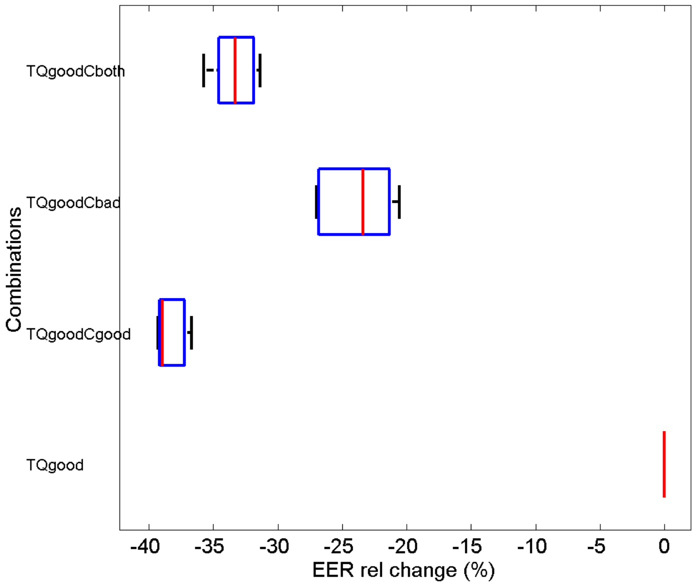


(a) Queries of good quality

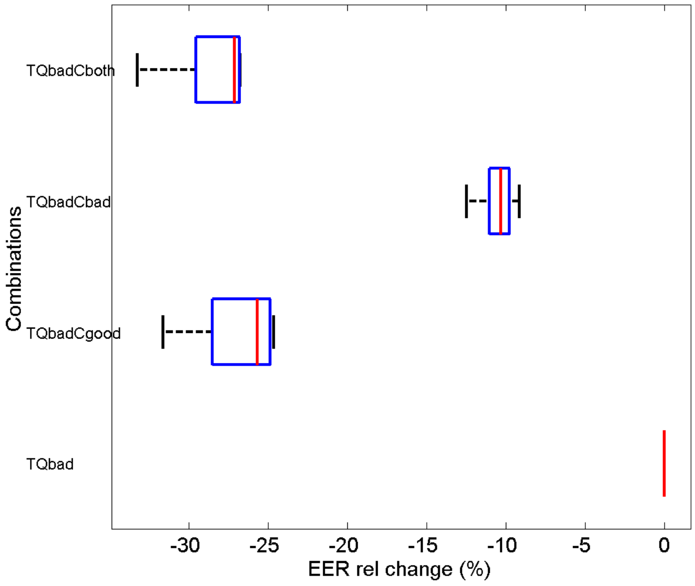


(b) Queries of bad quality

Figure 3.18: The boxplot of the relative change of EER on the 5-fold face verification experiment with {"Cgood", "Cbad", "Cboth"} using Gabor feature. (a) Results of good queries; (b) Results of bad queries.

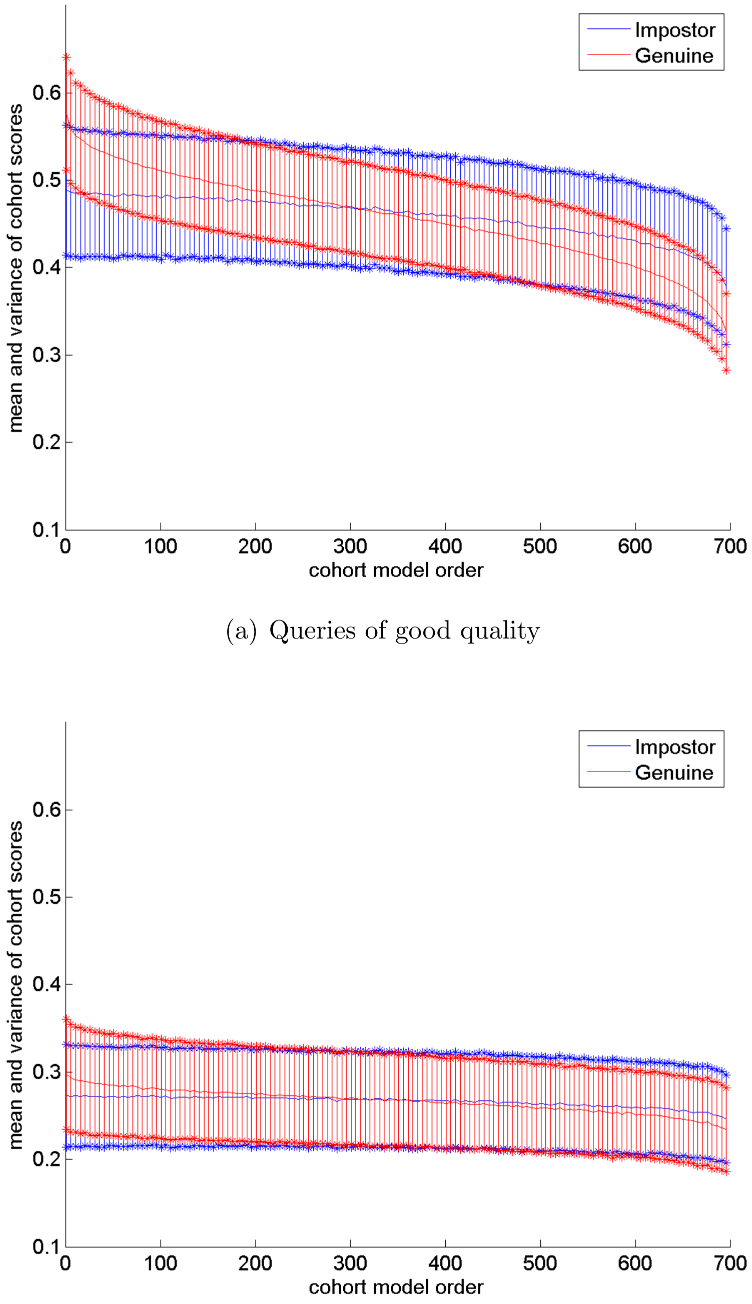


(a) Queries of good quality



(b) Queries of bad quality

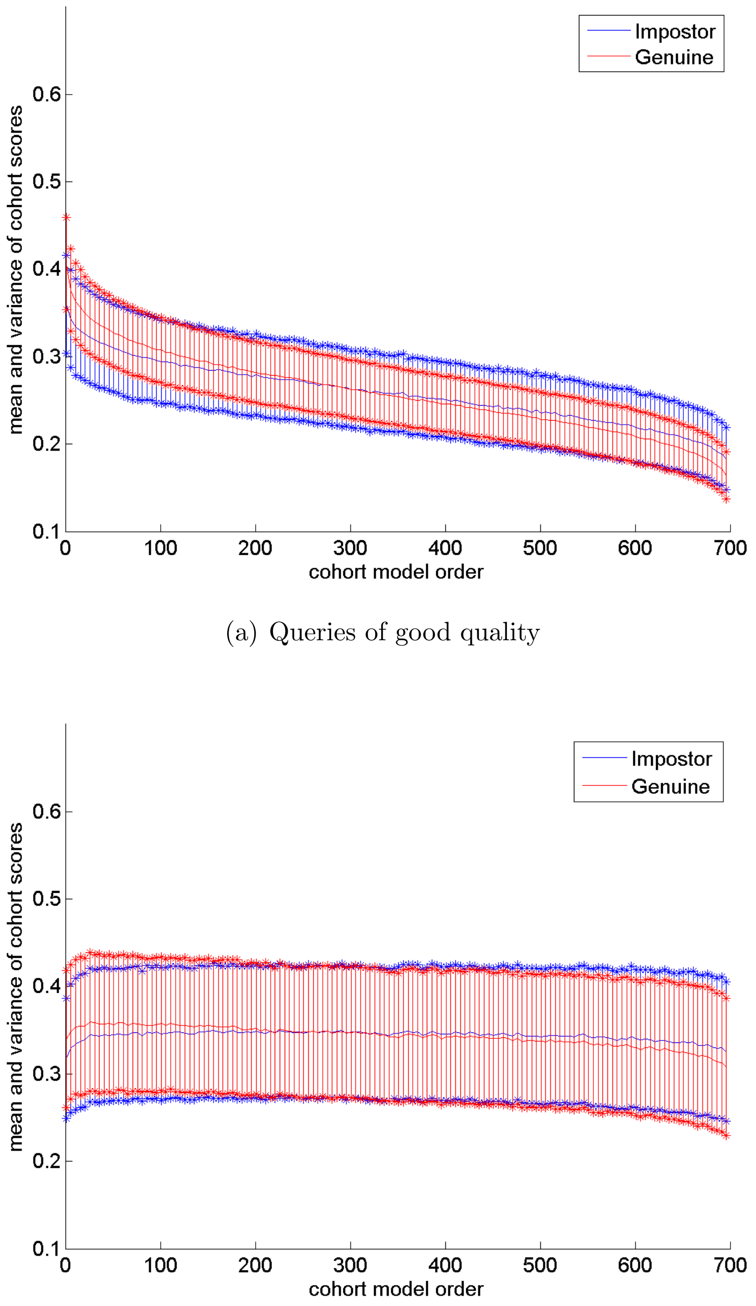
Figure 3.19: The boxplot of the relative change of EER on the 5-fold face verification experiment with {“Cgood”, “Cbad”, “Cboth”} using LBP feature.(a) Results of good queries; (b) Results of bad queries.



(a) Queries of good quality

(b) Queries of bad quality

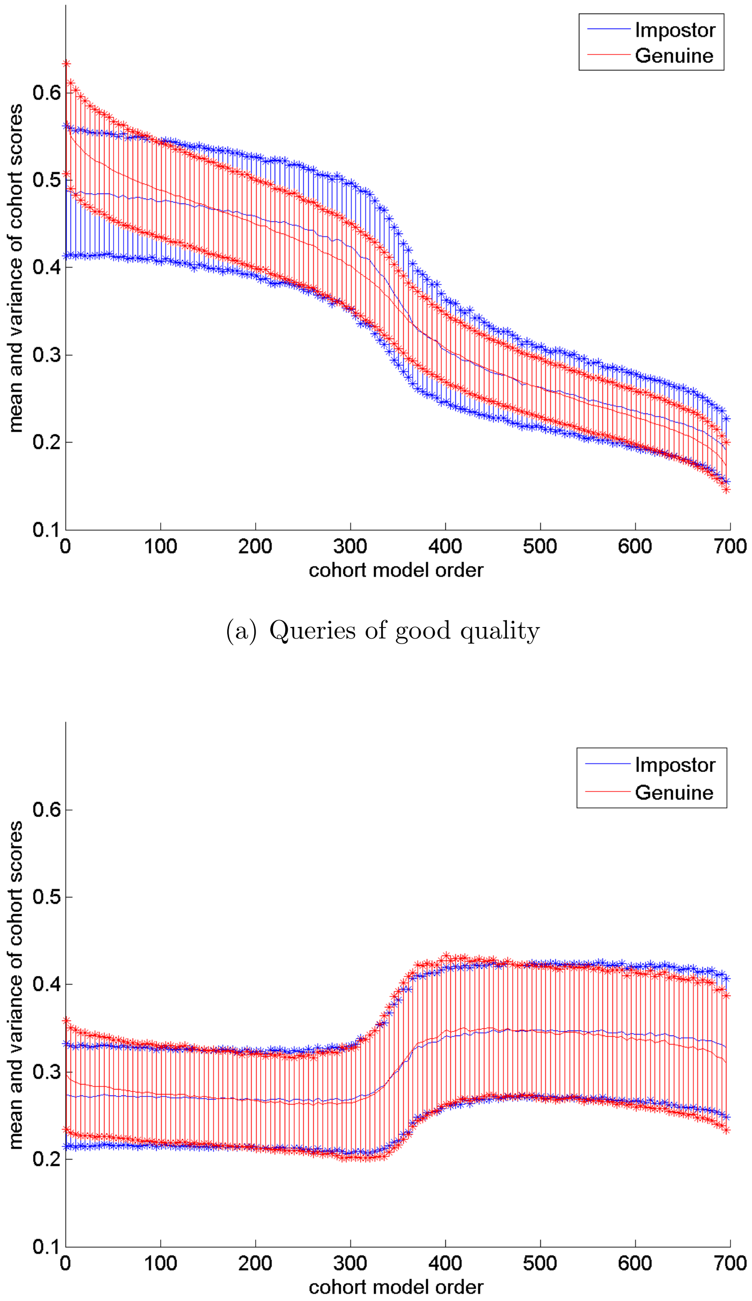
Figure 3.20: The distribution of cohort scores generated by ordered cohort samples using LBP with "Cgood". (a) Results of good queries; (b) Results of bad queries.



(a) Queries of good quality

(b) Queries of bad quality

Figure 3.21: The distribution of cohort scores generated by ordered cohort samples using LBP with "Cbad". (a) Results of good queries; (b) Results of bad queries.



(a) Queries of good quality

(b) Queries of bad quality

Figure 3.22: The distribution of cohort scores generated by ordered cohort samples using LBP with “Cboth”. (a) Results of good queries; (b) Results of bad queries.

Table 3.10: The 6 combinations of target, query and cohort samples.

Cohort condition	Good query	Bad query
Without cohort	TQgood	TQbad
Good cohort 1	TQgoodCgood1	TQbadCgood1
Good cohort 2	TQgoodCgood2	TQbadCgood2

Table 3.11: The cohort configuration of the 5 folds.

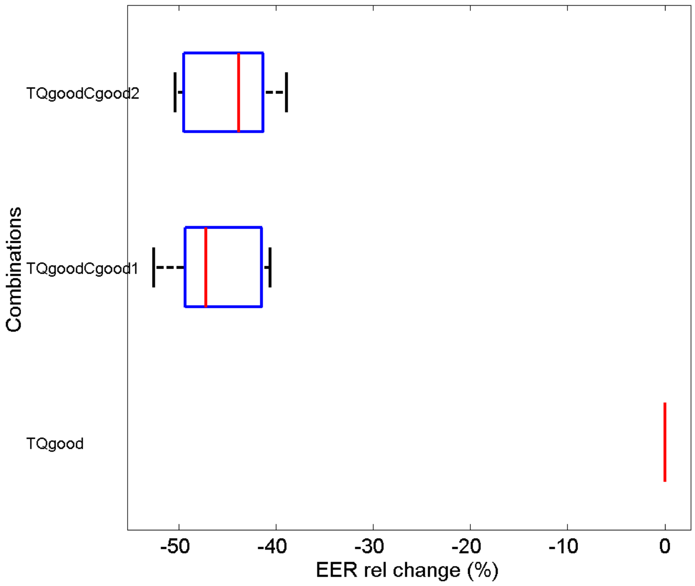
	Fold No	1	2	3	4	5
Cgood1	# cohort samples	694	716	702	718	712
	# subjects	71	63	64	71	61
	min # samples/subject	2	2	2	2	2
	max # samples/subject	20	24	24	24	24
Cgood2	# cohort samples	694	716	702	718	712
	# subjects	22	20	21	21	19
	min # samples/subject	22	32	28	26	32
	max # samples/subject	44	44	42	40	42

but also the minimum and maximum numbers of samples of each subject.

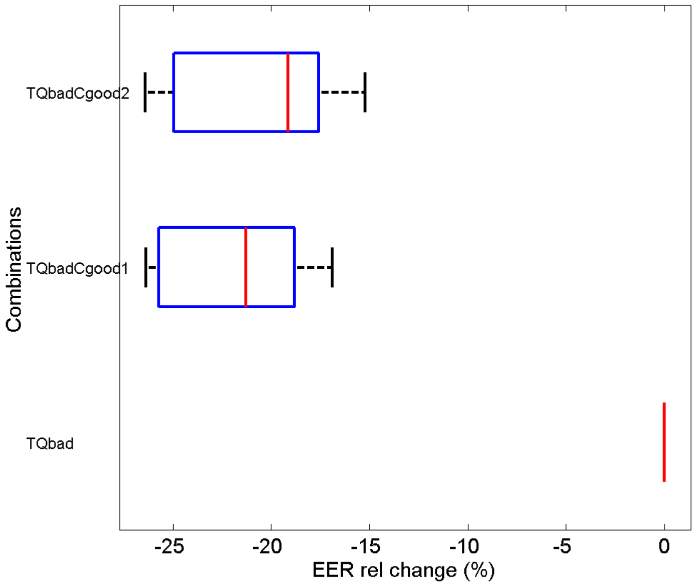
We summarize the results in Table 3.12. As observed, for matching either good or bad queries, cohort normalization with “Cgood1” and “Cgood2” leads to similar performance. The similar phenomenon is also observed in the relative change of EER as shown in Figure 3.23 and Figure 3.24.

Table 3.12: The mean EERs of the 5-fold experiment on face verification with two different cohort conditions.

Feature	Query	Czero	Cgood1	Cgood2
Gabor	Qgood	11.23%	6.07%	6.18%
	Qbad	28.67%	22.40%	22.73%
LBP	Qgood	7.46%	4.71%	4.80%
	Qbad	31.85%	23.48%	23.37%

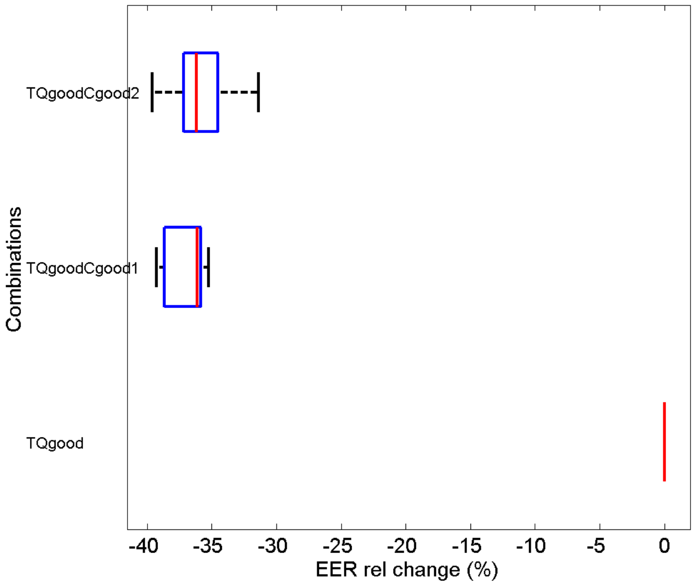


(a) Queries of good quality

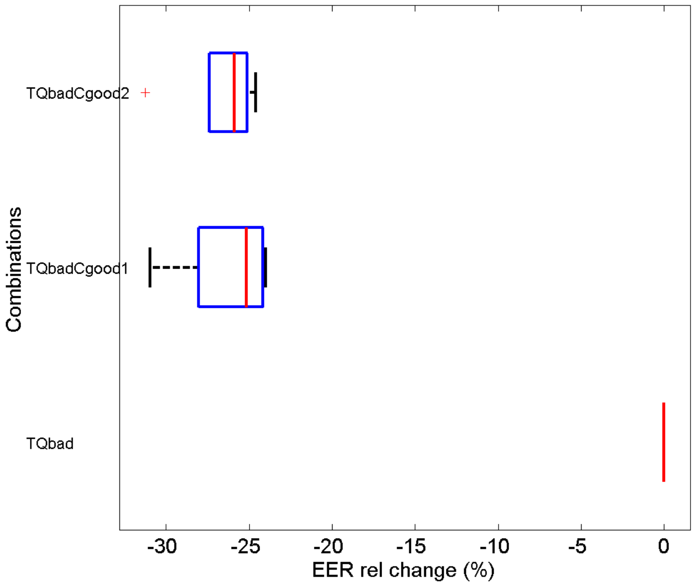


(b) Queries of bad quality

Figure 3.23: The boxplot of the relative change of EER on the 5-fold face verification experiment with {“Cgood1”, “Cgood2”} using Gabor feature. (a) Results of good queries; (b) Results of bad queries.



(a) Queries of good quality



(b) Queries of bad quality

Figure 3.24: The boxplot of the relative change of EER on the 5-fold face verification experiment with {"Cgood1", "Cgood2"} using LBP feature. (a) Results of good queries; (b) Results of bad queries.

3.4.4 Conclusion

In this work, we conduct an extensive experimental exploration on the FRGC ver2.0 database [70] to give proper answers of two interesting questions about cohort behavior on face recognition. The first one is "How does the quality of cohort samples affect the performance of a face recognition system?", while the second one is "Should we use a cohort set containing most possible subjects(each subject with fewest possible samples as possible) or a cohort set including fewest possible subjects (each subject with most possible samples)?". From the experimental results, we found that when matching a good query with a good template model, score normalization with good cohort samples can achieve much better performance than with bad cohort samples. While for the matching between a bad query and a good template model, cohort samples with "Cgood" and "Cboth" obtain similar performance. Furthermore, the results have shown that the selection of a cohort set between as many subjects as possible and as few subjects as possible has little impact on the generalization performance.

3.5 Summary

In this section, we investigated cohort score normalization for face recognition under challenging environments. First, we introduced cohort normalization to sparse presentation based undersampled face identification. Next we presented a novel and effective cohort normalization approach specifically for unconstrained face pair matching. Experiments performed on several well known face databases show the effectiveness of cohort score normalization on measuring the degradation effect under challenging environments. To better understand the cohort behavior, we investigated the impact of cohort sets of different sizes and cohort samples of different quality on the cohort normalization performance. We found that bigger cohort set size usually gives more stable and often better results to a point before the performance saturates. And cohort samples with different quality indeed produce different cohort normalization performance. Generally, using cohort samples of good quality can lead to much better performance the using bad cohort samples.

Chapter 4

Face Recognition across Alteration

Beauty and aesthetics have fascinated human beings from the very dawn of mankind, especially the attractiveness of facial appearance. A simple way to enhance the aesthetic appeal of human faces in frontal photographs is using some digital image processing tools to alter the facial appearance photos for the desired beauty. Note that these digital image alterations are instant and temporary. Nowadays, due to the advanced surgical technologies and the affordable cost, more and more people choose to take some plastic surgeries for long-lasting or even permanent facial alterations. By these medical procedures, people can either modify their appearance for aesthetic improvement or correct defects of a facial feature for functionality improvement. Both digital image alterations and plastic surgeries can be considered as intentional alterations. There are also some alterations introduced unintentionally, such as geometric alterations caused by the acquisition device or a bad printing process. As these alterations become more and more commoner, face recognition systems will be challenged to recognize individuals with alterations presented. In this section, we study the impact of these alterations on face recognition accuracy. Further, we develop some useful methods for detecting those alterations which can significantly affect the recognition performance. In the end, we present an effective algorithm for face recognition across plastic surgery.

4.1 Introduction

Face image alterations can be roughly classified into two categories: unintentional and intentional alterations. The unintentional category contains transformations that are typically introduced by either the acquisition or the printing device(e.g., lens distortion or change in the image aspect ratio). The second transformation involves those alterations which are intentionally introduced to make the facial appearance more attractive or to correct some defects, including temporal digital beautification and plastic surgery. In this work, we investigate different types of alterations which are likely to be found in practical cases.

4.1.1 Unintentional Alterations

To well illustrate each alteration, we apply the corresponding transformation to the original photo (as shown in Figure 4.1 (a)) at different levels (i.e., with different strength). We use a parameter p to describe the strength. A large value of p denotes a more significant alteration.

Barrel distortion. Barrel distortion is one of the most common types of lens distortions, which is a nonlinear and generally radial distortion and represents the typical defect that could be introduced by a low quality acquisition device. Barrel distortion is due to the fact that many wide angle lenses have higher magnification in the image center than at the periphery. This causes the image edges to shrink around the center and form a shape of a barrel as shown in Figure 4.2. In this transformation, a distortion with a strength of p is imposed on the original image while preserving the image size. The approach described in [84] has been adopted to implement this transformation. The value of p is increased in a step of 2% from 10% to 20%, i.e., $p \in \{0.10, 0.12, 0.14, 0.16, 0.18, 0.20\}$. An altered image obtained by applying the barrel distortion with $p = 0.20$ is shown in Figure 4.1 (b).

Vertical contraction. In this alteration, the original photo is vertically compressed while keeping the width fixed. In particular, the original height is reduced

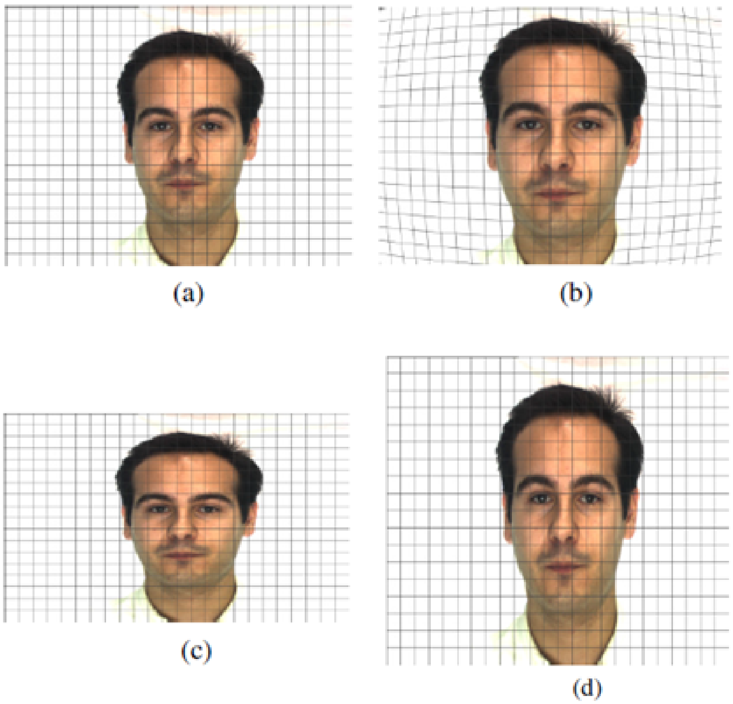


Figure 4.1: Examples of unintentional alterations. (a) Original image; (b) Altered image with barrel distortion; (c) Altered image with vertical contraction; (d) Altered image with vertical extension. A squared grid is superimposed to these original and altered images to better highlight the effect of these unintentional geometric transformations.

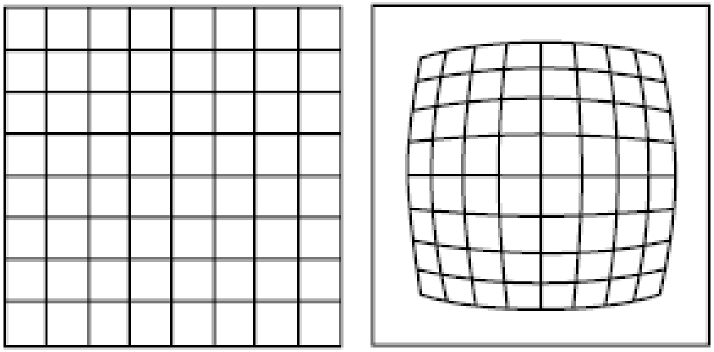


Figure 4.2: The original grid (left) and the effect of barrel distortion (right).

by a multiplying factor of $(1 - p)$. The values of p remain the same to those in the barrel distortion. Figure 4.1 (c) shows an altered image after vertical contraction with $p = 0.20$.

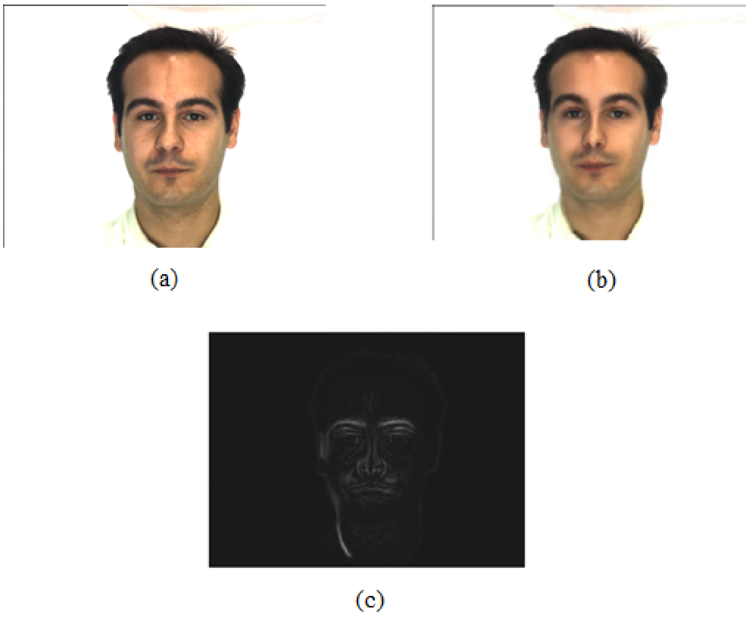
Vertical extension. On the contrary, in vertical extension, the height is increased by a multiplying factor of $(1 + p)$ while keeping the width invariable. Here too we increase the strength of extension from 10% to 20% in a step of 2%. An altered image after vertical extension with $p = 0.20$ is illustrated in Figure 4.1 (d). This alteration and the previous vertical contraction, which are essentially a modification of the face aspect ratio, could be unintentionally introduced when processing the image with a photo-editor tool or could be the result of a bad printing.

4.1.2 Intentional Alterations

For intentional alterations, we investigate the temporal digital beautification on face photos and the plastic surgery on facial appearance.

Digital beautification. This alteration usually includes those digital image processing procedures applied on face photos (e.g., using Photoshop to remove scars and to make the face skin look smoother). To obtain this alteration, we use LiftMagic [48], an instant cosmetic surgery and anti-aging makeover tool that produces realistic image beautification. The tool presents a very simple web interface that allows to load an image and to simulate different plastic surgery treatments at different levels. It makes available 17 treatments: 16 local treatments (e.g., eyelid fold enhancement) and one treatment integrating all the local ones. For each treatment, a specific selection bar is provided to personalize the strength of the modification. In this alteration, we consider only the integrated treatment and three different strengths obtained by positioning the selection bar at three approximately equidistant positions. We name the three levels "low", "medium" and "high". Figure 4.3 presents an original image and the corresponding altered image using LiftMagic with $p = high$. Here, we also illustrate the pixel difference between the original and digitally beautified images as shown in Figure 4.3 (c).

Plastic surgery. When a person undergoes plastic surgery, the facial features are reconstructive either globally or locally. Thus, plastic surgeries can be roughly



(a)

(b)

(c)

Figure 4.3: An example of digital beautification. (a) Original image; (b) Beautified image; (c) Pixel difference between the original and digitally beautified images.

grouped into two classes: local and global surgeries. Local surgeries are usually aimed at correcting defects, anomalies, or improving skin texture. These surgeries are also used for reshaping and restructuring facial features to improve the aesthetics, which can lead to varying amount of changes in the geometric distance between facial features. Differing from local plastic surgeries, global surgeries can completely change the facial appearance, skin texture or other facial geometric structures. People undergo this type of surgeries either for curing some functional damages or improving the aesthetic. Examples of local and global plastic surgeries can be found in Table 2.2. Figure 4.4 shows some common plastic surgeries. Next, we give a brief introduction about these main surgeries. In Figure 4.5, we give an example for each of these surgeries. The images are from the plastic surgery database built by Singh et al. [78]. A more detailed description of plastic surgeries can be found in [78].

1) Brow lift (forehead surgery): This surgery is usually performed to remove thick wrinkles from the forehead, thus giving people a younger look. It can also be used for old people who have vision problems caused by flagging eyebrows.

2) Blepharoplasty (eyelid surgery): It is generally recommended for patients who have vision problems due to the excessive growth of skin tissues on the eyelid. It can be also used for people with single-fold eyelids, especially those from Asian countries, to make their eyelids double-folds.

3) Otoplasty (ear surgery): This surgery is used for making the ears closer to the face and reducing the size of ears.

4) Rhinoplasty (nose surgery): With this surgery, people can straighten or narrow their nose to improve their facial appearance. In addition, for people whose nose bones are damaged due to accidents or people who have breathing problems due to the nose structure, rhinoplasty can be performed to reconstruct the nose.

5) Cheek implant: It is mostly aimed at improving the aesthetic, for example, malar augmentation, where a solid implant is fitted over the cheek bone.

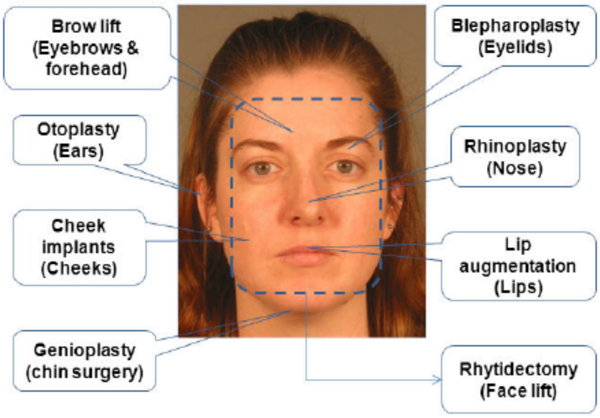


Figure 4.4: Some common plastic surgeries [35].

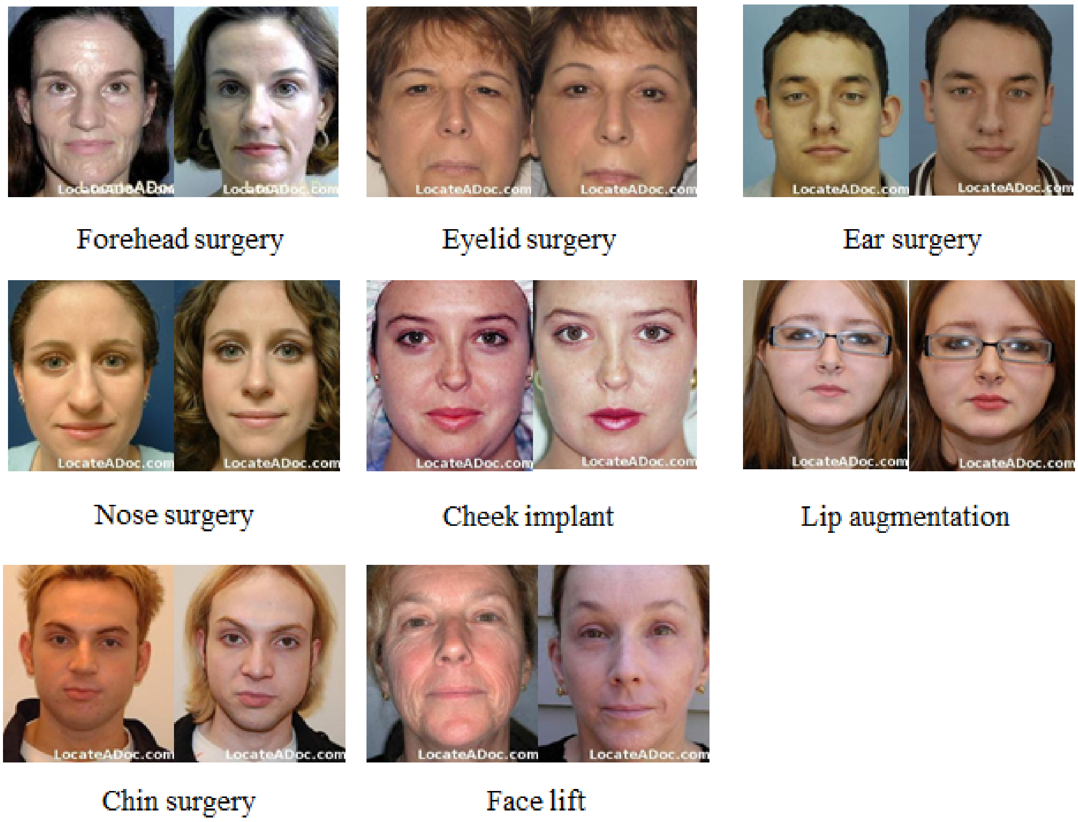
6) Lip augmentation: As is known, lips play an important role in an individual's beauty. Hence, people perform this surgery generally for shaping or enhancing their lips with injectable filler substances to get desired beauty.

7) Genioplasty (chin surgery): It is generally employed for reshaping the chin such as correcting bone damages, smooth rounding of the chin and reducing/augmenting chin bones.

8) Rhytidectomy (face lift): Differing from the above local surgeries, face lift is purely global surgery. It can be used to fight aging and make people look younger through tightening the face skin. Furthermore, face lift can be employed to treat patients with severe burns on face and neck.

4.2 Detection of Alteration in Face Photo

This work is framed into the context of automatic face recognition in electronic identity documents. We study the impact of alterations of the face images used for enrollment on the recognition accuracy. For the alterations, we consider both unintentional and digital intentional alterations. Our results show that state-of-the-art algorithms are sufficiently robust to deal with some alterations whereas



Forehead surgery

Eyelid surgery

Ear surgery

Nose surgery

Cheek implant

Lip augmentation

Chin surgery

Face lift

Figure 4.5: Examples of the main surgeries from the plastic surgery database [78].

other kinds of degradation can significantly affect the accuracy. Furthermore, we develop some useful methods to detect those alterations which can significantly affect the recognition performance.

4.2.1 Overview

In recent years, traditional identity documents have been replaced by electronic documents able to store biometric features to be used for machine-assisted identity verification [16, 17]. With the Berlin resolution (2002), the International Civil Aviation Organization (ICAO) selected the face as the primary globally interoperable biometric characteristic for machine-assisted identity confirmation in electronic Machine Readable Travel Documents (eMRTD) [1].

In order to facilitate the automatic identity verification process, the images stored in an ICAO compliant electronic document have to fulfill very restrictive quality standards, i.e., no elements that could compromise the recognition accuracy should be present. A number of indications about the geometric and photometric properties of the face images to be used in e-documents are given in the ISO/IEC 19794-5 standard following the guidelines initially proposed by ICAO. For instance the subject should have a well-controlled pose, a proper lighting, a natural expression, no accessories that could partially occlude some important facial characteristics.

Some of the countries issuing e-documents, acquire the face images of the subject at the enrollment station with a digital camera. Other countries (e.g., Italy) require the user to provide a printed face photo (ID format) to the issuing authority. While in the first case it is enough to verify (manually or in machine-assisted way) ISO/IEC19794-5 compliance [2], in the second case a further validation should be done to ensure that the face in the printed photo has not (intentionally or unintentionally) been altered.

In fact, several problems may raise in different scenarios if an altered face image are included in the document:

1) in a verification scenario, for instance in an automatic gate in an airport, the alterations may determine a high rate of false rejections, i.e., the system

would not recognize the person thus making necessary the human intervention.

2) in a watch-list scenario, where a list of subjects wanted by the police has to be checked in order to raise proper alarms, the presence of altered face images in the documents could imply missing the suspect that, in this case, could intentionally alter the face image to reduce the probability of being identified.

In this section, we present a study of the effects of image alterations on face recognition performance. More in detail, we include intentional digital alterations and unintentional geometric alterations. For the experiments, we use three reference recognition approaches: two commercial systems and one algorithm at the state-of-the-art are used. We further develop two facial descriptors (y-distance and triangle descriptors) specially used for detecting those alterations which can significantly degrade the recognition performance.

4.2.2 Impact of Alterations on Face Recognition Accuracy

Experimental settings. We evaluate the effects of some alterations described in Section 4.1, including unintentional alterations and digital beautification, on face recognition accuracy. Here, we focus on face verification. And our verification framework is shown in Figure 4.6. During the enrollment stage, the user needs to provide a printed face photo or a face photo captured by a digital camera to the authorities issuing electronic documents. The provided photo is then stored in the database as the e-document representing the corresponding user. Note that if the provided photo is a printed one, the authorities need to first scan it and then store it in the digital format. During the verification stage (e.g., in an airport's customs), a face image is first acquired with a digital camera. By matching the captured new face photo with the one stored in the user's e-document, the goal of face verification is to verify whether the two photos represent the same person. In our work, image alterations are performed on the provided face photo in the enrollment stage. For each of the three unintentional alterations (barrel distortion, vertical contraction and vertical extension), we consider six different strengths,

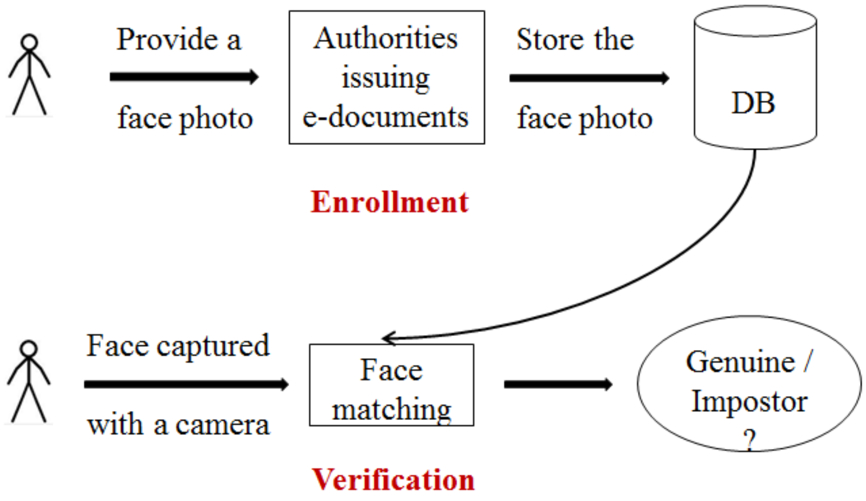


Figure 4.6: Our face verification framework for studying the impact of alterations on face recognition accuracy.

i.e., $p \in \{0.10, 0.12, 0.14, 0.16, 0.18, 0.20\}$, as described in Section 4.1. For digital beautification, the strength is three-level, i.e., $p \in \{low, media, high\}$. The experiments are conducted with three different state-of-the-art face recognition approaches: two commercial softwares (Neurotechnology VeriLookSDK 2.1 [64] (VL) and Luxand FaceSDK 4.0 [52] (LU)) and a SIFT-based matching algorithm [51] (SI).

The choice of a proper face database is here an important issue. In fact, in the context of electronic documents, face images are expected to be high quality; hence, variations caused by illumination, expressions, poses, etc. should not be presented in the selected database. The selected database is the AR face database [57]. This database consists of 4,000 frontal images taken under different conditions in two sessions, separated by two weeks. The images relevant to our study are well controlled and high quality images (with neutral expressions and good illumination), hence the poses 1 and 14 are selected for the tests. We denote them as $No1$ and $No14$ respectively (see Figure 4.7 for an example). In our test we assume that the images $No14$ are used for the enrollment (i.e., are stored in the e-documents), while the images $No1$ are used as probe (i.e., at the point of verification). The alterations are thus applied to images $No14$ to simulate the

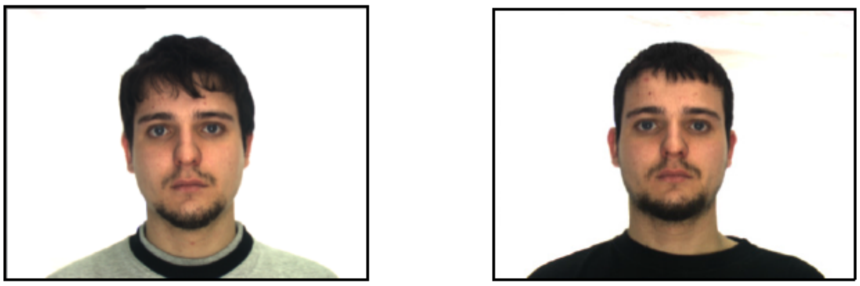


Figure 4.7: Two unaltered images of the same subject on the AR database (pose 1 on the left, pose 14 on the right).

inclusion in the document of an altered image. For the two commercial systems VL and LU, we directly use the original images. For the SIFT-based matching algorithm, before extracting the SIFT features, we first geometrically normalize the original image to 240 × 320 (240 is the width while 320 is the height) with the distance between the two eye centers $eyedist = 60$. For the SIFT descriptor, we divide the image into non-overlapping blocks of 16 × 16 size, and compute a 128D SIFT descriptor for each block. Finally, the cosine similarity is calculated as the matching score of two SIFT descriptors.

To evaluate the effects of the various alterations on face recognition accuracy, a systematic experimentation has been carried out. Starting from the original database, for each alteration, face images with different alteration strength have been generated by modifying the original images with different transformations (see Section 4.1). The performance evaluation of face recognition algorithms is based on a set of genuine and impostor recognition attempts. In a genuine recognition attempt, two face images of the same individual are compared, while in an impostor attempt, two images from different persons are compared. In each genuine/impostor attempt, the first image is supposed to have been acquired during an "enrollment" stage (and included into the document) and the second one during a "verification" stage. The performance indicators are FMR100 FMR1000 [53].

In the following definitions, each database DB consists of two sets of face images: DB_e (acquired during enrollment) and DB_v (acquired during verification).

The original database (without alterations) is denotes as $DB^o = \{DB_e^o, DB_v^o\}$. DB_e^o is made of all the original $No14$s of 120 subjects, while DB_v^o is composed of all the original $No1$s of 134 subjects. For genuine attempts, each $No14$ is compared against the $No1$ of the same subject; since only 118 subjects have both pose 1 and 14, the number of genuine attempts is 118. For impostor attempts, the $No14$ of one subject and all the $No1$ of the other subjects are compared. Hence, the total number of impostor attempts is $118 \times 118 - 118 = 13,806$.

As to the altered databases, for a given alteration a let $DB_a^p = \{(DB_e)_a^p, DB_v^o\}$ be a database that simulates enrollment face images reporting alteration a with a strength of p. For genuine attempts, the original $No1$ and the altered $No14$ from the same subject are compared. Impostor attempts are the same as in the original database DB^o.

Experimental results. The first experiment is performed on the unaltered database. The performance measured for the three systems (VL, LU and SI) is good as shown in Figure 4.8, hence they constitute a good test bed to evaluate the effects of alterations: in particular the measured EER is 0.003% (VL), 1.693% (LU) and 2.217% (SI).

The results of the barrel distortion are reported in Figure 4.9. It can be observed that both FMR100 and FMR1000 change slightly and irregularly as the degree of barrel distortion increases for LU and SI, while there is no significant performance change for VL. Overall this alteration has no noticeable effects on the recognition accuracy.

Figure 4.10 and Figure 4.11 illustrate the results of the vertical contraction and extension respectively. For both FMR100 and FMR1000, as the strength of the alterations increases, the accuracy of LU significantly decreases. SI shows a less noticeable performance drop than LU, while there is no significant performance change for VL.

Finally the results of the digital beautification are reported in Figure 4.12. For both FMR100 and FMR1000, this alteration produces a performance drop for all the three system (even if LU shows a less noticeable reduction of the recognition accuracy).

Overall the experimental results show that the barrel alteration does not affect significantly the recognition accuracy. This is probably due to the fact that in

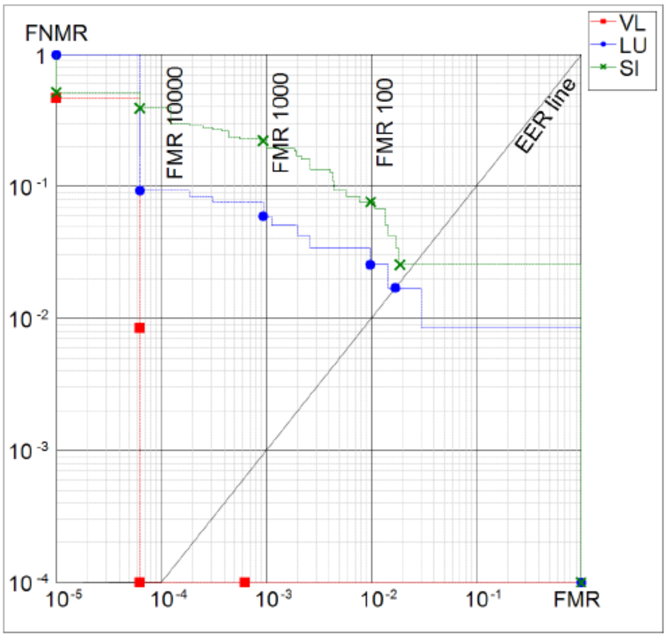


Figure 4.8: The DET curves of the three reference systems on the unaltered database.

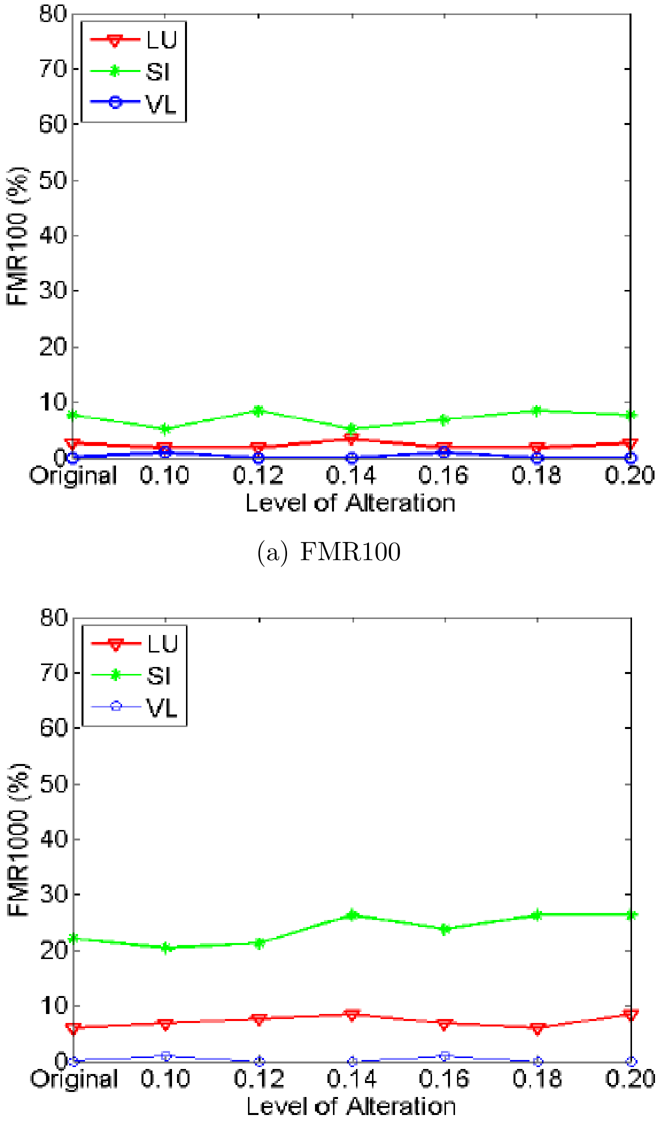
the central part of the image containing the face, the barrel distortion produces simply a sort of scaling effect, which is well handled by the algorithms analyzed.

Aspect ratio alteration is critical for some approaches (for instance the vertical contraction at the maximum strength causes a performance drop of FMR1000 of about 11 times for LU) while it is just slightly disturbing other systems. In particular, we believe that face recognition based on local features only is quite insensitive to global geometric changes.

Finally, alteration such as digital beautification, when applied with high strength, produces marked performance drop to all the system tested.

4.2.3 Detection of Alteration in Face Photo

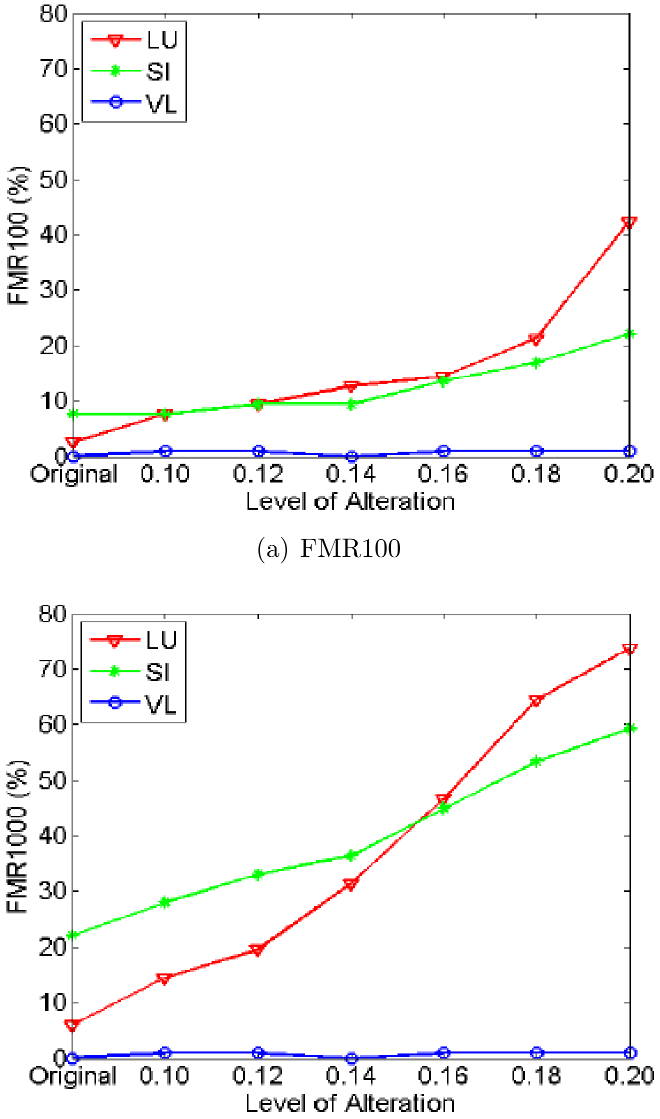
Methodology. As described above, state-of-the-art algorithms are able to overcome limited digital alterations but are sensitive to more relevant modifications. Thus, proper countermeasures are suggested to be taken to avoid storing digitally altered photo in eMRTD. In particular, the authorities issuing e-documents,



(a) FMR100

(b) FMR1000

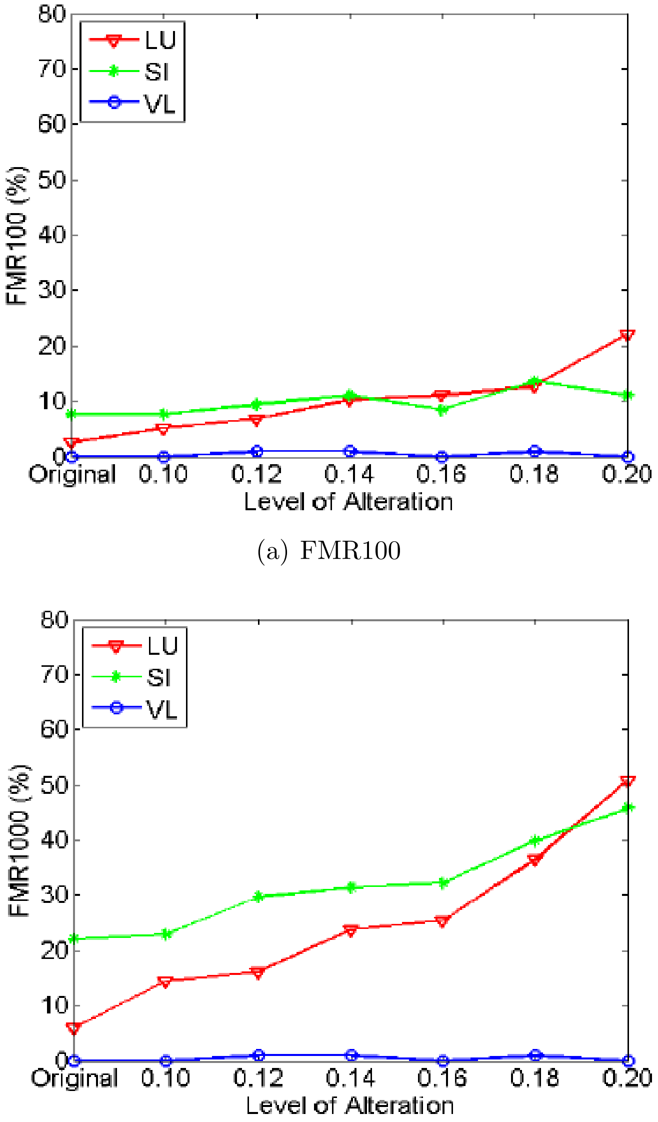
Figure 4.9: The performance comparison before and after barrel distortion. (a) Results of FMR100; (b) Results of FMR1000.



(a) FMR100

(b) FMR1000

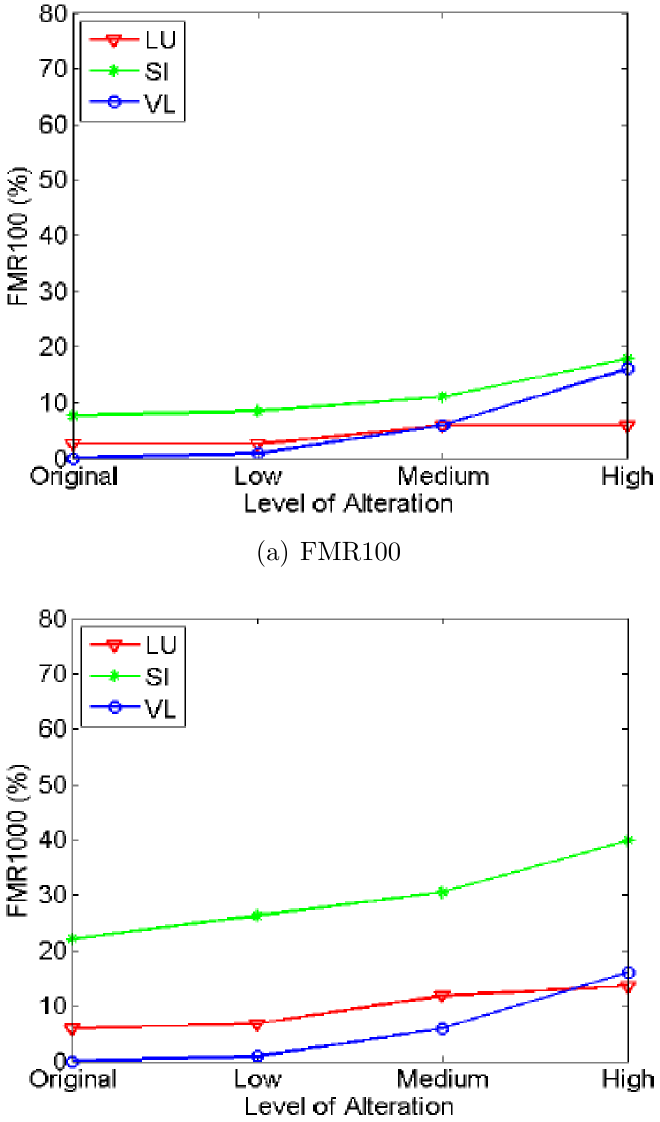
Figure 4.10: The performance comparison before and after vertical contraction. (a) Results of FMR100; (b) Results of FMR1000.



(a) FMR100

(b) FMR1000

Figure 4.11: The performance comparison before and after vertical extension. (a) Results of FMR100; (b) Results of FMR1000.



(a) FMR100

(b) FMR1000

Figure 4.12: The performance comparison before and after digital beautification. (a) Results of FMR100; (b) Results of FMR1000.

based on printed ID photos provided by citizens, should carefully check these photos to detect intentional and unintentional alterations. In this section, we focus on how to effectively detect those alterations which can significantly degrade the face recognition performance. As shown in Section 4.2.2, aspect ratio alteration is a critical problem for current face recognition systems. Hence, our work in this section is placed on developing effective techniques for the detection of vertical contraction and vertical extension.

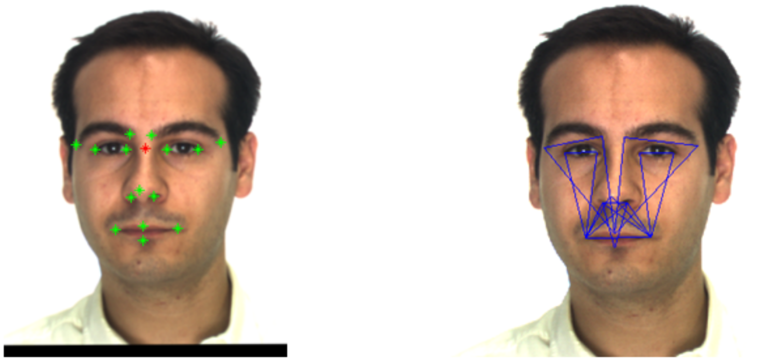
Our first investigation is to employ the current state-of-the-art local feature based algorithms for this task. Here, we use three well studied and widely used facial features: Gabor [49], LBP [6] and SIFT [15]. For the matching score, we compute the cosine similarity between two feature vectors. As is known, in both vertical contraction and vertical extension, the aspect ratio between the width and the height is changed. One intuitive method is to detect alterations through measuring this aspect ratio. Based on this idea, we develop two methods for this task: y-distance descriptor and triangle descriptor. Next, we give the details of both techniques.

The y-distance descriptor is calculated on the geometrically normalized face image, i.e., the image of size 240×320. Note that the coordinate of the midpoint of the two eye centers is fixed, i.e., $pm = (pm_x, pm_y) = (120, 144)$. We automatically locate 15 landmarks on the face region using a publicly available tool known as STASM [20, 62]. The 15 landmarks are shown by green stars in Figure 4.13 (a), while the red landmark is the fixed eye midpoint pm. Denote by $pi = (pi_x, pi_y), i = 1, ..., 15$, the 15 landmarks together with their coordinates, the y-distance descriptor is a vector $ydist = \{yd1, ..., ydi, ..., yd15\}$ with ydi computed as follows:

$$ydi = pi_y - pm_y \tag{4.1}$$

The similarity score between two y-distance descriptors $aydist = \{ayd1, ..., aydi, ..., ayd15\}$ and $bydist = \{byd1, ..., bydi, ..., byd15\}$ is calculated by

$$sc = \frac{1}{15}\sum_{i=1}^{15} sc(i) \tag{4.2}$$



(a) 15 points

(b) 8 triangles

Figure 4.13: The 15 landmarks and 8 triangles for the y-distance and triangle descriptors. (a) 15 landmarks; (b) 8 triangles.

where

$$sc(i) = \begin{cases} 0 & \frac{\|aydi - bydi\|}{eyedist} > 1 \\ 1 - \frac{\|aydi - bydi\|}{eyedist} & otherwise \end{cases} \tag{4.3}$$

Recall that $eyedist = 60$, through the division by the distance between the two eye centers, we can change the computed score into a normalized value. Here, we compute only the y-distance (i.e., the distance along the vertical axis), not the Euclidean distance (i.e., the distance along both the vertical and the horizontal axes). The reason is for both vertical contraction and vertical extension, the width keeps fixed, only the height gets changed. In other words, the x-distance descriptor keeps unchanged before and after these two aspect ratio alterations.

The triangle descriptor is based on the idea of congruent triangles, thus is calculated directly on the original image due to the scale-invariant property of congruent triangles. Similarly, we locate the same 15 landmarks to those used for computing the y-distance descriptor. With these 15 landmarks, we construct 8 triangles as shown in Figure 4.13 (b). For each triangle, we compute the radian values of the three angles. Denote by $trides = (tri1, ..., trii, ..., tri8)$ the triangle descriptor, where $trii$ represents the 3-angle vector of the i^{th} triangle and $trii = [trii_1; trii_2; trii_3]$. Hence the dimension of the triangle descriptor is 3×8. The similarity score between two triangle descriptors $atrides = (atri1, ..., atrii, ..., atri8)$ and $btrides = (btri1, ..., btrii, ..., btri8)$ is calculated as follows:

1) Calculate the discrepancy matrix $tridist = atrides - btrides$.

2) Compute the sum of the $tridist$ over each column leading to a 8D vector $trisumd = \{sumd1, ..., sumdi, ..., sumd8\}$.

3) The similarity score $sc = \frac{1}{8}\sum_{i=1}^{8}\left(1 - \frac{sumdi}{3\pi}\right)$.

The division by 3π is for the score normalization.

Experimental settings. We use the $No1$s and $No14$s of the 118 subjects for this experiment. The two aspect ratio alterations are performed on the $No14$s with six different strengths $p \in \{0.10, 0.12, 0.14, 0.16, 0.18, 0.20\}$. Our task is a binary classification problem. In particular, given two face images from the same individual, our goal is to determine whether there are some alterations presented on the second image in comparison with the first image. The performance evaluation of the conventional local feature based methods and our proposed methods is based on a set of no-alteration and with-alteration attempts. In a no-alteration attempt, the $No1$ and the original $No14$ of the same individual are compared, while in a with-alteration attempt, the $No1$ and the altered $No14$ of the same individual are compared. Thus, the number of both no-alteration and with-alteration attempts is 118. A good algorithm should classify the two types of attempts as correctly as possible. Here, we use EER as the performance indicator.

To compute the Gabor feature, we adopt five scales and eight orientations of the Gabor filters. The final Gabor feature vector is obtained by concatenating the responses at different pixels selected uniformly with a 10×10 down-sampling rate. For extracting LBP features, we divide each image into non-overlapping blocks of 8×8 size and extract 59-bin uniform LBP histograms for each block. For the SIFT descriptor, we also divide the image into non-overlapping blocks of 16×16 size, and compute a 128D SIFT descriptor for each block.

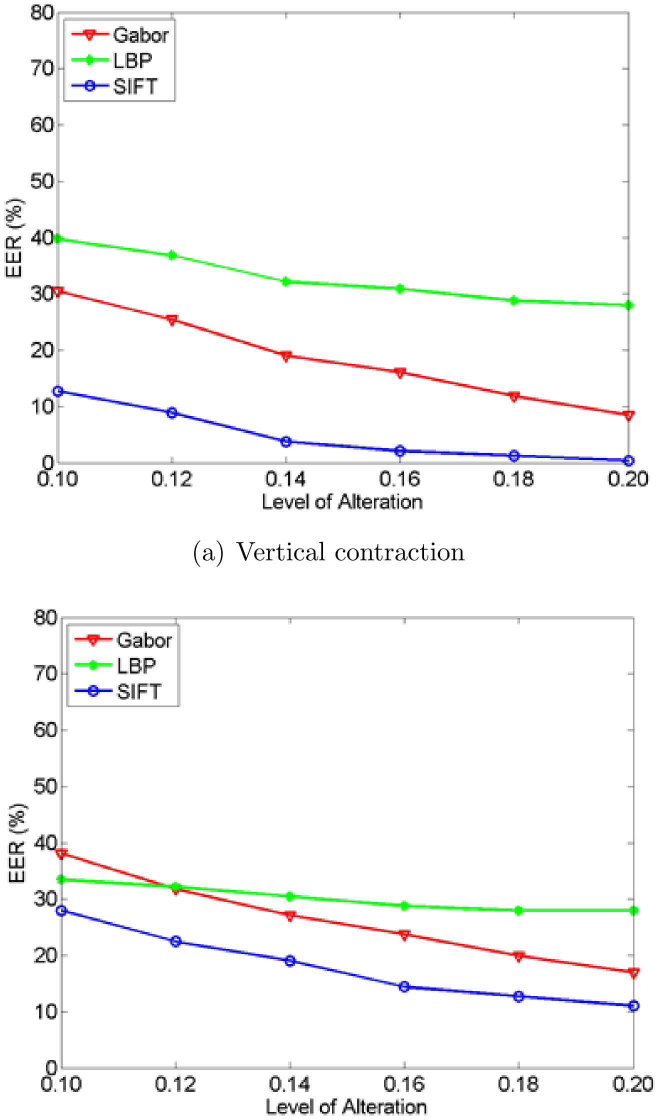
Experimental results. The results of using the conventional local feature based methods are illustrated in Figure 4.14. For both vertical contraction and extension, the SIFT feature achieves much better performance than Gabor and LBP. In particular, large strength vertical contraction can be detected effectively by SIFT, for example, with a strength $p \in \{0.14, 0.16, 0.18, 0.20\}$, the EERs are 3.814%, 2.119%, 1.271% and 0.424%, respectively. In comparison, SIFT leads to worse performance when detecting vertical extension than handling vertical contraction, for example, the EERs corresponding to strength $p = \{0.14, 0.16, 0.18, 0.20\}$ are 19.068%, 14.407%, 12.712% and 11.017%. As shown in Figure 4.14, when an alteration with a very low strength (e.g., vertical extension with $p = 0.10$) is presented on the face image, all the three methods fail to get a satisfied performance. For example, with a vertical extension of strength 0.10, the EERs of Gabor, LBP and SIFT are 38.136%, 33.475% and 27.966%, respectively. In addition, Gabor and LBP can not even well detect alterations with large strengths.

Figure 4.15 illustrates the results of using the proposed y-distance and triangle descriptors. For vertical contraction, y-distance achieves similar results to SIFT. With a lower strength $p = \{0.10, 0.12\}$, y-distance even exceeds SIFT, i.e., 7.627% and 4.661% of y-distance in comparison with 12.712% and 8.898% of SIFT. Furthermore, the y-distance descriptor can also produce good performance for detecting vertical extension, the EERs corresponding to strength $p = 0.14, 0.16, 0.18, 0.20$ are 4.661%, 4.237%, 2.966% and 0.847%. In comparison, the triangle descriptor leads to much worse performance than y-distance, especially when dealing with alterations of high strength.

4.2.4 Conclusion

In this section, we introduce two categories of image alterations: unintentional and intentional alterations. Further, we present a study of the effects of unintentional alterations and digital beautification on face recognition accuracy. The outcome of this study is that state-of-the-art algorithms are able to overcome limited digital alterations but are sensitive to more relevant modifications. Hence, proper countermeasures are suggested to be taken to avoid storing digitally altered photo in e-documents. Finally, we develop two facial descriptors (y-distance and triangle descriptors) specially used for detecting aspect ratio alterations. The comparison with three current state-of-the-art facial descriptors shows that y-distance can effectively detect both vertical contraction and extension.

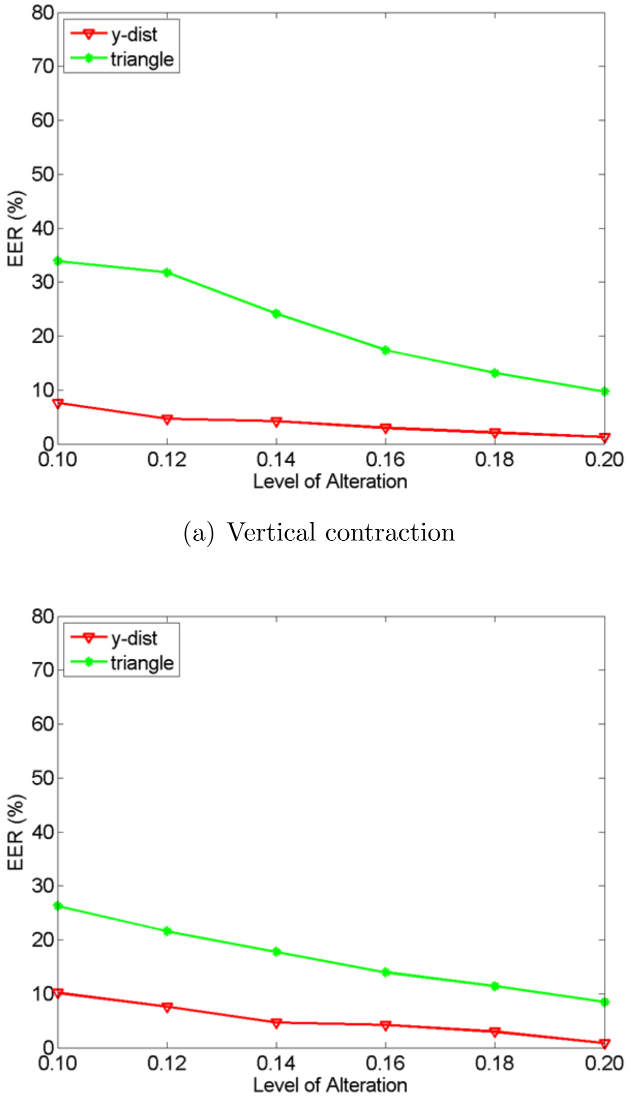
In the end, we suggest that authorities issuing e-documents, based on printed



(a) Vertical contraction

(b) Vertical extension

Figure 4.14: The EERs of classification between no-alteration and with-alteration attempts using conventional local feature based methods. (a) Vertical contraction; (b) Vertical extension.



(a) Vertical contraction

(b) Vertical extension

Figure 4.15: The EERs of classification between no-alteration and with-alteration attempts using our proposed y-distance and triangle descriptors. (a) Vertical contraction; (b) Vertical extension.

ID photos provided by citizens, carefully check these photos to detect intentional or unintentional alterations. To this purpose the officer workstation could be equipped with a software supporting the operator in comparing the scanned version of the ID photo with the live acquired face image. Automatic face recognition system (possibly based on both global and local features) could be used to issue warning in case of low matching scores and a graphical superimposition of the two face images could easily reveal to the officer the presence of alteration.

4.3 Structural Similarity based Image Quality Map for Face Recognition across Plastic Surgery

Variations in the face appearance caused by plastic surgery on skin texture and geometric structure, can impair the performance of most current face recognition systems. In this work, we proposed to use the Structural Similarity (SSIM) quality map to detect and model variations due to plastic surgeries. In the proposed framework, a SSIM index weighted multi-patch fusion scheme is developed, where different weights are provided to different patches in accordance with the degree to which each patch may be altered by surgeries. An important feature of the proposed approach, also achieving performance comparable with the current state-of-the-art, is that neither training process is needed nor any background information from other datasets is required. Extensive experiments conducted on a plastic surgery face database demonstrate the potential of SSIM map for matching face images after surgeries.

4.3.1 Overview

Plastic surgery becomes worldwide nowadays due to the advanced surgical technologies and the affordable cost. By these medical procedures, people can correct defects of a facial feature for functionality improvement or modify the appearance for aesthetic improvement. Face recognition across plastic surgery was first introduced to the biometric community by Singh et al. [78]. In the presented work, a database of face images related to various types of plastic surgeries was publicly released. Moreover, various existing algorithms were tested on this database. The significant performance degradation concluded that the current state-of-the-art face recognition algorithms cannot provide good performance for matching faces across plastic surgery.

To handle the challenges of face recognition after plastic surgery, Bhatt et al. proposed an evolutionary granular approach to extract discriminative information from non-disjoint face granules [14]. In [55], a recognition approach which integrates information derived from local region analysis was proposed to address this problem. Aggarwal et al. developed a sparse representation based local facial characteristic matching approach [5]. In the related work, a sequestered face image set was used to fulfill the multiple image requirement of sparse representation approach. In [34], a fusion approach by combining information from both the whole face and the ocular regions [68] was proposed to deal with the challenges of matching faces across variations caused by plastic surgeries. Very recently, Liu et al. employed an ensemble of Gabor Patch classifiers via Rank-Order list fusion algorithm and achieved very promising results [50].

As presented in [78], both local and global surgeries may result in varying amount of change in relative positioning of facial features and texture. Generally, the positions of these changed features and texture are unknown to a face recognition algorithm. However, it would be of great use if the positions could be extracted automatically. In such cases, we can less consider or ignore these changed features and texture. In this work, we consider variations caused by surgeries as a variety of distortions on the pre-surgery facial images. Further, we shall attempt to exploit the quality information implicated in the pre- and post-surgery images to detect and capture these variations.

An effective image quality tool to well model variations caused by surgeries should interpret well the degradation of both texture and structural information. In [89], an objective image quality metric based on the Structural Similarity (SSIM) index was developed for localized quality measurement. Given a reference image and its distorted version, through locally computing the SSIM index, this technique can provide a spatially varing quality map of the distorted image, which delivers much information about the quality degradation. In our work, we consider the pre-surgery image as a reference image and the post-surgery image as a distorted one. Then the SSIM quality map between the two images can be computed. Further, we employ this quality map in a patch level to control the contribution of each patch to the final matching score.

In face identification on the plastic surgery database, given a probe image y (post-surgery) and N gallery individuals (each with a pre-surgery image), the SSIM map between y and each gallery image is first calculated. Next, we perform the matching between y and each gallery image using our proposed SSIM index weighted multi-patch fusion scheme. In this scheme, the two images being compared together with their SSIM map are first divided into the same number of patches. When matching two corresponding patches, we calculate the mean SSIM index of the corresponding SSIM map patch as the weight for controlling the contribution of the patch to the final matching score. An intuitive motivation for employing the mean SSIM index of each patch as the patch weight is that, in the SSIM map of two images, the value of each pixel (SSIM index) represents the structural similarity of the two image regions within a local window of the corresponding pixel. By directly employing the SSIM index as the weight, we can give less weights (smaller SSIM index values) to those changed features and texture, and vice versa. After matching between y and the N gallery images, we can get a total of N SSIM index weighted matching scores. The final class label is the identity owing the highest matching score (similarity score).

The proposed approach employs the SSIM map to compute different weights for different image patches in accordance with the degree to which each patch is altered by surgeries. A significant advantage of the proposed approach is that neither training process, nor any background information from other databases is required. However, when matching faces of different individuals, for regions where the two faces differ most, lower weights are also assigned. To study the effect of our approach on genuine matching (comparing faces of the same person) and impostor matching (comparing faces of different individuals), we shall illustrate the score distributions. As it can be noticed from Figure 4.19, the overlap region between the genuine and impostor distributions is reduced by the SSIM weighting.

To thoroughly evaluate our proposed framework, we perform face matching in both a holistic manner and a component-wise manner. The proposed approach is evaluated on the plastic surgery database introduced in [78]. With our proposed approach, a significant improvement in recognition performance is observed.

4.3.2 Related Work

To quantitatively measure the image quality, several metrics have been proposed. The Structural Similarity (SSIM), proposed by Wang et al. [89], provides a spatially varying quality map of the two images being compared. The resulting measure allows to determine the location and degree of variations of the distorted image.

SSIM provides meaningful comparisons across different types of image distortions by separating the task of similarity measurement into three comparisons: luminance, contrast and structure. Suppose x and y are two image signals, which have been aligned with each other. The SSIM index between them is computed as follows:

$$SSIM\left(x,y\right)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)} \tag{4.4}$$

where μ_x and μ_y are the mean intensity values of signal x and y, while σ_x and σ_y are their corresponding standard deviations. σ_{xy} is the correlation coefficient between x and y. C_1 and C_2 are small positive constants used to avoid instability when the denominator is very close to zero.

For image quality assessment, it is useful to apply the SSIM index locally rather than globally. In [89], the local statistics μ_x, σ_x and σ_{xy} are calculated within a 11 × 11 circular-symmetric Gaussian window, which moves pixel-by-pixel over the whole image. At each step, the local statistics and SSIM index are computed within the local window. If one of the image signals being compared is regarded as perfect quality, then with such a windowing approach, a SSIM index map of the other image can be obtained. In the SSIM map, the value of each pixel is the SSIM index computed within the local window of the corresponding pixel. The SSIM index map provides a measurement of local image quality over space, in which the dynamic range of each pixel value is $[-1, 1]$. A pixel with value close to 1 means less distortions in the neighborhood of the pixel, while a pixel with a lower SSIM index value implies some variations or quality degradation within the local region of the pixel are detected in the distorted image. A Matlab implementation of the SSIM index algorithm is available online [60].

4.3.3 SSIM for Face Recognition across Plastic Surgery

In the proposed framework, the pre-surgery image is regarded as perfect quality and used for the gallery image in face identification, whereas the post-surgery image is viewed as a distorted image and used as a probe image. Denote the number of gallery individuals as N, that is we have a total of N gallery images. Given a probe image with some variations caused by surgeries, our face identification task aims to determine which gallery individual the probe image comes from. To get the final classification, we need to first calculate the similarity score (matching score) between the probe and each gallery image, and the identity owing the highest similarity score is the correct class label. Figure 4.16 illustrates our approach of matching the probe with one gallery individual, which roughly consists of the following two main steps: image alignment and SSIM index weighted multi-patch fusion. After alignment of the probe and the gallery images according to the eye centers, the SSIM map between the two images is computed. Next we divide the two images as well as their SSIM map into the same number of patches of the same size. The general idea of SSIM index weighted image matching is to employ this quality map in a patch level to control the contribution of each patch to the final matching score. The outline of using our approach for face identification across plastic surgery is shown in Figure 4.17. Next we present details of the main steps.

Image alignment. Images from the plastic surgery database are used in this work. For each subject, there is one pre-surgery image and one post-surgery image. All the images are of the same size 273×236 pixels. Like any other common face recognition algorithms, we first need to locate some fiducial landmarks and then normalize all the images according to the positions of these landmarks. Here, we choose to use eye centers for face alignment. Considering that a number of images in the plastic surgery database present some variations in the eye region due to expressions and poses, we choose to use the four eye corners to determine the eye centers. That is the mean position of the two left (right) eye corners is computed as the position of the left (right) eye center. To locate eye corners

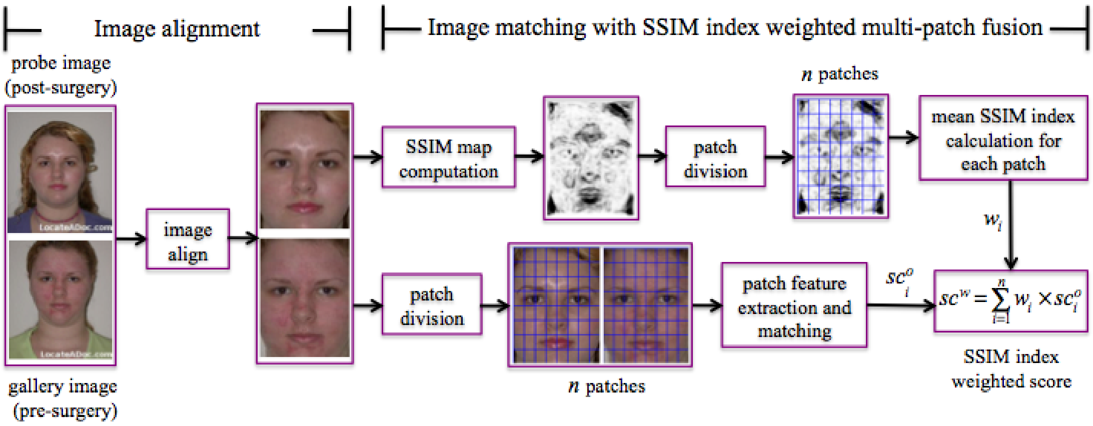


Image alignment
Image matching with SSIM index weighted multi-patch fusion
probe image (post-surgery)
gallery image (pre-surgery)
image align
SSIM map computation
patch division
n patches
mean SSIM index calculation for each patch
w_i
patch division
n patches
patch feature extraction and matching
sc_i^o
$sc^w = \sum_{i=1}^{n} w_i \times sc_i^o$
SSIM index weighted score

Figure 4.16: Outline of the process to compute the reliability weight, for face matching, from SSIM maps.

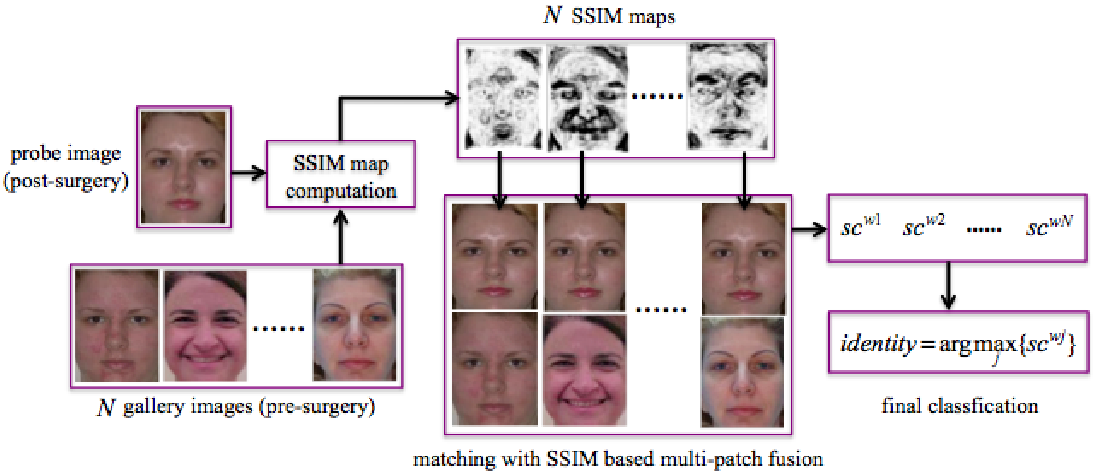

N SSIM maps
probe image (post-surgery)
SSIM map computation
N gallery images (pre-surgery)
matching with SSIM based multi-patch fusion
sc^{w1} sc^{w2} ...... sc^{wN}
$identity = \arg\max_j \{sc^{wj}\}$
final classfication

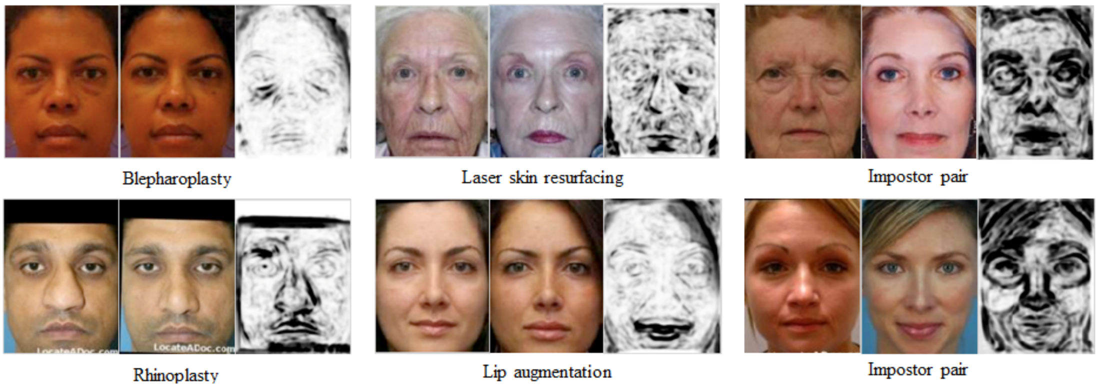
Figure 4.17: Outline of the process for face identification using SSIM-based image matching.

automatically, a publicly available tool known as STASM [20, 62] is employed. After the location of eye centers, the distance between the eye centers is set as 60 pixels. Finally, all the images are normalized to the size of 160 × 120.

SSIM index weighted multi-patch fusion. Given a pre-surgery image and a post-surgery image, we regard the pre-surgery image as a reference image, whereas the post-surgery image is viewed as a distorted image. Next the SSIM map of the post-surgery image is computed using the approach outlined in Section 4.3.2. For the parameters C_1 and C_2, we use the same values to those used in [89]. And the standard deviation of the 11×11 circular-symmetric Gaussian weighting function is set to 1.5. More details about how to calculate the SSIM map can be found in [89].

Figure 4.18 shows some SSIM index maps corresponding to some pre- and post-surgery image pairs. In the SSIM map, dark regions represent smaller SSIM index and larger distortions, while bright regions mean larger structural similarity and less quality degradation. As illustrated in Figure 4.18, after blepharoplasty (eyelid surgery), pouches in both lower eyelid regions are nearly removed. Obviously, the correspondig regions in the SSIM index map are darker than the neighborhood regions. Similar results can also be found in images after rhinoplasty (nose surgery) and lip augmentation. For global surgeries, we take laser skin resurfacing as an example. After this surgery, the chin, cheek and mouth regions present larger variations than other regions. Fortunately, these variations are well reflected in the SSIM map. From Figure 4.18, we can see the effectiveness of using SSIM index map to model both local and global variations caused by plastic surgery when matching a genuine pair (i.e, where the identity of the probe and gallery faces is the same).

Next we divide the two images as well as their corresponding SSIM index map into n patches of the same size. Denote the n patches as $\{p_1, p_2, ..., p_n\}$. For each patch p_i, feature extraction is then performed on the two images being compared. After feature extraction, we can employ a classifier c_i for each patch p_i. Denote the output of the classifier c_i as sc_i^o, i.e. the matching score of the i^{th} patches of the two images. In a conventional fusion scheme, the final matching score sc^o can



Blepharoplasty

Laser skin resurfacing

Impostor pair

Rhinoplasty

Lip augmentation

Impostor pair

Figure 4.18: The SSIM maps of some pre- and post-surgery image pairs.

be computed directly using the sum rule as follows.

$$sc^o = \sum_{i \in 1,\dots,n} sc_i^o \tag{4.5}$$

Note that this provides the same weights to all the patches. In our work, we give different weights to different patches according to the degree of their quality degradation. Specifically, the mean pixel value of each patch p_i in the SSIM map is calculated as the weight of the corresponding classifier c_i. We represent the weight of c_i as w_i, then in our SSIM index weighted multi-patch fusion scheme, the final weighted matching score sc^w is calculate as follows.

$$sc^w = \sum_{i \in 1,\dots,n} w_i \times sc_i^o \tag{4.6}$$

As described in Section 4.3.2, the values of each pixel in the SSIM map locate in a dynamic range of $[-1, 1]$. In other words, the calculated weights maybe smaller than zero. This is not in accordance with our objective, in which regions with large variations will be less considered or be ignored. There are two intuitive solutions to this problem. One can either normalize all the pixel values to $[0, 1]$ or normalize all the patch weights to $[0, 1]$. However, our experiments on the plastic surgery database show that both normalization solutions do not affect the recognition performance significantly in contrast with the original might minus pixel values or weights. Thereby, in all our experiments later, we directly use the original SSIM map for weight calculation.

Face identification using SSIM based patch fusion. As shown in Figure 4.17, given a probe image y and N gallery individuals, the SSIM maps of the probe y and each of the N gallery images are first calculated. After this, we can get a total of N SSIM maps. Next, SSIM index weighted multi-patch fusion scheme can be employed to match y with each of the N gallery images. Finally, we can get N SSIM index weighted matching scores $\{sc^{w1}, sc^{w2}, ..., sc^{wN}\}$. The final classification is performed as follows:

$$identity(y) = \arg \underset{j}{max}\{sc^{wj}\} \tag{4.7}$$

4.3.4 Analysis of SSIM Weighted Patch Fusion

The proposed approach employs the SSIM map in a patch level to provide different weights to different patches in accordance with the degree to which each patch is altered by surgeries. The advantages of the proposed approach are as follows:

1) Effectiveness for modeling variations caused by surgeries. Figure 4.18 well illustrated this.

2) No training cost. our approach does not employ any training procedures.

3) No background information employed. Background information [91] or cohort information [61], is extracted on an additional background dataset. Faces in the background dataset are disjoint by identity from the test faces. As is known, in most existing literature related to recognition across plastic surgery, a background dataset is collected to handle the problems caused by insufficient gallery images of each individual [5, 50]. Note that in our approach, we do not use any background information.

However, using our approach for matching faces of different individuals, also the regions where the two faces differ most are given lower weights. Two SSIM maps corresponding to two pre- and post-surgery impostor pairs (i.e., where the identity of the probe and gallery faces is not the same) are shown in Figure 4.18.

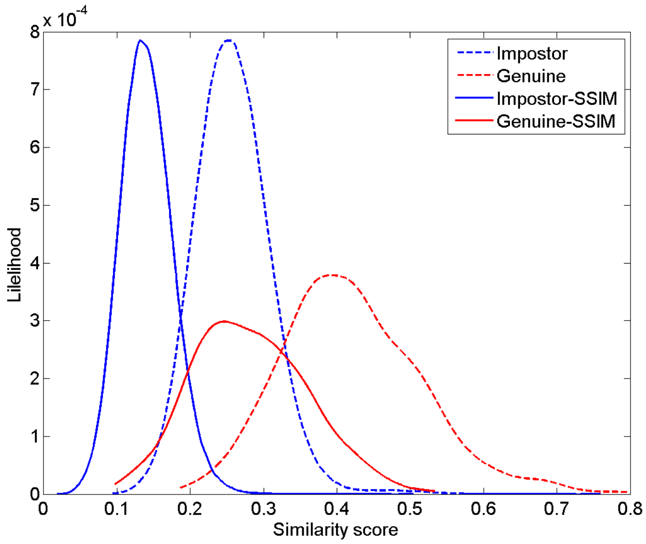


Figure 4.19: The score distributions of genuine and impostor before and after using SSIM quality.

For both impostor pairs, the two faces differ from each other significantly around the eye, nose and mouth regions. Even though these regions encode most of the discriminative information in faces, the SSIM mapping assigns lower weights in comparison with other regions, such as the forehead and cheek regions.

To well study the effect of the SSIM approach on genuine and impostor matchings, the two score distributions on the plastic surgery database are shown in Figure 4.19. The matching scores were computed on the entire faces by means of LBP features [6]. The total number of genuine and impostor scores are respectively N and $N \times (N-1)$, in this case $N = 784$. More details about the experimental settings can be found in Section 4.3.5. The dashed plots correspond to the distributions computed without using the SSIM map, while the solid plots correspond to the distributions resulting from the SSIM weighting. After the SSIM weighting, both the genuine and impostor scores decrease. Most importantly, the overlap region between the genuine and impostor distributions is reduced after the SSIM weighting.

4.3.5 Experiments and Results

The proposed approach was tested on a plastic surgery database containing $1,800$ images of 900 subjects. These images were collected from plastic surgery information websites, hence, many present a number of non-ideal factors such as duplicate entries, incomplete faces and non-frontal faces. Similarly to the pre-selection procedure used in [34], we excluded images which show the above negative factors. Finally, a dataset consisting of $1,576$ images from 784 subjects was selected, i.e., $N = 784$. In comparison with the evaluation schemes used in [78] which use 40% of the dataset for training and only the remaining 60% for testing, all our experiments were performed on the whole dataset.

For feature extraction, we employ two widely used facial features LBP [6] and Gabor [49]. In all our experiments, the size of each patch is set to be 8×8. When extracting LBP features, for each of the $(160 \div 8) \times (120 \div 8) = 300$ patches, we extract a 59-bin uniform LBP histogram. To compute the Gabor feature, we adopt five scales and eight orientations of the Gabor filters. After this, we can get a 40-dimension Gabor jet for each pixel. Directly using Gabor jets from all pixels in the 8×8 patch as the feature representation will result in a very large dimension feature vector $(64 \times 40 = 2,560)$, thus having a high computational cost. To address this problem, we use responses at a smaller set of pixels selected uniformly with a 8×8 down-sampling rate. The SSIM map is down-sampled using the same down-sampling rate. The down-sampled responses have only $(160 \div 8) \times (120 \div 8) = 300$ pixels, each of which corresponds to a 40-dimension Gabor jet. Here we use the down-sampled pixels to simulate the patches described above, and use the 40-dimension Gabor jet at each down-sampled pixel as the patch feature representation, thus having a much lower computational cost. For the matching score, we compute the cosine similarity between the two descriptor vectors.

To thoroughly validate the proposed approach, we perform face matching in both a holistic manner and a component-wise manner. We shall first treat the case of the holistic manner and return our discussion to the case of component-wise manner later.

Holistic matching. The whole face image is divided into a number of the above

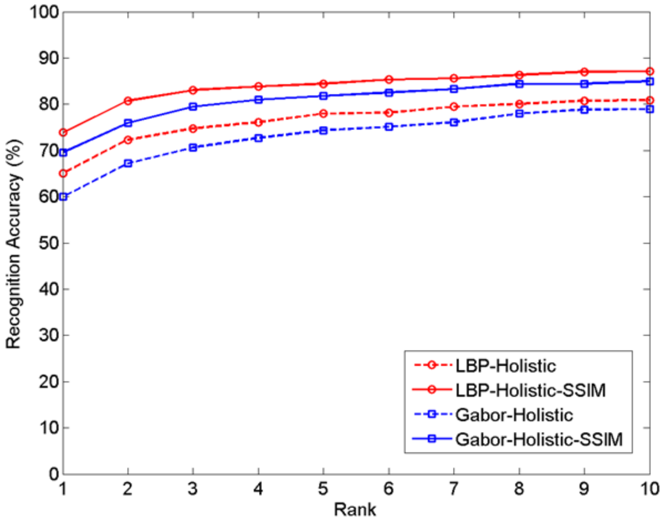


Figure 4.20: The CMC plots illustrating the performance using holistic manner with and without SSIM quality.

mentioned 8 × 8 patches. The accuracy is reported in terms of CMC curves. Figure 4.20 shows the CMC curves for: (a) holistic LBP without SSIM quality; (b) holistic LBP with SSIM quality; (c) holistic Gabor without SSIM quality; (d) holistic Gabor with SSIM quality. As expected, when using LBP and Gabor features in a holistic manner, SSIM weighted multi-patch fusion significantly outperforms approaches without using SSIM quality information. The Rank-1 accuracies of these scenarios were observed to be: (a) 65.05%; (b) 73.85%; (c) 59.95%; (d) 69.52%; SSIM quality improves the Rank-1 accuracy of LBP feature about 8.80%, while with Gabor feature, we can get an increased accuracy of almost 9.57%.

Component-wise matching. Seven facial regions are extracted, including: forehead, leftocolar, rightocular, nose, leftcheek, right cheek and mouth. Figure 4.21 shows the seven components and their size. The remaining process, including feature extraction and matching, is performed individually for each facial component. For the component-wise approach, we perform experiments using only LBP features. Figure 4.22 shows the individual CMC curves corresponding to the seven components. As we can see that, the proposed approach improves

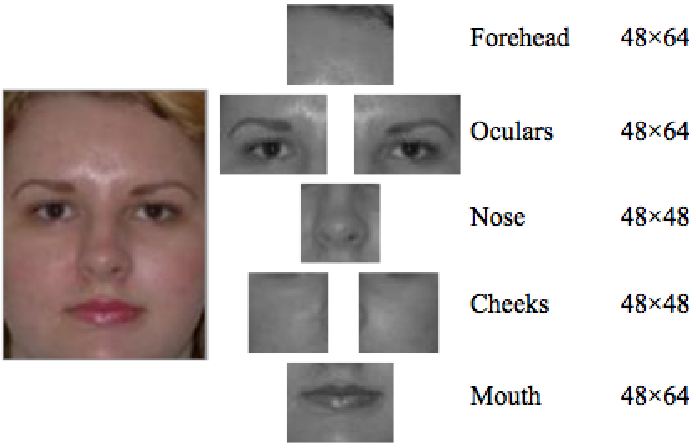
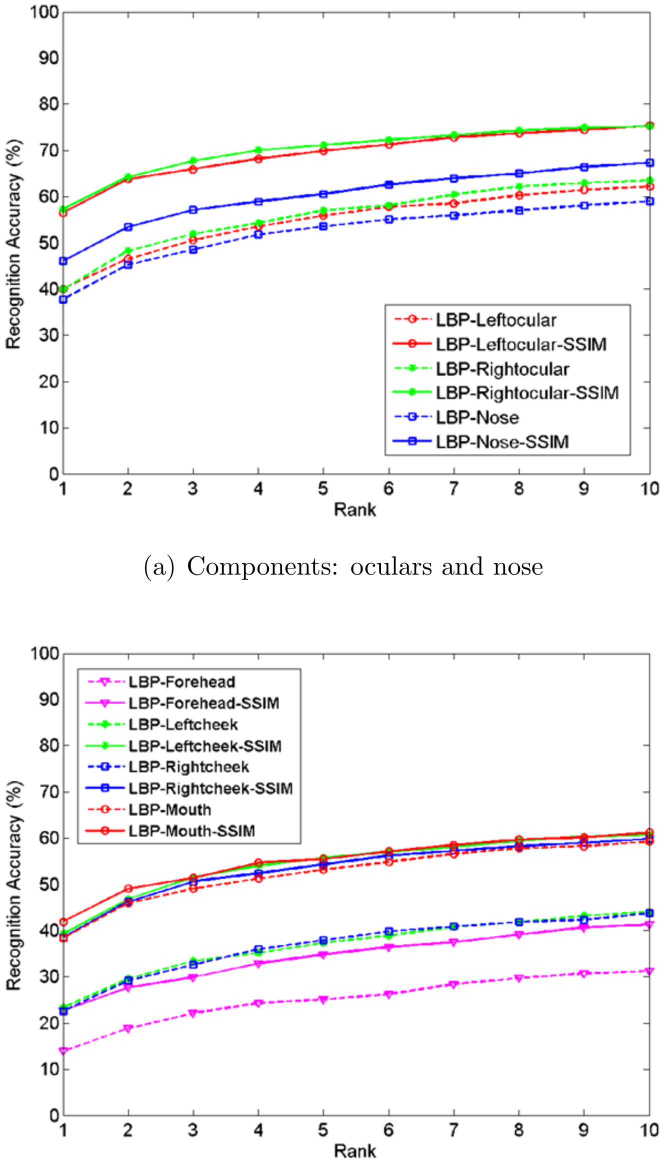


Figure 4.21: The 7 components used in the component-wise matching mechanism and their size.

the recognition accuracy significantly. The increased Rank-1 accuracy for all the seven components is here listed: forehead (8.80%), leftocular (16.46%), rightocular (17.35%), nose (8.29%), leftcheek (15.95%), rightcheek (15.94%) and mouth (3.57%).

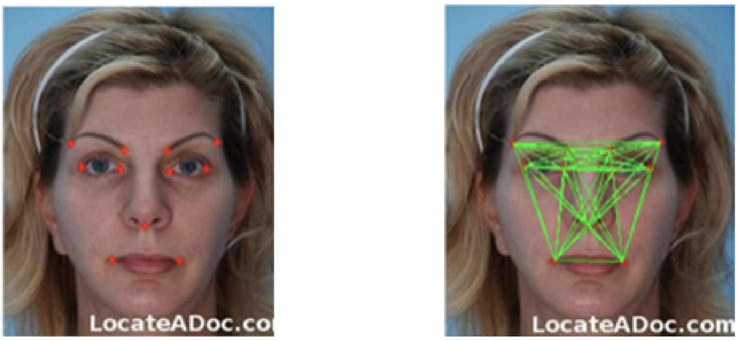
Next we fuse the outputs of the seven component classifiers for the final decision. Considering that matching faces using a component-wise manner might lose some useful information of the whole face geometric structure, we automatically locate 11 landmarks using STASM and construct 83 triangles using these 11 landmarks. For each triangle, we compute the radian values of the three angles. Further, we concatenate the three radian values of all the triangles to form a triangle descriptor representing the whole face geometric structure. Next the cosine similarity of two triangle descriptors from two faces is calculated for the geometric matching score. It is worth nothing that the triangle descriptor for each face is computed on the original image instead of the aligned one due to the scale-invariant property of congruent triangles. The 11 landmarks and the 83 triangles are illustrated in Figure 4.23. Finally, the rank-order list based fusion scheme proposed in [50] is employed to generate the final rank-order list by fusing the seven component classifiers and the triangle classifier. The weights for the seven component classifiers and the triangle classifier are respectively:



(a) Components: oculars and nose

(b) Components: forehead, cheeks and mouth

Figure 4.22: The CMC plots illustrating the performance of each individual component using LBP as feature. (a) Components: oculars and nose; (b) Components: forehead, cheeks and mouth.



(a) 11 landmarks

(b) 83 triangles

Figure 4.23: The 11 landmarks and 83 triangles for the calculation of the triangle descriptor. (a) 11 landmarks; (b) 83 triangles.

forehead (2); leftocular (4); rightocular (4); nose (2.5); leftcheek (3); rightcheek (3); mouth (3.5); triangle (4). The fusion results of component LBP features and triangle classifier as well as the result of using only the triangle classifier are illustrated in Figure 4.24. It can be observed that triangle descriptor shows some useful information. With SSIM index, a significant improvement is achieved in the component-wise manner. The Rank-1 accuracy increased from 69.13% to 77.55%.

Results on different individual plastic surgeries. We also report the effects of different kinds of plastic surgical procedures on the Rank-1 identification rate. We separately consider local and global surgeries in this experiment. Specifically, for local surgical procedures, we select from the plastic surgery database Brow lift, Blepharoplasty, Otoplasty and Rhinoplasty due to their large number of samples and organize other local surgeries with few samples into one group. We list the results of each of these surgeries as well as the average accuracy across all these local surgeries. Similarly, we report both the performance of the individual surgery and their average performance for global surgeries. However, in global surgeries, there is solely face lift which includes sufficient number of surgeries. Table 4.1 lists the results for: (a) holistic LBP; (b) component-wise LBP; (c) holistic Gabor without SSIM quality. And the results of using our proposed SSIM

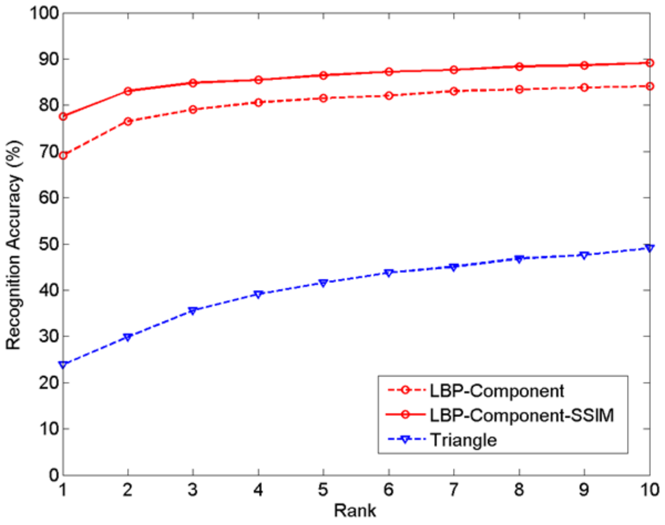


Figure 4.24: The CMC plots illustrating the performance using component-wise manner and solely triangle descriptor.

based patch fusion are shown in Table 4.2. As evaluated in the above experiments, our proposed framework significantly outperforms those methods without SSIM quality. For each of these three scenarios, we compute the increased accuracy of the average accuracies of all the local surgeries: (a) 9.17%; (b) 12.00%; (c) 7.29%. And the increased accuracies for the global surgeries are: (a) 8.36%; (b) 6.69%; (c) 9.75%. Furthermore, from Table 4.2, we can see that our approach can lead to much better performance when dealing with local surgeries than global ones.

Comparison with the state-of-the-art. Note that our experiments are performed without exploiting any outside dataset for providing auxiliary information [5, 50]. Hence, it is more reasonable to compare our approach with those existing approaches which do not use any cohort information. Table 4.3 shows the comparison results. For each algorithm, we show not only the Rank-1 accuracy but also whether it employs training and feature fusion or not. Furthermore, the number of gallery subjects and probe images used in each algorithm is illustrated for a better comparison. In [34], 661 pre-surgery images from the plastic surgery database and 568 images from the FRGC database [70] are used to form the gallery set, while the query images are the corresponding 661 post-surgery

Table 4.1: Effects of different plastic surgeries on Rank-1 accuracy for methods without SSIM quality.

Type	Surgery	LBP-Holistic	Gabor-Holistic	LBP-Component
Local	Brow lift	68.63%	54.90%	66.67%%
	Blepharoplasty	76.53%	65.31%	76.53%
	Otoplasty	63.16%	61.40%	71.93%
	Rhinoplasty	67.59%	60.00%	75.17%
	Others	62.16%	59.46%	72.97%
	All	**68.24**%	**60.71**%	**73.65**%
Global	Face lift	60.07%	58.66%	61.84%
	Others	65.79%	60.53%	71.05%
	All	**61.28**%	**59.05**%	**63.79**%

Table 4.2: Effects of different plastic surgeries on Rank-1 accuracy for methods with SSIM quality.

Type	Surgery	LBP-Holistic	Gabor-Holistic	LBP-Component
Local	Brow lift	76.47%	64.71%	76.47%
	Blepharoplasty	83.67%	77.55%	85.71%
	Otoplasty	73.68%	80.70%	82.46%
	Rhinoplasty	77.93%	68.97%	81.38%
	Others	71.62%	72.97%	75.68%
	All	**77.41**%	**72.71**%	**80.94**%
Global	Face lift	67.84%	65.02%	72.79%
	Others	76.32%	68.42%	76.32%
	All	**69.64**%	**65.74**%	**73.54**%

Table 4.3: The Rank-1 accuracies of our approach and those of the existing approaches on the plastic surgery database.

Algorithms	Rank-1	Training	Fusion	# gallerys	# probes
GNN [77, 78]	54.20%	yes	no	540	540
Verilook 3.2 [34]	73.90%	no	no	661PS + 568FRGC	661
Our approach (LBP)	**77.55%**	**no**	**no**	**784**	**784**
Evolutionary granular [14])	78.61%	yes	yes	540	540
Verilook 3.2 + Ocular [34]	81.50%	no	yes	661PS + 568FRGC	661

images. Verilook 3.2 is a commercial software from Neurotechnology [64]. From these results, we can see the effectiveness of our approach for face recognition across plastic surgery.

4.3.6 Conclusion

In this work, we introduced SSIM index weighted multi-patch fusion to face recognition across plastic surgery. Experimental results on the plastic surgery database highlight the effectiveness of our approach. In comparison with the existing approaches, neither training process is needed nor any background information is required. The good potential of SSIM quality map for face recognition after plastic surgery can be seen through our work.

4.4 Summary

In this section, we presented two categories of alterations: unintentional and intentional alterations. Further, we studied the impact of unintentional geometric alterations and intentional digital beautification of face photos on face recognition accuracy. From the experimental results, we found that state-of-the-art algorithms are able to overcome limited digital alterations but are sensitive to more relevant modifications. In addition, we developed two facial descriptors specially used for detecting aspect ratio alterations. Finally, we presented an effective algorithm for face recognition across plastic surgery. The proposed approach used the Structural Similarity quality map to detect and model variations caused by surgeries. Extensive experiments conducted on a plastic surgery face database demonstrated the potential of SSIM map for matching face images after surgeries.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The thesis investigated cohort score normalization, which aims to post-process the matching score using information from cohort samples, as means to improve the robustness of face recognition under challenging environments. In particular, we introduced cohort normalization to sparse presentation based undersampled face identification. We further presented a novel and effective cohort normalization approach specifically for unconstrained face pair matching. Experiments performed on several well known face databases show the effectiveness of cohort score normalization on measuring the degradation effect under challenging environments.

Furthermore, we advanced the state-of-the-art in cohort score normalization by providing a better understanding of the cohort behavior. By focusing on how to organize a good cohort set for face recognition, we investigated the impact of cohort sets of different sizes and cohort samples of different quality on the cohort normalization performance. In particular, we found that bigger cohort set size usually gives more stable and often better results to a point before the performance saturates. And cohort samples with different quality indeed produce different cohort normalization performance. Generally, using cohort samples of good quality can lead to much better performance the using bad cohort samples.

Face recognition across alteration is a very challenging problem for current face recognition algorithms. Alterations can be produced both unintentionally (e.g., by the acquisition or printing device) or intentionally (e.g., digital beautification or plastic surgery). In this thesis, we studied the impact of unintentional geometric alterations and intentional digital beautification of face photos on face recognition accuracy. From the experimental results, we found that state-of-the-art algorithms are able to overcome limited digital alterations but are sensitive to more relevant modifications. Furthermore, we developed two facial descriptors specially used for detecting aspect ratio alterations. Finally, we presented an effective algorithm for face recognition across plastic surgery. The proposed approach used the Structural Similarity quality map to detect and model variations caused by surgeries. Extensive experiments conducted on a plastic surgery face database demonstrated the potential of SSIM map for matching face images after surgeries.

5.2 Future Work

As biometric systems are expected to operate under challenging conditions, cohort normalization certainly offers an option to render the system more robust. As a future work, we would perform cohort normalization in a part-based manner, i.e., not focusing on the whole face recognition problem, but operating with image patches or facial parts. As shown in this work, cohort normalization can bring sparse representation based approach much robustness against various forms of degradation factors for undersampled face recognition. One interesting application of cohort score normalization might be face recognition from a single image per person.

In this work, we investigated the cohort behavior using polynomial regression based cohort normalization methods for the discriminative information, thus focusing on the discriminative cohort behavior. In the literature, there is much work focusing on organizing an additional cohort set and exploiting useful information as prior knowledge from these cohort samples. However, researchers try to exploit not only discriminative information but also some other helpful information such as identity and representative information from cohort samples. Thereby, it is interesting to get a thorough understanding of cohort behavior by using different cohort normalization approaches.

The good potential of SSIM quality map for face recognition after plastic surgery has been shown in our work. However, there are still several issues deserving more study on this direction.

1) Demand for more discriminative weights. The SSIM weighting can induce lower weights for regions where two different faces differ most. Hence, it is important to change the SSIM weights to make those regions more discriminative. A function may be learned mapping the SSIM weights to more discriminative values.

2) Extraction of holistic geometric information. For some plastic surgeries modifying the skin texture such as skin resurfacing and rhytidectomy, the geometric structure of the whole face will be of great importance. For instance, in our approach, the use of the triangle descriptor might be improved by employing more landmarks located in the outline of the face.

Bibliography

[1] Biometric deployment of machine readable travel documents, 2004. ICAO. 98

[2] Iso/iec 19794-5, information technology - biometric data interchange formats - part 5: Face image data, 2011. 98

[3] G. Aggarwal, N.K. Ratha, and R.M. Bolle. Biometric verification: Looking beyond raw similarity scores. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, pages 31–36, 2006. 31, 56, 73

[4] G. Aggarwal, N.K. Ratha, R.M. Bolle, and R. Chellappa. Multi-biometric cohort analysis for biometric fusion. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5224–5227, 2008. 31, 34, 56, 73

[5] G. Aggarwal, S. Biswas, P. Flynn, and K. Bowyer. A sparse representation approach to face matching across plastic surgery. In *IEEE Workshop on Applications of Computer Vision*, pages 113–119, 2012. 72, 115, 122, 129

[6] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004. 13, 14, 60, 77, 108, 123, 124

[7] American Aesthetic. `http://www.surgery.org/`. 27

[8] R. Auckenthaler, M. J. Carey, and H. L1oyd-Thomas. Score normalization

for text-independent speaker verification systems. *Digital Signal Processing*, 10:42–54, 2000. 6, 33

[9] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, 2002. 13

[10] P.N. Belhumeur and D.J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–277, 1996. 12

[11] P.N. Belhumeur and D.J. Kriegman. Face recognition under varying lighting conditions using self quotient image. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 819–824, 2004. 12

[12] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 12, 13

[13] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *British Machine Vision Conference*, 2012. 56

[14] H. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, and A. Noore. Evolutionary granular approach for recognizing faces altered due to plastic surgery. In *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, pages 720–725, 2011. 115, 131

[15] M. Bicego, A. Lagorio, E. Gross, and M. Tistarelli. On the use of sift features for face authentication. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 35–35, 2006. 13, 60, 108

[16] T. Bourlai, A. Ross, and A.K. Jain. On matching digital face images against passport photos. In *International Conference on Biometrics, Identity and Security*, pages 1–10, 2009. 98

[17] T. Bourlai, A. Ross, and A.K. Jain. Restoring degraded face images for matching faxed or scanned photos. *IEEE Transactions on Information Forensics and Security*, 6(2):371–384, 2011. 98

[18] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10): 1042–1062, 1993. 13

[19] C.J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998. 17

[20] C++ software library of STASM. http://www.milbo.users.sonic.net/stasm/. 108, 120

[21] T. Chen, W. Yin, X. Zhou, D. Comaniciu, and T. Huang. Total variation models for variable lighting face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1519–1524, 2006. 12

[22] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013. 39, 72

[23] W.H. Deng, J.N. Hu, and J. Guo. Extended src: undersampled face recognition via intra-class variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012. 6, 39

[24] D.L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006. 41

[25] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley Sons, 2012. 16

[26] Find a Doctor. http://www.locateadoc.com/. 27

[27] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition edition, 2007. 12

[28] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, pages 498–505, 2009. 67, 69

[29] T. Hasan and J.H.L. Hansen. A study on universal background model training in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1890–1899, 2011. 71

[30] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007. 3, 6, 24, 55, 60, 72

[31] A.K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross. Biometric: A grand challenge. In *IEEE international Conference on Pattern Recognition*, pages 935–942, 2004. 30

[32] A.K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005. 30

[33] A.K. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2008. 1, 19, 20, 46

[34] R. Jillela and A. Ross. Mitigating effects of plastic surgery: fusing face and ocular biometrics. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 402–411, 2012. 115, 124, 129, 131

[35] R. Jillela and A. Ross. Mitigating effects of plastic surgery: Fusing face and ocular biometrics. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems*, pages 402–411, 2012. XI, 96

[36] D.J. Jobson, Z. Rahman, and G.A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997. 12

[37] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002. 17

[38] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, pages 365–372, 2009. 55

[39] N. Kumar, A. Berg, P.N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011. 71

[40] L.I. Kuncheva and C.J. Whitaker. *Multiple Classifier Systems*, chapter Feature subsets for classifier combination: an enumerative experiment, pages 228–237. Springer Berlin Heidelberg, 2001. 13

[41] Shen L and Bai L. A review on gabor wavelets for face recognition. *Pattern analysis and applications*, 9(2-3):273–292, 2006. 13

[42] S. Lawrence, C.L. Giles, A. Tsoi, and A. Back. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997. 13

[43] T.S. Lee. Image representation using 2-d gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996. 13

[44] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013. 71

[45] S.Z. Li and A.K. Jain. *Encyclopedia of Biometrics*. Springer, 2009. 1

[46] S.Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. 10

[47] S. Liao, A.K. Jain, and S.Z. Li. Partial face recognition: alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, 2013. 71

[48] LiftMagic - Instant cosmetic surgery and anti-aging makeover tool. http://makeovr.com/liftmagic. 93

[49] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476, 2002. 13, 14, 46, 47, 60, 77, 108, 124

[50] X. Liu, S. Shan, and X. Chen. Face recognition after plastic surgery: a comprehensive study. In *Asian Conference on Computer Vision*, pages 565–576, 2012. 115, 122, 126, 129

[51] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 13, 15, 100

[52] Luxand Inc. Luxand Web Site. http://luxand.com. 100

[53] D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, 2nd edition, 2009. 20, 101

[54] B.S. Manjunath, R. Chellappa, and C.V.D. Malsburg. A feature based approach to face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 373–378, 1992. 13

[55] M. D. Marsico, M. Nappi, D. Riccio, and H. Wechsler. Robust face recognition after plastic surgery using local region analysis. In *Image Analysis and Recognition*, pages 191–200, 2011. 115

[56] W.D. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25(1):71–102, 1987. 31

[57] A. Martinez and R. Benavente. The ar database. Technical Report 24, CVC Technical Report, 1998. 20, 46, 100

[58] A.M. Martnez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002. 10, 13

[59] A.M. Martnez and A.C. Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001. 21

[60] MATLAB implementation of SSIM. https://ece.uwaterloo.ca/~z70wang/research/ssim/. 118

[61] A. Merati, N. Poh, and J. Kittler. User-specific cohort selection and score normalization for biometric systems. *IEEE Transactions on Information Forensics and Security*, 7(4):1270–1277, 2012. VIII, 35, 36, 39, 42, 46, 58, 59, 61, 62, 63, 78, 122

[62] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conference on Computer Vision*, pages 504–513, 2008. 108, 120

[63] S. Milborrow and F. Nicolls. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579, 2012. 71

[64] Neurotechnology Inc. Neurotechnology web site. http://www.neurotechnology.com. 100, 131

[65] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 13, 14

[66] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by vi? *Vision Research*, 37(23):3311–3326, 1997. 38

[67] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M.R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J.-L. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, S. Garcia-Salicetti, L. Allano, B. Ly-Van, B. Dorizzi, J. Kittler, T. Bourlai, N. Poh, F. Deravi, R. Ng, M. Fairhust, J. Hennebert, A. Humm, M. Tistarelli, L. Brodo, J. Richiardi, A. Drygajlo, H. Ganster, F. Sukno, S.-K. Pavani, A. Frangi, L. Akarun, and A. Savran. The multi-scenario multi-environment biosecure multimodal

database (bmdb). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1097–1111, 2010. 35

[68] U. Park, R.R. Jillela, A. Ross, and A.K. Jain. Periocular biometrics in the visible spectrum. *IEEE Transactions on Information Forensics and Security*, 6(1):96–106, 2011. 115

[69] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000. 21, 50

[70] P.J. Phillips, P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, L. Hoffman, J. Marques, J. Min, and W.J. Worek. Overview of the face recognition grand challenge. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005. 7, 24, 72, 74, 75, 89, 129

[71] N. Pinto, J.J. DiCarlo, and D.D. Cox. How far can you get with a modern face recognition test set using only simple features? In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2591–2598, 2009. 71

[72] N. Poh and S. Bengio. Database, protocol and tools for evaluating score-level fusion algorithms in biometric authentication. *Pattern Recognition*, 39 (2):223–233, 2005. 62

[73] N. Poh and J. Kittler. *Advances in Biometrics*, chapter On the use of log-likelihood ratio based model-specific score normalisation in biometric authentication, pages 614–624. Springer Berlin Heidelberg, 2007. 30, 33

[74] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Transactions on Image Processing*, 15(11):3349–3361, 2006. 5

[75] A.E. Rosenberg and S. Parthasaraty. Speaker background models for connected digit password speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 81–84, 1996. 31

[76] A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang, and F.K. Soong. The use of cohort normalized scores for speaker verification. In *International Conference on Spoken Language Processing*, 1992. 31

[77] R. Singh, M. Vatsa, and A. Noore. Face recognition with disguise and single gallery images. *Image and Vision Computing*, 27:245–257, 2009. 131

[78] R. Singh, M. Vatsa, H. Bhatt, S. Bharadwaj, A. Noore, and S. Nooreyezdan. Plastic surgery: a new dimension to face recognition. *IEEE Transactions on Informtion Forensics and Security*, 5(3):441–448, 2010. XI, 5, 27, 95, 97, 114, 115, 116, 124, 131

[79] J.A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 17

[80] X. Tan, S.C. Chen, Z.H. Zhou, and F. Zhang. Recognizing partially occluded, expression variant faces from single training image per person with som and soft knn ensemble. *IEEE Transactions on Neural Networks*, 16(4): 875–886, 2005. 13

[81] X. Tan, S. Chen, Z.H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006. 5, 39

[82] S. Tulyakov, Z. Zhang, and V. Govindaraju. Comparison of combination methods utilizing t-normalization and second best score model. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, pages 1–5, 2008. 34

[83] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 13

[84] G. Vass and T. Perlaki. Applying and removing lens distortion in post production. In *Hungarian Conference on Computer Graphics and Geometry*, 2003. 91

[85] P. Viola and M.J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 17

[86] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012. 39

[87] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9): 1222–1228, 2004. 13, 38

[88] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, 70(1):91–104, 2006. 13, 38

[89] Z. Wang, Bovik A.C, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7, 115, 117, 120

[90] L. Wiskott, R. Fellous, N. Kruger, and C. von Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 13

[91] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *Asian Conference on Computer Vision*, pages 88–97, 2009. 55, 60, 67, 69, 122

[92] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *IEEE International Conference on Computer Vision*, pages 897–902, 2009. 55

[93] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10): 1978–1990, 2011. 71

[94] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 38, 46, 47

[95] A.Y. Yang, S.S. Sastry, A. Ganesh, and Y. Ma. Fast 1-minimization algorithms and an application in robust face recognition: A review. In *IEEE International Conference on Image Processing*, pages 1849–1852, 2010. 38

[96] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *European Conference on Computer Vision*, pages 448–461, 2010. 39

[97] P. Yang, S. Shan, W. Gao, S.Z. Li, and D. Zhang. Face recognition using ada-boosted gabor features. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 356–361, 2004. 17

[98] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 497–504, 2011. 56

[99] C. Zhang and Z.Y. Zhang. A survey of recent advances in face detection. Technical report, Microsoft Research, 2010. 11

[100] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: a literature survey. *ACM Computing Survey*, 35(4):399–458, 2003. 1, 5, 14