Face recognition based on multi-scale local features

Geng, Cong

2012

Geng, C. (2012). Face recognition based on multi-scale local features. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/50725>

<https://doi.org/10.32657/10356/50725>

NANYANG TECHNOLOGICAL UNIVERSITY

Face Recognition Based on Multi-scale Local Features

A Thesis Submitted to the
School of Electrical and Electronic Engineering
of the Nanyang Technological University

by

Geng Cong

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2012

Abstract

The ability to recognize human faces is a demonstration of incredible human intelligence. Over the last four decades, attempts from diverse areas are made to replicate this outstanding visual perception of human beings in machine recognition of faces. Within the face recognition literature, researchers have centered the debate on how human beings perceive human faces and this has become an important and active research area. Psychologists concluded that holistic and local feature based approaches are dual routes to the face recognition. Although holistic based approaches have attained certain level of maturity, in general they require a preprocessing procedure to normalize the face image variations in pose, scale and illumination. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. As a result, most approaches work on the normalized face images based on the manually identified landmarks. The recognition performance deteriorates considerably if the manual process is replaced by an automatic landmark detection algorithm. Moreover, global features are sensitive to image variations in scale, facial expression, pose and occlusion. Most of the holistic approaches are dependent on the training databases because knowledge about the face discrimination is generalized by machine learning from the face samples. A representative training database is necessary, which, however, is not available in many applications.

In contrast to holistic methods, local feature based approaches are more robust to variations in pose, scale, expression and occlusion. Elastic bunch graph matching (EBGM) and active appearance models (AAM) fall into this category. In EBGM, faces are represented as graphs, with nodes positioned at facial landmarks and edges labeled with distance vectors. Each node contains a set of Gabor wavelet coefficients. The AAM can match a class of deformable objects, for example faces with different expressions.

However, the performances of both EBGM and AAM depend on a good selection of facial landmarks, which are often annotated manually. This makes the approaches semi-automatic and labor consuming.

One of the very fundamental problems arising when analyzing face images originates from the fact that face structures appear in different ways, depending upon the scales of observations. First, local facial structures are shown at different levels of scales, ranging from skin textures at fine scales, through eyes and mouth represented at median scales to the shape of face contours at large scales. Second, the characteristic or the description of a local structure is strongly dependent on the scale at which the structure is modeled. Third, it is unknown in advance what the proper scales are to describe different local structures of unknown face images. To cope with these problems, an image representation that explicitly incorporates the notion of scale is a crucially important tool. The scale invariant feature transform (SIFT) detects distinct local structures from images and selects appropriate scales to describe them automatically. It has shown good performance on object detection and some other machine vision applications. Recently, some initial attempts apply the SIFT algorithm in the face recognition task. However, there are still many outstanding issues and problems that need to be addressed and solved if I am to leverage the idea of SIFT and some of its good properties to solve the challenging face recognition problem. Although some local feature based approaches achieve better recognition performance than holistic approaches, their computational burden is much heavier.

In this thesis, I firstly analyze the merits and deficiencies of SIFT and propose new strategies for feature extraction and image matching, which leads to a multi-scale local feature extraction and matching framework (LFEM) that overcomes some limitations of SIFT in solving the face recognition problem. In the original SIFT algorithm, a keypoint is selected only if it is larger than 26 neighbor points or smaller than all of them. This keypoint detection method works well for rigid visual objects, which have sharp transitions between different sides of an object. In other words, there are distinct corner structures with high contrast in such objects. However, human faces are non-rigid, round and smooth. There are few obvious blobs and corners with high contrast, as the intensity changes in face images are gradual and slow in the most areas. On the other hand, the

shape of the structures could be complex and some structures are close to each other or overlap. As a result, many local structures in the smooth area such as forehead, cheeks and chin cannot be detected due to the strict condition of comparing the 26 neighbors. Hence I propose a new approach to keypoint detection which captures the information of many facial structures in the smooth area such as forehead, cheeks and chin. A partial descriptor is designed to represent the keypoints whose support areas exceed the face image. My proposed detection approach and my partial descriptor strategy produce a rich number of keypoints. As a significant keypoint should be distinct from others in terms of either its location or the image structures of its neighborhood, I further propose to remove keypoints based on their distinctiveness. A two-stage image matching scheme and a strategy of keypoint search for the nearest subjects are developed to cater for the identification task. It circumvents the problem that the most similar local structures to the probe image disperse to many different gallery images. A training procedure is developed to achieve higher recognition accuracies if multiple training samples per subject are available. It contains template selection, template synthesis and unstable keypoint removal to meet different requirements of face recognition applications.

Furthermore, I propose a fully automatic face recognition framework based on both the local and global features (FAFF). As mentioned above, holistic approaches are very popular in recent years due to their good performance and low computational complexity. However, the holistic approaches require a preprocessing procedure to normalize the face image variations in pose and scale, which is usually done based on the localizations of semantic local features such as eyes, corners of mouth, nostrils and so on. However, in many real applications the appearances of these semantic features are not distinct or missing due to changes in expressions, occlusions, illuminations or image noise. Hence most of the research papers work on the manually pre-normalized faces. The recognition performance deteriorates considerably if the manual process is replaced by an automatic landmark detection algorithm. In contrast to holistic approaches, the local feature based approach LFEM proposed by me is more robust to image variations in pose and scale. Unlike the holistic approaches, the face normalization is an integrated part of the LFEM approach. And it achieves significantly better recognition performance than many popular holistic approaches, which is validated in the experiments. However, its

computational time is much longer. For instance, to fulfill the face recognition task, one must search all the images in the database and compare each local feature in every image, which causes very heavy computational burden. To solve these two critical problems in holistic and local feature based approaches, I propose a fully automatic face recognition framework based on both the local and global features. To speed up the local feature based approach, I propose to integrate the local feature based approach and the holistic approach in a cascaded way. The holistic approach is used as a filter to retrieve some candidate face images from the whole gallery set. The selected face images have higher probabilities matching to the probe, than the remains in the gallery. They form a new gallery set with reduced size, on which I perform the local feature based approach for face recognition. The reduction in the size of the gallery set speeds up the recognition process of LFEM. To solve the alignment problem in holistic approaches, I design a face alignment scheme based on multi-scale local features instead of semantic facial features. In face images, even if the appearances of facial components change drastically, there are many other non-semantic features holding distinct information, which can be utilized in the alignment process. Hence, I propose to align face images based on non-semantic multi-scale local features. This fully automatic framework not only speeds up the local feature based approach, but also improves the recognition accuracy comparing with the holistic and local approaches as shown in the experiments.

Acknowledgments

First and foremost, I would like to express my great appreciation to my advisor, Dr. Xudong Jiang, for his invaluable directions and support throughout my research efforts towards this dissertation. His insights and suggestions to the research area enlightened me in various detailed aspects throughout the work. His directions with regards to the development of my academic thinking and writing were inspiring and helpful for my current and future work.

Secondly, I would give my thankfulness to all students of our workstation resource lab for their patience of listening and their sparkling minds.

Finally, I wish to say thanks to my family and all the people who provided me with continued support and understanding.

Contents

Abstract	i
Acknowledgments	v
List of Figures	x
List of Tables	xiv
1 Introduction	1
1.1 An Overview of Face Recognition	1
1.2 Motivation for This Research	4
1.2.1 The Importance of Face Recognition and its Mechanism	4
1.2.2 Research Motivation	5
1.3 Major Results	9
1.4 Dissertation Organization	9
2 Literature Survey	11
2.1 Holistic Methods	11
2.1.1 Subspace Feature Extraction Approaches	12
2.1.1.1 Eigenfaces Method	12
2.1.1.2 Linear Discriminants: Fisherfaces	14
2.1.1.3 Locality Preserving Projections: Laplacianfaces	16
2.1.1.4 Marginal Fisher Analysis	19
2.1.1.5 Eigenfeature Regularization and Extraction	21
2.1.2 Holistic Feature Matching Approaches	22
2.1.2.1 Bayesian Methods	23
2.1.2.2 Support Vector Machines	25
2.1.2.3 Sparse Representation	25

2.2	Feature/Component based Approaches	29
2.2.1	Earlier Approaches	29
2.2.2	Gabor Features	31
2.2.3	Shape Based Approaches	34
2.2.4	Local Binary Pattern	34
2.2.5	Local Feature Detectors and Descriptors	37
2.2.5.1	Local Feature Detectors	37
2.2.5.2	Local Feature Descriptors	49
2.2.5.3	Scale Invariant Feature Transform	51
2.2.6	Face Recognition based on Multi-scale Local Features	57
2.3	Databases	65
2.4	Summary	66
3	Multi-scale Local Feature Extraction	68
3.1	Introduction	68
3.2	Keypoint Detection and Scale Selection	72
3.3	Descriptor Representation	76
3.4	Insignificant Keypoint Removal	78
3.4.1	Volume-SIFT Approach	78
3.4.2	Eliminating Similar Keypoints	81
3.5	Experimental Results	84
3.5.1	Validation of the Keypoint Detector	86
3.5.2	Validation of Feature Detection and Description in Face Recognition	96
3.5.2.1	Results on FERET Database I	96
3.5.2.2	Results on FERET Database II	97
3.5.2.3	Results on ORL, GT and AR Databases	99
3.5.3	Validation of Insignificant Keypoint Removal	100
3.6	Summary	102

4	Image Matching Approach and Its Application	103
4.1	Image Matching Approach	104
4.1.1	Introduction	104
4.1.2	Search the k -Nearest Neighbors of the Nearest Subject and Affine Transform Estimation	108
4.1.3	Further Matching between Probe and Each Candidate Gallery Image	111
4.2	Training for Multiple Image Samples per Subject	113
4.2.1	Template Selection	113
4.2.2	Unstable Keypoint Removal	114
4.2.3	Template Synthesis	115
4.3	Experimental Results	117
4.3.1	Validation of Image Matching Approach	117
4.3.1.1	Results on FERET Database I	117
4.3.1.2	Results on FERET Database II	118
4.3.1.3	Results on ORL database	119
4.3.1.4	Results on GT database	120
4.3.2	Validation of Template Selection and Synthesis	121
4.3.2.1	Validation of Template Selection for Multiple Samples per Subject	121
4.3.2.2	Validation of Unstable Keypoint Removal	123
4.3.2.3	Validation of Template Synthesis	124
4.4	Summary	125
5	Fully Automatic Face Recognition Framework Based on the Multi-scale Local Features	127
5.1	Introduction	127
5.2	Framework Outline	130
5.3	Face Alignment	132
5.3.1	Generate the Common Face Template	132
5.3.2	Establish the Feature Correspondences	135
5.3.3	Face Alignment by Similarity Transformation	136

5.4	Fully Automatic Face Recognition Framework based on Local and Global Features	137
5.4.1	Eigenfeature Regularization and Extraction	137
5.4.2	Fully Automatic Face Recognition Framework based on both the Local and Global Features	139
5.5	Experimental Results	141
5.5.1	Validation of My Face Alignment Strategy	141
5.5.1.1	Results on AR Database	141
5.5.1.2	Results on GT Database	143
5.5.1.3	Results on ORL Database	145
5.5.2	Validation of My Fully Automatic Face Recognition Framework .	147
5.6	Summary	149
6	Conclusions	151
6.1	Principle Contributions	154
6.2	Future Directions	155
	References	157
	List of Publications	174

List of Figures

1.1	Structure of a Face Recognition System	5
1.2	Facial Components Modeled at Different Scales	7
2.1	The concept of PCA/KLT. (a) Solid lines: the original basis; dashed lines: the PCA basis. The dots are selected at regularly spaced locations on a straight line rotated at 30° and then perturbed by isotropic 2D Gaussian noise. (b) The projection (1D reconstruction) of the data using only the first principal component [1].	13
2.2	The real part of the Gabor kernels at five scales and eight orientations. .	32
2.3	The basic LBP operator [2].	35
2.4	SUSAN corners are detected by segmenting a circular neighborhood into similar (orange) and dissimilar (blue) regions. Corners are located where the relative area of the similar region (USAN) reaches a local minimum below a certain threshold [3].	38
2.5	Illustration of pixels examined by the FAST detector [3].	39
2.6	The corner point $\mathbf{p}$ is the starting point of the edge-based region detector which exploits nearby edge information.	41
2.7	Illustration of the components of the Hessian matrix and Hessian determinant [3].	43
2.8	Construction of intensity-based regions [3].	45
2.9	Overview of local feature detectors [3].	47

2.10	(a) For each octave of scale space, the initial image is repeatedly convolved with Gaussian to produce the set of scale space images shown on the left. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images on the right. After each octave, the Gaussian image is down-sampled by a factor of 2, and the process repeated; (b) Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with cross) to its 26 neighbors in 3×3 regions at the current and adjacent scales (marked with circles).	52
2.11	Facial Components Modeled at Different Scales	57
3.1	Problem of the SIFT keypoint detection.	74
3.2	Extrema of the DoG images are detected by comparing a pixel (marked with cross) to its 8 neighbors at the current scale and the corresponding pixels at the adjacent scales (marked with circle).	75
3.3	(a) Initial keypoints detected by the original method (in blue/big dots) and extra keypoints detected by the proposed approach (in red/small dots). (b) Remaining blue keypoints after keypoint removal by low contrast (as described in Section 2.2.5.3); (c) Remaining blue keypoints after further keypoint removal based on the ratio of principal curvatures (as described in Section 2.2.5.3).	76
3.4	An illustration of partial-descriptor: blocks marked with cross are considered as invalid.	77
3.5	Symmetric Gaussian blobs with different scales and their corresponding one dimensional illustrations.	80
3.6	One-dimensional outputs of I_1 and I_2 .	81
3.7	Keypoints detected by SIFT in artificially produced images: (a) Initial keypoints detected in the clean image; (b) Initial keypoints detected in the image imposed with $\pm 4\%$ random noises; (a') Keypoints remain in (a) after applying a threshold on the minimum contrast; (b') Keypoints remain in (b) after applying a threshold on the minimum contrast; (a'') The final keypoints remain in (a') after further applying a threshold on the ratio of principal curvatures; (b'') The final keypoints remain in (b') after further applying a threshold on the ration of principal curvatures.	83

3.8	(a) Sample images used in the holistic approaches; (b) Sample images used in the multi-scale local feature based approaches (SIFT and my proposed frameworks).	85
3.9	Database. (a) Viewpoint change, (b) Viewpoint change, (c) Zoom+rotation, (d) Zoom+rotation, (e) Image blur, (f) Image blur, (g) JPEG compression, (h) Light change. The left most image of each set is used as the reference image in the experimental comparisons.	87
3.10	Performance evaluation of various keypoint detectors (Harris-Affine (haraff) [4], Hessian-Affine (hesaff) [4], maximally stable extremal regions (mseraf) [5], intensity-based regions (ibraff) [6, 7], edge-based detectors (ebraff) [8, 7], Difference-of-Gaussians (DoG) [9, 10], and the proposed) under viewpoint change. (a) Repeatability score (overlap 40%, normalized size = 30 pixels). (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	88
3.11	Performance evaluation of various keypoint detectors under viewpoint change for the textured scene. (a) Repeatability score. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	89
3.12	Performance evaluation of various keypoint detectors under scale change for the structured scene. (a) Repeatability score for scale change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	90
3.13	Performance evaluation of various keypoint detectors under scale change. (a) Repeatability score for scale change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches. . .	91
3.14	Performance evaluation of various keypoint detectors under blur change for the structured scene. (a) Repeatability score for blur change. (b) Absolute number of corresponding regions. (c) Matching score. (d) Absolute number of correct matches.	92

3.15	Performance evaluation of various keypoint detectors under blur change for the textured scene. (a) Repeatability score for blur change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	93
3.16	JPEG compression. (a) Repeatability score for different JPEG compression. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	94
3.17	Illumination change. (a) Repeatability score. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.	95
3.18	Recognition rate of different keypoint removal approaches obtained in (a) the first experimental setting (b) the second experimental setting.	101
4.1	The process of keypoints matching strategy proposed by Lowe: the number of keypoints matched in the right image is 4 and the number of keypoints matched in the wrong image is 5.	106
4.2	The process of keypoints matching strategy proposed by us: the number of keypoints matched in the right image is 16 and the number of keypoints matched in the wrong image is 7.	107
4.3	The geometric relationship between points p, p' , and points q, q'	109
4.4	Generalized Hough Transform bins with different values.	110
4.5	An illustration of the process to remove keypoints with low repeatability.	116
4.6	An illustration of the process of template synthesis.	116
5.1	Fully automatic face recognition framework based on both the local and global features.	131
5.2	Mean face $\mathbf{m}$ with (a) SIFT keypoints; (b) Symmetric keypoints $\mathcal{K}$	133
5.3	(a) Sample images before alignment; (b) Sample images after alignment.	136
5.4	Sample images of occluded face.	142
5.5	Cumulative matching curves of different alignment approaches obtained by ERE on (a) AR database; (b) GT database; (c) ORL database.	145

List of Tables

3.1	Recognition rate on FERET database I	97
3.2	Average recognition rate and standard deviation on FERET Database II	98
3.3	Recognition rates on ORL, GT and AR databases	100
4.1	Recognition rate with single template per subject on FERET database I	118
4.2	Average recognition rate and its standard deviation with single template per subject on FERET database II	119
4.3	Average recognition rate and its standard deviation with single template per subject on ORL database	119
4.4	Recognition rate with multiple templates per subject on GT database . .	121
4.5	Recognition rate on ORL, GT and AR databases	122
4.6	Average number of keypoints per template and recognition rate under different threshold T	123
4.7	Recognition rate of single template on ORL database	124
5.1	Recognition rate on AR database	143
5.2	Recognition rate on GT database	144
5.3	Recognition rate on ORL database	146
5.4	Percentage of the gallery images selected by ERE where FAFF achieves the same recognition performance as LFEM on the original database at r_1 and the best performance at r_2	147
5.5	Recognition rate on AR, GT and ORL databases	148

Chapter 1

Introduction

The topic of this PhD dissertation is the face recognition based on the multi-scale componential features and statistical learning. As the ability to recognize human faces is a demonstration of incredible human intelligence, over the last three decades researchers from diverse areas have been making attempts to replicate this outstanding visual perception of human beings in machine recognition of faces [11]. However, there are still substantial challenging problems such as intraclass variations in 3D poses, facial expressions, make-up and lighting conditions as well as occlusions and cluttered backgrounds.

This chapter provides a general introduction of this PhD dissertation. In Section 1.1 I briefly review current face recognition technologies and their relations. Section 1.2 explains the motivation for the researches on this topic and the focus of my research work. Then I present the major results of my research work in Section 1.3. Finally Section 1.4 describes the organization of this dissertation.

1.1 An Overview of Face Recognition

Face recognition is an important research problem spanning numerous fields and disciplines, which has attracted researchers who have different backgrounds: psychology, pattern recognition, neural networks, computer vision and graphics. Generally, face recognition can be formulated as follows: identification, verification and watch-list screening.

Identification is to match a biometric record from a probe against an entire database of similar biometric records to determine the probe's identity: a one-to-many comparison. The verification process confirms that a person is who he or she claims to be by matching the biometric record against that of his or her claimed identity: a one-to-one comparison. In watch-list screening, images captured by watch list cameras are automatically compared against criminal or other agency databases. If a match is found, alerts are sent out from the comprehensive surveillance solution to the common security interface. This watch list combines both identification and verification processes.

The literature on face recognition is vast and diverse. Often, a single system involves techniques motivated by different principles. The usage of a mixture of techniques makes it difficult to classify these systems based purely on what types of techniques they use for feature representation or classification. To have a clear and high-level categorization, face recognition algorithms can be classified into two different groups as follows:

- Component/Feature based approaches. The general idea of these methods is to first locate several facial component (features), and then classify the faces by combining and comparing the corresponding local statistics.
- Holistic matching methods. These methods use the whole face image as the raw input to a recognition system.

The earliest approaches in computer recognition of faces are mainly component based approaches [12, 13, 14]. Some of these approaches need manually entering feature points into the computer, which is very labor-consuming. And in the early component based approaches, the features are distances and angles between points such as eye corners, mouth extremities, nostrils and chin top. However, there is no method which can automatically detect facial components as accurately as human beings, various errors occur during the

component detection process. Due to above shortcomings, these purely geometry-based approaches are not robust to automatic face recognition issues.

Images of faces, represented as high-dimensional pixel arrays, often belong to a manifold of intrinsically low dimension. If the whole face region is used as input in holistic methods, it is very computationally expensive. Much of the surface of a face is smooth and has regular texture. Therefore, per-pixel sampling is unnecessarily dense: The value of a pixel is typically highly correlated with the values of the surrounding pixels. And the appearance of faces is highly constrained, for example, any frontal view of a face is roughly symmetrical, has eyes on the sides, nose in the middle. A vast proportion of the points in the image space does not represent physically possible faces. Thus in holistic matching methods, I mainly review those techniques which project face images into low-dimensional subspaces [1].

Holistic methods work well for classifying frontal views of faces in normal conditions. However, they are not robust against pose changes, illumination variations or occlusions. Moreover, generalizability problem remains a huge bottleneck for most holistic methods. To solve these problems in holistic approaches, in the mid 1990s, researchers paid attention to local feature based approaches again. But the local features are different from the componential features used in the early days. These local features or local statistics are extracted from the whole face image, which do not represent semantic facial components such as eyes, nose and mouth explicitly. Local feature based approaches are more robust to variations in pose, illumination conditions and occlusions. Moreover, to extract local features from face images, one does not need a training stage necessarily, which has naturally avoided the generalizability problem.

1.2 Motivation for This Research

The work described in this dissertation focuses on the 2D face recognition using multi-scale local features. The area being studied is: The extraction and matching of multi-scale local features, which finally leads to a fully automatic face recognition framework.

At the beginning of my research for the 2D face recognition problem, I find that most of the approaches are based on the holistic approaches, such as discriminant subspace extraction, which inherently suffer from the generalizability problem due to the unpredictable distribution of the real-world testing images, which might differ dramatically from that of the training samples. To solve this problem, I turn to local feature based approaches, where training stage is not needed necessarily. However, there are many problems in local feature based approaches, such as half-automation, instability, sensitivity to noise and so on. In the following, I first discuss the importance of face recognition in the biometrics world. After that, I state my motivations in using multi-scale local features for 2D face recognition.

1.2.1 The Importance of Face Recognition and its Mechanism

Many recent events, such as terrorist attacks, exposed serious weakness in most sophisticated security systems. The current state-of-the-art technologies of using a personal identification number or password for these purposes are inadequate because they are revealable, transferable and difficult to manage/remember. Various government agencies are now more motivated to improve security data systems based on body or behavioral characteristics, often called biometrics [15]. It is a very attractive technology, because it can be integrated into any application requiring security or access control, effectively eliminating risks associated with less advanced technologies that are based on what a person has or knows rather than whom a person really is. The most common biometrics are fingerprints and iris. However, iris recognition is expensive to implement and

not very accepted by people. Fingerprints require that the subject cooperate in making physical contact with the sensor surface. This raises issues of how to keep the surface clean and germ-free in a high-throughput application. On the contrary, face recognition seems to be a good compromise between reliability and social acceptance and balances security and privacy well. Moreover, because of the wide range of commercial and law enforcement applications together with the availability of feasible technologies after more than 40 years of research, face recognition has received significant attention.

Figure 1.1: Structure of a Face Recognition System

A face recognition system generally consists of four processing parts as depicted in Figure 1.1: Face detection, face alignment, facial feature extraction and face classification. *Face detection* provides information about the location and scale of each detected face. In *face alignment*, facial components, such as eyes, nose, mouth and facial outline are located, and thereby the input face is normalized in geometry and photometry. In *feature extraction*, features useful for distinguishing between different persons are extracted from the normalized face. In *face classification*, the extracted feature vector of the input face is matched against those of enrolled faces in the database, outputting the identity of the face when a match is found with a sufficient confidence or as an unknown face otherwise.

1.2.2 Research Motivation

From the number of paper in face recognition, I can see that holistic methods are dominant in this field compared with component/feature based methods. Although human

beings can easily recognize face images, it is unclear what features or image structures are used in the human intelligence for this recognition task. Therefore, holistic approaches do not explicitly utilize the image structure information in feature extraction. They take all the pixels of a face image as initial features and extract a set of reliable and discriminative features based on machine learning from an available database. It is true that if I collect enough large number of face images as training samples which cover all the variations, I can get very high recognition accuracy. However, firstly, it is infeasible to collect all the images for the same identity under a variety of conditions; secondly, it is awkward to do training in such a way, as we human beings do not need so many different images of the same person to distinguish face identities. In practice, only a small number of face images are for training compared with the number of the real-world testing face images with unpredictable distributions, which might differ dramatically from that of the training samples. This is the generalizability bottleneck for most holistic approaches. Moreover, in general, holistic approaches require a preprocessing procedure to normalize the face image variations in pose, scale and illumination. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. The performance of the alignment process heavily influences the final recognition result. Hence most approaches work on the normalized face images based on the manually identified landmarks. The recognition performance deteriorates considerably if the manual process is replaced by an automatic landmark detection algorithm.

In contrast to holistic methods, local component/feature based approaches are more robust to variations in pose, scale, expression and occlusion. In most holistic approaches, the 2D image matrices must be previously transformed into 1D image vectors. Such a matrix-to-vector transform causes the loss of some structure information hiding in original 2D images. Thus, it is more reasonable to extract features directly from 2D face images instead of 1D vectors. Moreover, training stage is not necessary to extract

the local features, which avoids the generalizability problem in most holistic approaches. Hence, I focus on face recognition using local features. A local feature is an image pattern which differs from its immediate neighborhood. It is usually associated with a change of an image property or several properties simultaneously, although it is not necessarily localized exactly on this change. Local features can be points, but also edges or small image patches. Typically, some measurements are taken from a region centered on a local feature and converted into descriptors. The descriptors can be used for various applications.

Figure 1.2: Facial Components Modeled at Different Scales

The question is what types of local features I should use. One of the very fundamental problems that arises when analyzing face images originates from the fact that faces appear in different ways depending upon the scale of observation. Intuitively, facial components are modeled at different levels of scales, ranging from skin textures and facial wrinkles at fine scale, through eyes and mouth represented at median scale, to the whole face region at larger scale, as shown in Figure 2.11. Notably, the types of facial component description are strongly dependent on the scale at which the face is modeled. In certain controlled situations, appropriate scales for analysis may be known a priori. Under

other circumstance, it isn't obvious at all how to determine in advance what are the proper scales. One such example is a vision system with the task of analyzing unknown face images. Besides the inherent multi-scale properties of those face images, such a system has to face the problems that the perspective mapping gives rise to size variations, that noise is introduced in the image formation process, and that the available data are two-dimensional data sets reflecting only indirect properties of a three-dimensional world. To be able to cope with these problems, an image representation that explicitly incorporates the notion of scale is a crucially important tool whenever interpreting images automatically.

In computer vision and image processing, these insights have led to the construction of multi-scale representations of image data, obtained by embedding any given signal into a one-parameter family of derived signals. This family should be parameterized by a scale parameter and be generated in such a way that fine-scale structures are successively suppressed when the scale parameter is increased. A main intention behind this construction is to obtain a separation of the image structures in the original image, such that fine scale image structures only exist at the finest scales in the multi-scale representation. Thereby, the task of operating on the image data is simplified, provided that the operations are performed at sufficiently coarse scales where unnecessary and irrelevant fine-scale structures have been suppressed.

However, the representation of multi-scale by itself, contains no explicit information about what image structures should be regarded as important or what scales are proper for treating those. Hence, unless early judgments can be made about what image structures should be regarded as significant, a substantial expansion of the amount of data is obtained to be interpreted by later steps. Thus, it is essential to complement multi-scale process by explicit mechanisms for automatic scale selection. Hence, I focus on face recognition using multi-scale local features, where the scale of each feature is determined by its intrinsic structure automatically.

1.3 Major Results

The first part of my research is face recognition based on the multi-scale local feature extraction and matching framework (LFEM). While some basic tools in this framework are inherited from the Scale Invariant Feature Transform (SIFT) algorithm, this work investigates and contributes to all major steps in the feature extraction and image matching. The second part of my research is a fully automatic face recognition framework (FAFF) based on both the local and global features. In this framework, I propose to align the input face images using multi-scale local features for the holistic approach, which serves as a filter to narrow down the database for further fine matching. The computationally heavy local feature based approach is then applied on the narrowed database. This fully automatic framework not only speeds up the local feature based approach, but also improves the recognition accuracy comparing with the holistic approach, the local feature based approach and their parallel combination.

My efforts on the local feature extraction, description and matching for face images have resulted in the following contributions:

- (i) Design of a multi-scale local feature extraction framework.
- (ii) Design of a two-stage local feature matching strategy.

Three applications: face recognition, template selection and synthesis for multiple templates per subject, and face alignment are proposed based on my multi-scale local feature extraction and matching framework (LFEM). Finally I obtain a fully automatic face recognition framework composed of face alignment and identification.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows. In Chapter 2 I review the backgrounds of the face recognition algorithms, including holistic methods and feature

/ component based methods, where I introduce the deficiencies and limitations of the current face recognition algorithms. In Chapter 3, I present my research works on the multi-scale local feature extraction and description. New approaches to keypoint detection, partial descriptor and insignificant keypoint removal are proposed specifically for human face images, a type of non-rigid and smooth visual objects. In Chapter 4 I present my work for the multi-scale local feature matching algorithm. A strategy of keypoint search for the nearest subject and a two-stage image matching scheme are developed for the face identification task. They circumvent the problem that local structures matched with those in probe disperse into many different gallery images. And in this chapter I introduce one application: Template selection and synthesis, based on the feature extraction and matching strategies described in Chapters 3 and 4. Chapter 5 presents my fully automatic face recognition framework based on both the local and global features. Firstly, I describe my face alignment approach based on multi-scale local features. Given an unaligned face image resulting from a face detector and a set of aligned face images in the data set, I build an automatic transformation mechanism, under which the unaligned face image can be precisely aligned for the subsequent holistic recognition process. Then I integrate the local feature based approach LFEM introduced in Chapters 3 and 4 and the holistic approach ERE in a cascaded way. I use ERE as a filter to retrieve some candidate images which form a gallery set with reduced size, where I perform LFEM for face recognition. This framework not only speeds up the LFEM approach, but also achieves better recognition performance than LFEM, ERE and their parallel combination. Finally in Chapter 6 I conclude my works and predict some future directions of my research.

Chapter 2

Literature Survey

In this chapter I review the backgrounds of face recognition technologies and provide an in-depth introduction for 2D face recognition. As mentioned in Section 1.1, the algorithms of face recognition technologies are classified into two categorizations: component/feature based approaches and holistic matching approaches. Although the earliest works in computer recognition of faces are mainly component based approaches [12, 13], I introduce holistic approaches first for a better organization. After that, I present the early component based approaches and the state-of-art local feature based approaches. By the analysis of these technologies, I find that local feature based approaches are an important but immature area for 2D face recognition. Improvements are necessary for the local feature extraction and matching.

2.1 Holistic Methods

The image of face often belongs to a manifold of intrinsically low dimension, which is represented as high-dimensional pixel arrays. If the whole image pixels are used as input, it is very computationally expensive. Because much of the surface of a face is smooth and has regular texture, per-pixel sampling is in fact unnecessarily dense: The value of a pixel is typically highly correlated with the values of the surrounding pixels. Moreover, the appearance of faces is highly constrained, for example, any frontal view of a face

is roughly symmetrical, has eyes on the sides, nose in the middle. A vast proportion of the points in the image space does not represent physically possible faces. Thus in holistic matching methods, I firstly review those techniques which extract low dimensional features from high dimensional images, namely subspace approaches [1]. After that, I introduce some popular classification approaches, which are based on the low dimensional subspace features.

2.1.1 Subspace Feature Extraction Approaches

The performance of face recognition methods using subspace projection is directly related to the characteristics of their basis images. Then the central issue is which low-dimensional subspaces are the most relevant or informative for the recognition task. Below I study some of the popular subspace feature extraction approaches.

2.1.1.1 Eigenfaces Method

Turk and Pentland [16] proposed the technique of “Eigenfaces” which is the first application of Principal Component Analysis (PCA) [17] to face recognition. PCA, based on extracting the desired number of principal components of the multi-dimensional data, is a dimensionality reduction technique. Figure 2.1 illustrates its idea. The axis labeled ϕ_1 corresponds to the direction of maximum variance and is selected as the first principal component. In a two-dimensional case, the second principal component is then determined uniquely by the orthogonality constraints; in a higher-dimensional space the selection process would continue, guided by the variances of the projections. PCA is closely related to the Karhunen-Loève Transform (KLT) [18], which was derived in the signal processing context as the orthogonal transform with the basis $\Phi = [\phi_1, \dots, \phi_N]^T$ that for any $k \leq N$ minimizes the average L_2 reconstruction error for data points x :

$$\epsilon(x) = \left\| x - \sum_{i=1}^k \phi_i^T x \phi_i \right\|. \quad (\text{Eq. 2.1})$$

Figure 2.1: The concept of PCA/KLT. (a) Solid lines: the original basis; dashed lines: the PCA basis. The dots are selected at regularly spaced locations on a straight line rotated at 30° and then perturbed by isotropic 2D Gaussian noise. (b) The projection (1D reconstruction) of the data using only the first principal component [1].

The formulations of PCA and KLT are identical under the assumption that the data is zero-mean. The basis vectors in KLT can be calculated as below. Let $\mathbf{X}$ be the $N \times M$ data matrix whose columns $x_1, \dots, x_M$ are observations of a signal embedded in $\mathbb{R}^N$; in the context of face recognition, M is the number of available face images and $N = mn$ is the number of pixels in an image. The KLT basis Φ is obtained by solving the eigenvalue problem $\Lambda = \Phi^T \Sigma \Phi$, where Σ is the covariance matrix of the data $\Sigma = \frac{1}{M} \sum_{i=1}^M x_i x_i^T$, $\Phi = [\phi_1, \dots, \phi_m]^T$ is the eigenvector matrix of Σ , and Λ is the diagonal matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_N$ of Σ on its main diagonal, so that ϕ_j is the eigenvector corresponding to the j -th largest eigenvalue. Then it can be shown that the eigenvalue λ_i is the variance of the data projected on ϕ_i .

Thus to perform PCA and extract k principal components of the data, one must project the data onto Φ_k - the first k columns of the KLT basis Φ , which correspond to the k highest eigenvalues of Σ . Another important property of PCA is that it de-correlated

the data: the covariance matrix of $\Phi_k^T X$ is always diagonal. The main properties of PCA are summarized as below:

$$x \approx \Phi_k y, \Phi_k^T \Phi_k = I, E[y_i y_j]_{i \neq j} = 0 \quad (\text{Eq. 2.2})$$

namely, approximate reconstruction, orthonormality of the basis Φ_k and de-correlated principal components $y_i = \phi_i^T x$, respectively. Figure 2.1 illustrates these properties, where PCA is successful in finding the principal manifold.

In the ‘‘Eigenfaces’’ technique [16], the dimension of the basis vectors constructed by PCA was the same as that of the input face images. Hence, these basis vectors were named ‘‘Eigenfaces’’. Each face image was projected into the principal subspace; the coefficients of the PCA expansion were averaged for every subject, resulting in a single k -dimensional representation of that identity. When a test image was projected into the subspace, Euclidean distances between its coefficient vector and those representing each subject were computed. The simplest method for determining which face class provides the best description of an input face image is to find the face class that minimizes the Euclidean distance.

2.1.1.2 Linear Discriminants: Fisherfaces

The PCA approach provides us with features that capture the main directions along which face images differ the most. However, the variations between the images of the same face due to expression, illumination and viewing direction are almost always larger than image variations due to change in face identity. In other words, since PCA doesn’t use class membership information, examples of the same class or different classes are treated the same way. Belhumeur *et al.* [19] proposed to solve this problem with Fisher’s Linear Discriminant (FLD) method, which uses the class membership information in finding eigenfaces where the variations among different faces are emphasized while the variations of the same face are de-emphasized.

Given a set of properly normalized w-by-h face images, a training set of column image vectors X_{ij} can be formed, where $X_{ij} \in \mathbb{R}^{n=wh}$. Let it contain p persons and q_i sample images for person i . The number of total training sample is $l = \sum_{i=1}^p q_i$. For face recognition, each person is a class with prior probability of c_i . The within-class scatter matrix is defined by

$$S_w = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T \quad (\text{Eq. 2.3})$$

where $\bar{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. The between-class scatter matrix S^b and the total scatter matrix S^t are defined by

$$S^b = \sum_{i=1}^p c_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T \quad (\text{Eq. 2.4})$$

$$S^t = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})^T \quad (\text{Eq. 2.5})$$

where $\bar{X} = \sum_{i=1}^p c_i \bar{X}_i$. $c_i = \frac{1}{p}$, if all classes have equal prior probability.

Let S^g , $g \in t, w, b$ represent one of the above scatter matrices. If the elements of the image vector and the class mean vector are regarded as features, these preliminary features will be de-correlated by solving the eigenvalue problem $\Lambda^g = \Phi^{gT} S^g \Phi^g$, where $\Phi^g = [\phi_1^g, \dots, \phi_n^g]$ is the eigenvector matrix of S^g , and Λ^g is the diagonal matrix of eigenvalues $\lambda_1^g, \dots, \lambda_n^g$ corresponding to the eigenvectors. FLD selects the linear subspace Φ which maximizes the ratio

$$\frac{\Phi^T S^b \Phi}{\Phi^T S^w \Phi}. \quad (\text{Eq. 2.6})$$

Intuitively, FLD finds the projection of the data in which the classes are most linearly separable.

It can be shown that the dimension of Φ is at most $p - 1$. The Fisherfaces algorithm first reduces the dimensionality of the data with PCA, and then applies FLD to further reduce the dimensionality to $p - 1$, since in practice S^w is usually singular.

2.1.1.3 Locality Preserving Projections: Laplacianfaces

PCA is an eigenvector method designed to model linear variation in high-dimensional data. PCA performs dimensionality reduction by projecting the original n -dimensional data onto the $k (< n)$ -dimensional linear subspace spanned by the leading eigenvectors of the data's covariance matrix. Its goal is to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data and for which the coefficients are pairwise de-correlated. For linearly embedded manifolds, PCA discovers the dimensionality of the manifold and produces a compact representation. In contrast, LDA is a supervised learning algorithm. LDA searches for the project axes on which the data points of the same class are close to each other while requiring data points of different classes are far from each other. Unlike PCA which performs in an orthogonal linear space, LDA encodes information in a linearly separable space using bases that are not necessarily orthogonal. It is generally believed that algorithms based on LDA are superior to those based on PCA, but when the training database is small, PCA outperforms LDA, and also that PCA is less sensitive to different training data sets than LDA. However, the structure of the manifold on which the face images possibly reside is not explicitly considered by neither of them. In [29], He *et al.* proposed a new approach to face representation and recognition, which explicitly considers the manifold structure. Specifically, the manifold structure is modeled by a nearest-neighbor graph which preserves the local structure of the image space. A face subspace is obtained by Locality Preserving Projections (LPP) [30]. Each face image in the image space is mapped to a low-dimensional face subspace, which is characterized by a set of feature images (Laplacianfaces). The face subspace preserves local structure and has more discriminating power than the PCA approach for classification purpose.

Considering the problem of representing all of the vectors in a set of n d -dimensional samples $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, with zero mean, by a single vector $\mathbf{y} = [y_1, y_2, \dots, y_n]$ such that y_i

represents $\mathbf{x}_i$. To be specific, a linear mapping is constructed from the $d - dimensional$ space to a line. $\mathbf{w}$ is the transformation vector. That is, $\mathbf{w}^T \mathbf{x}_i = y_i$. In face recognition, a face image is denoted by the vector $\mathbf{x}_i$. The first step of LPP algorithm is to construct the adjacency graph $\mathbf{G}$ of n nodes, such that node i and j are linked if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close with respect to each other in any of the belowing two conditions:

- (i) k-nearest neighbors: Nodes i and j are linked by an edge, if i is among k-nearest neighbors of j or vice-versa.
- (ii) ϵ -neighbors: Nodes i and j are linked by an edge if $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \epsilon$, where $\|\cdot\|$ is the usual Euclidean norm.

Next stage is to construct the weight matrix $\mathbf{S}$, which is a sparse symmetric $n \times n$ matrix with weights s_{ij} if there is an edge between nodes i and j , and 0 if there is no edge. Two alternative criterions to construct the weight matrix are as follows:

- (i) Heat-Kernel: $s_{ij} = \exp(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t})$, if i and j are linked.
- (ii) $s_{ij} = 1$, if nodes i and j are linked by an edge.

The objective function of LPP is:

$$\min \sum_{ij} (y_i - y_j)^2 s_{ij}, \quad (\text{Eq. 2.7})$$

where y_i is the one-dimensional representation of x_i . To solve the objective function, one need to solve the belowing generalized eigenvector problem:

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{w} = \lambda \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{w} \quad (\text{Eq. 2.8})$$

Where $\mathbf{D}$ is the diagonal matrix with entries as $D_{ii} = \sum_j s_{ji}$ and $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the laplacian matrix. The transformation matrix $\mathbf{w}$ is formed by arranging the eigenvectors of Equation Eq. 2.8 ordered according to their eigenvalues, $\lambda_1 < \lambda_2, \dots, < \lambda_l$.

In [31], Noushath *et al.* proposed a method called mixture-of-Laplacianfaces (or LPP mixture model), which obtains several sets of Laplacianfaces through Expectation-Maximization (EM) learning of Gaussian Mixture Models (GMM). As LPP based approaches construct the face subspace without using the available face identity information, Wang and Li [32] proposed a semi-supervised locality preserving learning method for face representation and recognition, called Semi-supervised Laplacianface (S-Laplacianface). It is essentially based on the Laplacianface method. It builds an adjacency graph which best reflects the class relationship and the geometry of the face manifold between the sample points. The projections composed of a set of basis functions are obtained by preserving such a graph structure. It tries to preserve the local geometric structure as Laplacianfaces, and in the meanwhile requires the basis functions to satisfy the pairwise constraints refined by the user. In [33], Cai *et al.* proposed an appearance based face recognition method, called orthogonal Laplacianface. Their algorithm is based on the locality preserving projection (LPP) algorithm. LPP is nonorthogonal, and this makes it difficult to reconstruct the data. The orthogonal locality preserving projection (OLPP) method produces orthogonal basis functions and has more locality preserving power than LPP. Since the locality preserving power is potentially related to the discriminating power, the OLPP has more discriminating power than LPP. Experimental results on Yale, ORL and PIE databases demonstrate the effectiveness of their proposed algorithm. The Laplacianfaces based approaches have been observed to significantly outperform the popular Eigenfaces and Fisherfaces methods. However, all of these methods are inefficient in that when handling the associated eigen equations, the computation and the memory complexities of the problem go up exponentially with the dimensionality of the training image vectors. In [34], Niu *et al.* developed the 2D Laplacianfaces method to reduce the complexity of Laplacianfaces. In extensive experiments on various databases, the one-dimensional Laplacianfaces algorithm is outperformed by this 2D Laplacianfaces.

2.1.1.4 Marginal Fisher Analysis

[35] summarized PCA, LDA and LPP under a common framework graphically. And a new dimensionality reduction algorithm, Marginal Fisher Analysis (MFA) is developed to overcome the limitations of LDA. In MFA, the intrinsic graph is formulated to characterize intra-class compactness, and the penalty graph is designed for inter-class separability. If one vertex is among the k_1 -nearest neighbors of the other and the elements of the pair belong to the same class in the intrinsic graph, a vertex pair is connected. For each class, in the penalty graph, the k_2 -nearest vertex pairs in which one element is in-class and the other is out-of-class are connected. Compared to LDA, MFA has the following advantages:

- (i) The number of available projection directions is much larger than that of LDA;
- (ii) There is no assumption on the data distribution;
- (iii) The interclass margin better characterizes the separability of different classes than the interclass scatter in LDA.

$G = \{X, W\}$ is an undirected weighted graph with vertex set X and similarity matrix $W \in \mathbb{R}^{N \times N}$. The diagonal matrix D is defined as $D_{ii} = \sum_{j \neq i} W_{ij}$, $\forall i$. $L = D - W$ is the Laplacian matrix of a graph G . The objective function is as below:

$$y^* = \arg \min_{y^T B y = d} \sum_{i \neq j} \|y_i - y_j\|^2 W_{ij} = \arg \min_{y^T B y = d} y^T L y, \quad (\text{Eq. 2.9})$$

where d is a constant and B is the constraint matrix, y_i is the low-dimensional representation of vertex x_i .

Assuming the low-dimensional vector representations of the vertices can be obtained from a linear projection as $y = X^T w$, where w is the unitary projection vector. By

following the graph embedding formulation, intra-class compactness is characterized from the intrinsic graph by the term

$$\tilde{S}_c = \sum_i \sum_{i \in N_{k_1}^+(j) \text{ or } j \in N_{k_1}^+(i)} \|w^T x_i - w^T x_j\|^2 = 2w^T X(D - W)X^T w, \quad (\text{Eq. 2.10})$$

where $W_{ij} = \begin{cases} 1, & \text{if } i \in N_{k_1}^+(j) \text{ or } j \in N_{k_1}^+(i); \\ 0, & \text{else.} \end{cases}$. $N_{k_1}^+(i)$ indicated the index set of the k_1 nearest neighbors of the sample x_i in the same class. Inter-class separability is characterized by a penalty graph with the term

$$\tilde{S}_p = \sum_i \sum_{(i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j)} \|w^T x_i - w^T x_j\|^2 = 2w^T X(D^p - W^p)X^T w, \quad (\text{Eq. 2.11})$$

where $W_{ij} = \begin{cases} 1, & \text{if } (i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j); \\ 0, & \text{else.} \end{cases}$. $P_{k_2}(c)$ is a set of data pairs that are the k_2 nearest pairs among the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$.

The algorithmic procedure of MFA is stated as below:

- (i) PCA projection.
- (ii) Constructing the intra-class compactness and inter-class separability graphs.
- (iii) Marginal Fisher Criterion. From the linearization of the graph embedding framework, the Marginal Fisher Criterion is

$$w^* = \arg \min_w \frac{w^T X(D - W)X^T w}{w^T X(D^p - W^p)X^T w} \quad (\text{Eq. 2.12})$$

- (iv) Output the final linear projection direction as

$$w = W_{PCA} w^*, \quad (\text{Eq. 2.13})$$

where W_{PCA} is the transformation matrix of PCA.

The MFA can be extended to Kernel Marginal Fisher Analysis (KMFA) and Tensor Marginal Fisher Analysis (TMFA).

2.1.1.5 Eigenfeature Regularization and Extraction

A new approach was presented in [36] called eigenfeature regularization and extraction (ERE). Experimental results in [36, 37] show that ERE consistently outperforms other holistic approaches.

Given a set of properly normalized w -by- h face images, a training set of column image vectors X_{ij} is formed, where $X_{ij} \in \mathbb{R}^{n=wh}$. Let the training set contain p persons and q_i sample images for person i . The number of total training sample is $l = \sum_{i=1}^p q_i$. For face recognition, each person is a class with prior probability of c_i . The within-class scatter matrix is defined by $\mathbf{S}^w = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$, where $\bar{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. By solving the eigenvalue problem, the equation as follow is obtained:

$$\mathbf{\Lambda} = \mathbf{\Phi}^{wT} \mathbf{S}^w \mathbf{\Phi}^w, \quad (\text{Eq. 2.14})$$

where $\mathbf{\Phi}^w = [\phi_1^w, \dots, \phi_n^w]$ is the eigenvector matrix of $\mathbf{S}^w$, and $\mathbf{\Lambda}^w$ is the diagonal matrix of eigenvalues $\lambda_1^w, \dots, \lambda_n^w$ corresponding to the eigenvectors.

The eigenspace $\mathbb{R}^n$ spanned by eigenvectors $\{\phi_k^w\}_{k=1}^n$ is decomposed into three subspaces: a reliable face space $\mathbf{F} = \{\phi_k^w\}_{k=1}^m$, an unstable noise space $\mathbf{N} = \{\phi_k^w\}_{k=m+1}^r$ and a null space $\mathbf{\emptyset} = \{\phi_k^w\}_{k=r+1}^n$, where r is the rank of $\mathbf{S}^w$. To determine the start point of the noise dominant region $m+1$, firstly the median of nonzero eigenvalues $\lambda_{med}^w = \text{median}\{\forall \lambda_k^w | k \leq r\}$ is found. The distance between λ_{med}^w and the smallest nonzero eigenvalue is $d_{m,r} = \lambda_{med}^w - \lambda_r^w$. The start point of the noise region $m+1$ is estimated by

$$\lambda_{m+1}^w = \max\{\forall \lambda_k^w | \lambda_k^w < (\lambda_{med}^w + \mu(\lambda_{med}^w - \lambda_r^w))\}, \quad (\text{Eq. 2.15})$$

where μ is a constant. The optimal value of μ is slightly larger or smaller than 1 for different applications.

The eigenspectrum is regularized by replacing the noise dominating λ_k^w in $\mathbf{N}$ with a model $\alpha/(k + \beta)$ and replacing the zero λ_k^w in the null space $\emptyset$ with a constant. The regularized eigenspectrum $\tilde{\lambda}_k^w$ is given by

$$\tilde{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m; \\ \frac{\alpha}{k+\beta}, & m \leq k \leq r; \\ \frac{\alpha}{r+1+\beta}, & r < k \leq n, \end{cases} \quad (\text{Eq. 2.16})$$

where $\alpha = \frac{\lambda_1^w \lambda_m^w (m-1)}{\lambda_1^w - \lambda_m^w}$, and $\beta = \frac{m \lambda_m^w - \lambda_1^w}{\lambda_1^w - \lambda_m^w}$. The training data are transformed to $\tilde{Y}_{ij} = \tilde{\Phi}_n^{wT} X_{ij}$, where

$$\tilde{\Phi}_n^w = [\phi_1^w / \sqrt{\tilde{\lambda}_1^w}, \dots, \phi_n^w / \sqrt{\tilde{\lambda}_n^w}]. \quad (\text{Eq. 2.17})$$

After the feature regularization, a new total scatter matrix is formed by vectors $\tilde{Y}_{ij}$ of the training data as $\tilde{\mathbf{S}}^t = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (\tilde{Y}_{ij} - \bar{Y})(\tilde{Y}_{ij} - \bar{Y})^T$, where $\bar{Y} = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} \tilde{Y}_{ij}$.

The regularized features $\tilde{Y}_{ij}$ will be decorrelated for $\tilde{\mathbf{S}}^t$ by solving the eigenvalue problem similar to Eq. 5.6. The eigenvectors in the eigenvector matrix $\tilde{\Phi}_n^t = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_n^t]$ are sorted in a descending order of the corresponding eigenvalues. The dimensionality reduction is performed by keeping the eigenvectors with the d largest eigenvalues $\tilde{\Phi}_d^t = [\tilde{\phi}_k^t]_{k=1}^d = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_d^t]$, where d is the number of features usually selected by a specific application. The feature regularization and extraction matrix $\mathbf{U}$ is given by $\mathbf{U} = \tilde{\Phi}_n^w \tilde{\Phi}_d^t$, which transforms a face image vector X , $X \in \mathbb{R}^n$, into a feature vector F , $F \in \mathbb{R}^d$, by $F = \mathbf{U}^T X$.

2.1.2 Holistic Feature Matching Approaches

In the previous section, I have reviewed several popular subspace feature extraction approaches. Now, I will survey some matching approaches for the face recognition application, which use the features extracted by the subspace approaches.

2.1.2.1 Bayesian Methods

The standard eigenface approach was extended to a Bayesian approach [38, 39] by using a probabilistic measures of similarity, instead of the nearest-neighbor eigenspace matching methods. This method differs from traditional algorithms in two important aspects: firstly, it uses a Bayesian classifier instead of the nearest-neighbor classifier; secondly, it casts the multi-class problem of distinguishing among images of different subjects into the problem of distinguishing between intra-personal and extra-personal difference images.

The classifier proposed by Moghaddam *et al.* defines the subspace spanned by the difference between two face images. The difference image is the signed arithmetic difference between respective pixels in the source images. Using such difference images two distinct and mutually exclusive classes can be defined. Φ^I representing intra-personal difference between multiple images of the same individual, and Φ^E representing extra-personal variations in matching two different individuals. It is assumed that both classes are Gaussian distributed and seek to obtain estimates of the likelihood functions $P(\Delta|\Phi^I)$ and $P(\Delta|\Phi^E)$ for a given intensity difference $\Delta = X_{ij} - X_{st}$. The similarity is then expressed in terms of probability:

$$S(X_{ij} - X_{st}) = P(\Delta \in \Phi^I) = P(\Phi^I|\Delta) \quad (\text{Eq. 2.18})$$

where $P(\Phi^I|\Delta)$ is the a posteriori probability given by Bayes rule, using estimates of the likelihoods $P(\Delta|\Phi^I)$ and $P(\Delta|\Phi^E)$.

Given these likelihoods, a similarity score $S(X_{ij}, X_{st})$ is evaluated between a pair of images directly in terms of intra-personal a posteriori probability as given Bayes rule:

$$S(X_{ij}, X_{st}) = P(\Phi^I|\Delta) = \frac{P(\Delta|\Phi^I) \cdot P(\Phi^I)}{P(\Delta|\Phi^I) \cdot P(\Phi^I) + P(\Delta|\Phi^E) \cdot P(\Phi^E)} \quad (\text{Eq. 2.19})$$

where the priors $P(\Phi^I)$ is set to reflect specific operating conditions or other sources of a priori knowledge regarding the two images being matched [38, 40, 39, 41].

A standard face recognition task, essentially an p -ary classification problem for p individuals Bayesian formulation in Equation Eq. 2.19 is casted into a binary pattern classification problem with Φ^I and Φ^E . This problem is then solved using the maximum a posterior (MAP) rule - two images are determined to belong to the same individual if $P(\Phi^I|\Delta) > P(\Phi^E|\Delta)$, or in equivalent, if $S(X_{ij}, X_{st}) > 1/2$ [39]. An alternative probability similarity measure is defined in simpler form using the intra-personal likelihood alone [39],

$$S'(X_{ij}, X_{st}) = P(\Delta|\Phi^I) \quad (\text{Eq. 2.20})$$

thus leading to the maximum likelihood (ML) recognition as opposed to MAP recognition in Equation Eq. 2.19. Experimental results in [39] show that this simplified ML measure is almost as effective as its MAP counterpart in most cases.

To estimate likelihoods $P(\Delta|\Phi^I)$ and $P(\Delta|\Phi^E)$, in [38, 39, 41], authors proposed an efficient density estimation method, which divides the vector space $\mathbb{R}^n$ into a principal subspace $\mathbb{F}$ and a residual subspace $\bar{\mathbb{F}}$ based on PCA to form a low-dimensional estimate of the complete likelihood which can be evaluated using only the first m components, where $m \ll n$. The main contribution of the approach [39] is the replacement of the unreliable eigenvalues in $\bar{\mathbb{F}}$ by a constant. This solves the singularity problem of the covariance matrix, but it does not well solve the over-fitting problem. The constant used to replace the eigenvalues in $\bar{\mathbb{F}}$ is estimated as the average eigenvalue over $\bar{\mathbb{F}}$. The high-dimensional face image and the limited number of training samples result in over-fitting problem because eigenvalues in $\bar{\mathbb{F}}$ are unreliable and so is their arithmetic average. In [42], Jiang *et al.* proposed an approach that selects a more robust constant, which is done by replacing the average eigenvalue with the maximum of the eigenvalues available in the subspace $\bar{\mathbb{F}}$. Experimental results in [42] show that their approach consistently outperforms the conventional ML method [39].

2.1.2.2 Support Vector Machines

After Phillips [43] introduces support vector machines (SVM) for face recognition, various methods of SVM have been proposed for face identification or verification [44, 45, 46]. In this theory, a SVM finds the hyperplane that separates the largest possible fraction of points of the same class on the same side, given a set of points belonging to two classes, while maximizing the distance from either class to the hyperplane. This hyperplane is called as optimal separating hyperplane (OSH) by Vapnik [47] which minimizes the risk of misclassifying not only the examples in the training set but also the unseen examples of the test set. Firstly, features are extracted using PCA and then the discrimination functions between each pair are learned by SVMs. Subsequently, the disjoint test set enters the system for recognition. [45] compares SVM with the standard eigenface approach using nearest center classification (NCC) criterion for face recognition and the former was found to perform better. However, SVMs can get over-trained and result in the loss of generality. Moreover, they employ various kernels with many parameters. How to choose these parameters and kernel functions are open questions in the literature. This makes the optimization space very extensive, without the guarantee that it has been fully explored to search the best solution.

2.1.2.3 Sparse Representation

In the computer vision literature, a vast number of feature extraction methods have been investigated to find projections that better separate the classes in lower dimensional spaces, or feature spaces, as surveyed in Section 2.1.1. Traditionally, when feature extraction is used following simple classifiers such as nearest neighbor (NN) and nearest subspace (NS), feature transformation is critical to the whole system. This has led to the development of a wide variety of increasingly complex feature extraction methods. With so many proposed features and so little consensus about which are better or worse,

researchers lack guidelines to decide which features to use. Wright *et al.* [48] proposed a framework for face recognition, within which the theory of compressed sensing implies that *the precise choice of feature space is no longer critical*: Even random features contain enough information to correctly classify any test image. What is critical is that the dimension of the feature space is sufficiently large and that the sparse representation is correctly computed.

A basic problem in object recognition is to use labeled training samples from k distinct object classes to correctly determine the class to which a new test sample belongs. The given n_i training samples from the i th class are columns of a matrix $A_i \doteq [v_{i,1}, v_{i,2}, \dots, v_{i,n_i}] \in \mathbb{R}^{m \times n_i}$. In the field of face recognition, a $w \times h$ gray-scale image are defined as the vector $v \in \mathbb{R}^m$ ($m = wh$) given by stacking its columns; the columns of A_i are then the training face images of the i th subject.

Given sufficient training images of the i th object class, $A_i = [v_{i,1}, v_{i,2}, \dots, v_{i,n_i}] \in \mathbb{R}^{m \times n_i}$, any test sample $y \in \mathbb{R}^m$ from the same class will approximately lie in the linear span of the training samples associated with object i :

$$y = \alpha_{i,1}v_{i,1} + \alpha_{i,2}v_{i,2} + \dots + \alpha_{i,n_i}v_{i,n_i}, \quad (\text{Eq. 2.21})$$

for some scalars, $\alpha_{i,j} \in \mathbb{R}$, $j = 1, 2, \dots, n_i$.

Since the membership i of the test sample is unknown, a new matrix A is defined for the entire training set as the concatenation of the n training samples of all k object classes:

$$A \doteq [A_1, A_2, \dots, A_k] = [v_{1,1}, v_{1,2}, \dots, v_{k,n_k}]. \quad (\text{Eq. 2.22})$$

Then, the linear representation of y can be rewritten in terms of all training samples as

$$y = Ax_0 \quad \in \mathbb{R}^m, \quad (\text{Eq. 2.23})$$

where $x_0 = [0, \dots, 0, \alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,n_i}, 0, \dots, 0]^T \in \mathbb{R}^n$ is a coefficient vector whose entries are zero except those associated with the i th class.

Conventionally, this difficulty is resolved by choosing the minimum ℓ^2 -norm solution:

$$(\ell^2) : \quad \hat{x}_2 = \arg \min \|x\|_2 \quad \text{subject to } Ax = y. \quad (\text{Eq. 2.24})$$

To seek the sparsest solution to $y = Ax$, one should solve the following optimization problem:

$$(\ell^0) : \quad \hat{x}_0 = \arg \min \|x\|_0 \quad \text{subject to } Ax = y, \quad (\text{Eq. 2.25})$$

where $\|\cdot\|_0$ denotes the ℓ^0 -norm, which counts the number of nonzero entries in a vector. However, the problem of finding the sparsest solution of an under-determined systems of linear equation is NP-hard and difficult even to approximate. Recent development in the emerging theory of sparse representation and compressed sensing reveals that if the solution x_0 sought is sparse enough, the solution of the ℓ^0 -minimization problem is equal to the solution to the following ℓ^1 -minimization problem:

$$(\ell^1) : \quad \hat{x}_1 = \arg \min \|x\|_1 \quad \text{subject to } Ax = y. \quad (\text{Eq. 2.26})$$

This problem can be solved in polynomial time by standard linear programming methods.

The authors in [48] proposed two models based on Eq. 2.23 to deal with small dense noise and occlusion or corruption. To account for small possibly dense noise, the model Eq. 2.23 can be modified as

$$y = Ax_0 + z, \quad (\text{Eq. 2.27})$$

where $z \in \mathbb{R}^m$ is a noise term with bounded energy $\|z\|_2 < \varepsilon$. The sparse solution x_0 can still be approximately recovered by solving the following ℓ^1 -minimization problem

$$(\ell_s^1) : \quad \hat{x}_1 = \arg \min \|x\|_1 \quad \text{subject to } \|Ax - y\|_2 \leq \varepsilon. \quad (\text{Eq. 2.28})$$

In many practical face recognition scenarios, the test image y could be partially corrupted or occluded. In this case the model Eq. 2.23 should be modified as

$$y = y_0 + e_0 = Ax_0 + e_0, \quad (\text{Eq. 2.29})$$

where $e_0 \in \mathbb{R}^m$ is a vector of errors - a fraction, ρ , of its entries are nonzero. The nonzero entries of e_0 model which pixels in y are corrupted or occluded. Since $y_0 = Ax_0$, Eq. 2.29 can be rewritten as

$$y = [A, I] \begin{bmatrix} x_0 \\ e_0 \end{bmatrix} \doteq Bw_0 \quad (\text{Eq. 2.30})$$

Here, $B = [A, I] \in \mathbb{R}^{m \times (n+m)}$. The sparsest solution w_0 is recovered from solving the following extended ℓ^1 -minimization problem:

$$(\ell_e^1) : \quad \hat{w}_1 = \arg \min \|w\|_1 \quad \text{subject to } Bw = y. \quad (\text{Eq. 2.31})$$

That is, in Eq. 2.26, the image matrix A is now replaced with the extended matrix $B = [A, I]$ and x with $w = [x, e]$.

Once the sparse solution $\hat{w}_1 = [\hat{x}_1, \hat{e}_1]$ is computed, setting $y_r \doteq y - \hat{e}_1$ recovers a clean image of the subject with occlusion or corruption compensated for. To identify the subject, the residual $r_i(y)$ in is slightly modified, it is computed against the recovered image y_r :

$$r_i(y) = \|y_r - A\delta_i(\hat{x}_1)\|_2 = \|y - \hat{e}_1 - A\delta_i(\hat{x}_1)\|_2. \quad (\text{Eq. 2.32})$$

Experimental results in [48] showed that exploiting sparsity is critical for the high-performance classification of high-dimensional data such as face images. With sparsity properly harnessed, the choice of features becomes less important than the number of features. Moreover, occlusion and corruption can be handled uniformly and robustly within the same classification framework.

In general, the present state-of-art holistic based face recognition approaches use subspace based methodologies. Several advances and challenging problems of subspace approaches were presented in [49]. It is known that global techniques work well for classifying frontal views of faces. However, they are not robust against pose changes, illumination variations, occlusions or expression variations. To solve this problem, many component/feature based methods which extract local facial features were proposed. I

will review some popular face recognition methods based on the local features in the following section.

2.2 Feature/Component based Approaches

In contrast to holistic methods, local feature based methods extract facial features from different levels of locality. The general idea of face recognition methods using local features is to first locate several facial features (components), and then classify the faces by comparing and combining the corresponding local statistics. Due to increasing interest, stand-alone sections were specifically devoted to local feature based methods in recent surveys [11, 50]. And a comparative study of local feature based approach for face recognition was reported in the literature [51]. Following I will review some popular local feature based methods for face recognition.

2.2.1 Earlier Approaches

Bledsoe [52] proposed a face recognition system which is one of the earliest works in computer recognition of faces. In this system, the positions of the feature points on the face were located and entered into the computer by a human operator. Given a set of feature point distances of an unknown person, nearest neighbor or other classification rules were used for identifying the label of the test image. This system could handle wide variations in head rotation, tilt, image quality and contrast, because feature extraction is manually done. Another landmark work on face recognition is reported in the doctoral dissertation of M.D. Kelly [12]. The framework of Kelly's algorithm is similar to that of Bledsoe's, but is significantly different in that it does not involve any human intervention. In this approach, the body and close up head images were used for recognition. Once the body and head have been outlined, ten measures are extracted. The body measures include heights, widths of the head, neck, shoulders and hips. The face measures include

width of the head and distance between eyes, top of head to eyes, between eyes and nose and the distance from eyes to mouth. To identify the class label of the test image, the nearest neighbor rule was applied. In [13], another method was proposed with the use of geometrical parameterizations, i.e., distances and angles between points such as eye corners, mouth extremities, nostrils and chin top. Once the eyes, nose and mouth are approximately located in two stages (a coarse-grain stage and a fine stage), more accurate information is extracted from four smaller regions (left and right eye, nose and mouth), scanning at higher resolution, and using the “best beam intensity” which is based on the local area histogram obtained in the coarse-grain stage. A set of 16 facial parameters which are ratios of distances, areas and angles is extracted. To eliminate scale and dimension differences the components of the resulting vector are normalized. A simple distance measure is used to check for similarity between an image of the test set and the image in the reference set.

If a system needs manually entering feature points into the computer, the labor costs should be high. Moreover, when there are changes due to expressions or occlusions, the recognition rate decreases drastically if a system only uses geometrical features (distances and angles between facial components). And till now there is no method which can automatically detect facial components as accurately as human beings, various errors occur during the component detection process. Due to above shortcomings, these early approaches are not robust to automatic face recognition issues.

Currently, Kumar et. al. [53, 54] proposed the use of describable visual attributes for face verification and image search. Describable visual attributes are labels that can be given to an image to describe its appearance. They built 73 face attribute classifiers, which include gender, age, hair color, eye shape, nose size, nose shape, mouth shape and so on. These features are very similar to those used in the early approaches for face recognition. To collect attribute and identity labels, they use the Amazon Mechanical Turk (MTurk) service, by which they collected over 145,000 triply-verified positive

attribute labels, for about \$6,000. Then they learnt 73 attribute classifiers by Support Vector Machines with RBF kernels. For face verification, a verification classifier is further trained. In order to decide whether two face images I_1 and I_2 are from the same identity, they train a verification classifier V that compares attribute vectors $C(I_1)$ and $C(I_2)$ and returns $v(I_1, I_2)$, the verification decision. These vectors are constructed by concatenating the result of n different attribute classifiers. Let $a_i = C_i(I_1)$ and $b_i = C_i(I_2)$ be the outputs of the i th trait classifier for each face ($1 \leq i \leq n$). Let $p_i = \langle |a_i - b_i|, a_i b_i \rangle$, by concatenation of these p_i for all n attribute classifier outputs forms the input to the verification classifier V :

$$v(I_1, I_2) = V(\langle p_1, \dots, p_n \rangle) \quad (\text{Eq. 2.33})$$

For the classification function, they use an SVM with an RBF kernel for V , trained using libsvm [55].

The Attribute Face Service (AFS) can be found online ¹, which lets you submit images for face detection and compute attributes on the detected faces.

2.2.2 Gabor Features

Gabor filters, which are comparable to the receptive fields of simple cells in the mammalian visual cortex [56], are spatially localized and selective to spatial orientations and scales. Gabor filters have been adopted in many face recognition applications [57, 58, 59, 60], due to their biological relevance and computational properties.

The Gabor wavelets [61] (kernels, filters) can be defined as follows:

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-(\|k_{\mu,\nu}\|^2 \|z\|^2 / 2\sigma^2)} [e^{ik_{\mu,\nu}z} - e^{-\sigma^2/2}] \quad (\text{Eq. 2.34})$$

where μ and ν define the orientation and scale of the Gabor kernels, $z = (x, y)$, $\|\cdot\|$ denotes the norm operator, and the wave vector $k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \quad (\text{Eq. 2.35})$$

¹Available at: <http://afs.automaticfacesystems.com/>

where $k_\nu = k_{max}/f^\nu$ and $\phi_\nu = \pi\nu/8$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain.

Figure 2.2: The real part of the Gabor kernels at five scales and eight orientations.

The Gabor kernels in Eq. 2.34 are all self-similar since they can be generated from one filter, the mother wavelet, by scaling and rotation via the wave vector $k_{\mu,\nu}$. In most face recognition applications, one would use Gabor wavelets of five different scales, $\nu \in \{0, \dots, 4\}$, and eight orientations, $\mu \in \{0, \dots, 7\}$. In this way, face images are represented in a multi-scale manner. Fig. 2.2 shows the real part of the Gabor kernels at five scales and eight orientations and their magnitude, with the following parameters: $\sigma = 2\pi$, $k_{max} = \pi/2$, and $f = \sqrt{2}$. The kernels exhibit desirable characteristics of spatial frequency, spatial locality, and orientation selectivity.

Among all the Gabor feature based approaches, Elastic Bunch Graph Matching (EBGM) [58] is one of the most popular. In this algorithm, the authors attached a set

of jets called the bunch graphs derived from different face images to each node. Please refer to [62] for the details. The general idea of EBGM is summarized as below:

- (i) Preprocessing with Gabor wavelets: Face images are convolved with Gabor wavelets. The set of convolution coefficients of different orientations and frequencies at one image pixel is called a jet.
- (ii) Landmark localization: This is the bunch graph creation step. The edges are labeled with distance information and nodes are labeled with wavelet responses locally bundled in jets. The graph structure (a set of nodes plus edges) and the assignment of the corresponding labels (jets and distances) for one initial image is designer-provided, which are known as model graphs and are stored in the database. Stored model graphs can be matched to new images to generate image graphs, which can be incorporated into gallery set.
- (iii) Comparing jets: Measuring the similarity for two jets is fundamental and is required for both landmark localization and face graph similarity measurement. The final phase similarity measure is given in [62].
- (iv) Recognition: After having extracted model graphs from the gallery images and image graphs from the probe images, recognition is performed by comparing an image graph to all model graphs and picking the one with the highest similarity values. The similarity function used for comparing the graphs is an average over the similarities between pairs of corresponding jets. A person is recognized correctly if the correct model yields the highest graph similarity measure.

In general, the matching finds the fiducial points quite accurately. But mispositioning occurs, which causes wrong matching. The success of the EBGM system is due to its resemblance to the human visual system. However, this method is very tedious in its

practical implementations as all the initial face graphs are to be created manually. Although this method is robust to pose and expression changes, its sensitivity with respect to illumination variations is yet to be tested. And if some facial parts are occluded, no fiducial points are detected in the occlusion area, which leads to large recognition error rate. Furthermore, if more number of fiducial points are selected for improving accuracy then the system can be computationally complex and manually very labor intensive.

2.2.3 Shape Based Approaches

The methods that I review in this category is shape based approaches for face recognition. Cootes *et al.* proposed Active Shape Model (ASM) in [63]. They described methods for building models which learn patterns of variability from training set. The proposed flexible model is an iterative refinement algorithm similar to active contour models (snakes) [64]. The key difference is that it can deform and fit according to the training data, which allow considerable variability but is still specific to the class of structures they represent. A new statistical based Active Appearance Model (AAM) is proposed in [65]. Their models of appearance can match to any of a class of deformable objects (e.g., any face with any expression, rather than one person's face with a particular expression). To increase the speed of their matching process, they pre-computed the derivatives of the residual of the match between the model and the target image and use them to compute the update steps in an iterative matching algorithm [65]. However, the performances of both ASM and AAM depend on a good selection of facial landmarks, which are often annotated manually. This makes the approaches semi-automatic and labor consuming.

2.2.4 Local Binary Pattern

The original Local Binary Pattern (LBP) operator, introduced by Ojala *et al.* [66], is a powerful means of texture description. The operator labels the pixels of an image by

thresholding the 3×3 -neighborhood of each pixel based on the value of the center pixel and considering the result as a binary number. Then the histogram of the labels is used as a texture. See Fig. 2.3 for an illustration of the basic LBP operator.

Figure 2.3: The basic LBP operator [2].

During the LBP operation, the value of current pixel, g_c , is applied as a threshold to each of the neighbors g_p , ($p = 0, \dots, P - 1$), to obtain a binary number. Hence, the information extracted by LBP is invariant to local gray-scale variations of the image. Then a local binary pattern is obtained by first concatenating these binary numbers and then converting the sequence into the decimal number. Using circular neighborhoods and linearly interpolating the pixel values allow the choice of any radius R and number of pixels P in the neighborhood, to generate an operator.

$$LBP_{P,R}(x, y) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (\text{Eq. 2.36})$$

where $s(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$

In [2, 67, 68, 69], LBP operators were applied to various face recognition problems. Generally, the face image is first divided into small regions from which the LBP features are extracted and concatenated into a feature histogram representing the face image. Because the face images can be seen as composition of micro-patterns which are invariant to monotonic gray scale transformations, a global description of the face image is obtained

by combining these micro-patterns. In the matching step, [2, 69] used a nearest neighbor classifier directly in the computed feature space with Chi square as a dissimilarity measure. In [67], authors considered the face recognition task as a two-class problem by classifying every two face images as intra-personal or extra-personal ones and outputting a similarity describing how confident two face images are of one person. Similarly, the Chi square distance between corresponding LBP histograms is used for classification. The Adaboost is applied to select the most significant intra/extra features from a large feature set.

Recently, LBP features are extended to represent face images in a multi-scale manner. In [70], LBP features were generated from images after convolution with a set of Gabor kernels instead of original images. The Gabor kernels are at five different scales, which leads to the multi-scale characteristics of the resulting LBP features. In [71], an alternative but a much simple discriminative descriptor containing the information from a multi-scale analysis is proposed. Firstly, LBP operators at R scales are applied to a face image. The resulting LBP images are cropped to the same size and divided into non-overlapping sub-regions, where a set of histograms is computed for each sub-region. By concatenating these histograms into a single histogram, the final multi-scale regional face descriptor is obtained. This regional facial descriptor is used to measure the face similarity by summing the similarities between all the regional histograms. PCA is employed to reduce the dimension of the descriptor. And LDA is followed to derive discriminative facial features. Thus a regional discriminative facial descriptor is defined by projecting the histogram information into LDA space. After the projection, the similarity measurement is obtained by summing the similarity, i.e. normalized correlation, of regional discriminative descriptors.

Obviously, local binary patterns are invariant under any monotonic transformation of the pixel gray values. However, LBP is significantly affected by non-monotonic gray

value transformations. Unfortunately, faces are not flat, in contrast to the flat surfaces where texture images are usually captured. Therefore, non-monotonic gray value transformations (manifested by shadows and bright spots) occur in face images, which cannot be handled by LBP. Moreover, in most LBP-based methods, face images are divided into small sub-regions. In this case, facial components may be partitioned into different sub-regions in various face images, which renders low similarity score during matching process.

2.2.5 Local Feature Detectors and Descriptors

Unlike Gabor features or LBP operators extracting features in a manner of pixel-wise, local feature detectors [72, 73, 3] extract image patterns which differ from their immediate neighborhood. Obviously, blob-like and corner-like structures are candidates of local features.

2.2.5.1 Local Feature Detectors

A. Corner Detectors

A vast number of corner detector methods have been proposed in the literature. The SUSAN (Smallest Univalued Segment Assimilating Nucleus) corner detector is a morphological approach introduced by Smith and Brady [74]. For each pixel in the image, a circular neighborhood of fixed radius is considered around it. The center pixel is referred to as the nucleus whose intensity value is used as reference. Then, depending on whether all other pixels within this circular neighborhood have similar intensity values as the nucleus or different intensity values, they are partitioned into two groups. In this way, each image pixel has associated with a local area of similar brightness (coined USAN), whose size contains vital information about the structure of the image at that pixel. In more or less homogeneous parts of the image, almost the entire circular neighborhood is covered by a local area of similar brightness. Near edges, this ratio drops to 50%, and

near corners it reduces to about 25%. Hence, corners are detected as locations in the image where the number of pixels with similar intensity value in a local neighborhood reaches a local minimum and is below a predefined threshold (Fig. 2.4).

Figure 2.4: SUSAN corners are detected by segmenting a circular neighborhood into similar (orange) and dissimilar (blue) regions. Corners are located where the relative area of the similar region (USAN) reaches a local minimum below a certain threshold [3].

The FAST detector, introduced by Rosten and Drummond in [75, 76] is built on the SUSAN detector. In FAST, only pixels on a circle of 16 pixels around the corner candidate are compared (see Fig. 2.5). Initially pixels 1 and 2 are compared with a threshold, then 3 and 4 as well as the remaining ones at the end. The pixels are classified into three subsets: Dark, similar, and brighter. The ID3 algorithm [77] is used to select the candidate corner pixel, which is measured by the entropy of the positive and negative corner classification responses at this pixel. This results in a very efficient detector but not invariant to scale changes. An extension to a multi-scale detector by scale selection with the Laplacian function was proposed in [78].

The Harris corner detector [79] is based on the second moment matrix often used to detect features and to describe local image structures. The gradient distribution of a pixel in a local neighborhood is described by this matrix as:

$$M = \sigma_D^2 g(\sigma_I) * \begin{bmatrix} I_x^2(\mathbf{x}, \sigma_D) & I_x(\mathbf{x}, \sigma_D) I_y(\mathbf{x}, \sigma_D) \\ I_x(\mathbf{x}, \sigma_D) I_y(\mathbf{x}, \sigma_D) & I_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \quad (\text{Eq. 2.37})$$

Figure 2.5: Illustration of pixels examined by the FAST detector [3].

where $I_x(\mathbf{x}, \sigma_D) = \frac{\partial}{\partial x} g(\sigma_D) * I(\mathbf{x})$, $g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$. Gaussian kernels of scale σ_D , which is the differentiation scale, are used to compute the local image derivatives. By smoothing with a Gaussian window of scale σ_I , which is the integration scale, the computed derivatives are then averaged in the neighborhood of the pixel. The principal signal changes in two orthogonal directions in a neighborhood around the point defined by σ_I are represented by the eigenvalues of this matrix. Then corners are found as locations in the image where the signal varies significantly in both directions (both eigenvalues are large). Harris proposed to avoid computing the eigenvalues directly by measuring the cornerness:

$$\text{cornerness} = \det(M) - \lambda \text{trace}(M), \quad (\text{Eq. 2.38})$$

where $\det(M)$ is the determinant and $\text{trace}(M)$ is the trace of the matrix M . Since the trace of a matrix is equal to the sum of its eigenvalues and the determinant corresponds to the product, it is clear that high values of the cornerness measure correspond to both eigenvalues being large.

Mikolajczyk and Schmid [4] extended the Harris detector to scale invariant Harris-Laplace detector and affine invariant Harris-Affine detector. In the scale invariant Harris-Laplace detector, a multi-scale Harris corner detector is initialized to determine the location of candidate corners. The characteristic scale is then determined based on scale selection as proposed by Linderberg [80]. As the name Harris-Laplace suggests, the

Laplacian is applied to select appropriate scales. In the affine invariant Harris-Affine detector, by the Harris-Laplace detector, a set of initial points are extracted at their characteristic scales. The iterative estimation of elliptical affine regions as proposed by Lindeberg [81, 82] allows to obtain affine invariant corners. These regions are ellipses instead of circular. The procedure consists of the following steps:

- (i) Detect the initial region by the Harris-Laplace detector.
- (ii) Estimate the shape by the second moment matrix.
- (iii) Normalize the affine region to a circular one.
- (iv) Re-detect the new location and scale in the normalized image.
- (v) Go back to stage 2 if the eigenvalues of the second moment matrix of the new keypoint are not the same.

In [8, 7], a more heuristic technique is proposed to obtain affine invariance by exploiting the geometry of the edges in the proximity of a Harris corner. In this algorithm, a corner point $\mathbf{p}$ [79] and its nearby edge [83] are detected at multiple scale. Suppose along the edge as depicted in Fig. 2.6, there are two points $\mathbf{p}_1$ and $\mathbf{p}_2$ moving away from the corner in both directions. Their relative speed is linked with the equality of relative affine invariant parameters l_1 and l_2 :

$$l_i = \int abs(|\mathbf{p}_i^{(1)}(s_i)\mathbf{p} - \mathbf{p}_i(s_i)|)ds_i \quad (\text{Eq. 2.39})$$

with s_i an arbitrary curve parameter ($i = 1, 2$), $\mathbf{p}_i^{(1)}(s_i)$ the first derivative of $\mathbf{p}_i(s_i)$ with respect to s_i , $abs()$ the absolute value and $|\dots|$ the determinant. In this case, the areas between the joint $\langle \mathbf{p}, \mathbf{p}_1 \rangle$ and the edge, between the joint $\langle \mathbf{p}, \mathbf{p}_2 \rangle$ and the edge is the same. l is used to refer to $l_1 = l_2$.

Figure 2.6: The corner point $\mathbf{p}$ is the starting point of the edge-based region detector which exploits nearby edge information.

For each l , the points $\mathbf{p}_1(l)$, $\mathbf{p}_2(l)$ and the corner $\mathbf{p}$ define a parallelogram $\Omega(l)$ which is spanned by the vectors $\mathbf{p}_1(l) - \mathbf{p}$ and $\mathbf{p}_2(l) - \mathbf{p}$ (see Fig. 2.6). One or a few parallelograms are selected to maximize:

$$\begin{aligned} Inv_1 &= abs \left(\frac{|\mathbf{p}_1 - \mathbf{p}_g \quad \mathbf{p}_2 - \mathbf{p}_g|}{|\mathbf{p} - \mathbf{p}_1 \quad \mathbf{p} - \mathbf{p}_2|} \right) \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - (M_{00}^1)^2}} \\ Inv_2 &= abs \left(\frac{|\mathbf{p} - \mathbf{p}_g \quad \mathbf{q} - \mathbf{p}_g|}{|\mathbf{p} - \mathbf{p}_1 \quad \mathbf{p} - \mathbf{p}_2|} \right) \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - (M_{00}^1)^2}} \end{aligned} \quad (\text{Eq. 2.40})$$

with

$$\begin{aligned} M_{pg}^n &= \int_{\Omega} I^n(x, y) x^p y^q dx dy \\ \mathbf{p}_g &= \left(\frac{M_{10}^1}{M_{00}^1}, \frac{M_{01}^1}{M_{00}^1} \right) \end{aligned} \quad (\text{Eq. 2.41})$$

with M_{pg}^n the n th order, $(p + q)$ th degree moment computed in the region $\Omega(l)$. $\mathbf{p}_g$ is the center of gravity of the region weighted with intensity $I(x, y)$. $\mathbf{q}$ is the corner of the parallelogram opposite to the corner point $\mathbf{p}$ (as shown in Fig. 2.6).

In the 3D world, corner based features do not necessarily correspond to real corners. However, in computer vision, the goal is to extract stable features. In many independent

evaluations [84, 85, 72], the Harris detector was identified as the most stable one. There are also multi-scale and scale and affine invariant extensions of this approach. The affine transformation model is invariant to viewpoint changes only in case of locally planar regions and assuming the camera is relatively far from the object. However, corners are often found near object boundaries as this is where the intensity change usually occurs. Hence, the region extraction process is often based on measurements on non-planar structures. In these cases, the viewpoint invariance will be limited and also the robustness to background changes will be affected. Alternatively, the SUSAN detector can be used, which is more efficient but also more sensitive to noise.

On the positive side, corners are better localized in the image plane compared to other types of features. However, their scale is not well defined as a corner structure changes very little over a wide range of scales. The reason why scale selection still works with the Harris detector is that the feature point is localized not exactly on the corner edge but slightly inside the corner structure.

B. Blob Detectors

After corners, blobs which provide complementary features to the corners are the second most distinctive local features. Firstly, a derivative-based method “the Hessian detector” is reviewed. Then the scale-invariant and affine invariant extensions of this method coined Hessian-Laplace and Hessian-Affine are considered. Finally, a salient region detector which is based on the entropy of the intensity probability distribution is described.

Hessian matrix is the second 2×2 matrix obtained from the Taylor expansion of the image intensity function $I(x)$, as below:

$$H = \begin{bmatrix} I_{xx}(x, \sigma_D) & I_{xy}(x, \sigma_D) \\ I_{xy}(x, \sigma_D) & I_{yy}(x, \sigma_D) \end{bmatrix}, \quad (\text{Eq. 2.42})$$

where I_{xx} , I_{xy} and I_{yy} are the second-order Gaussian smoothed image derivatives. Please note the filters based on the determinant and the trace of this matrix. The latter is

the Laplacian, a separable linear filter, which has one major drawback in the context of blob extraction. Local maxima are often found near contours or straight edges, where the signal change is only in one direction. These maxima are less stable because their localization is more sensitive to noise or small changes in neighboring texture. To solve this problem, a location and scale is selected for which the trace and the determinant of the Hessian matrix simultaneously reach a local extremum.

Figure 2.7: Illustration of the components of the Hessian matrix and Hessian determinant [3].

The feature detection process based on the Hessian matrix is illustrated in Fig. 2.7. Given the original image (as shown in upper left), the second-order Gaussian smoothed image derivatives (lower part) are firstly computed, which are then used to compute the determinant of the Hessian (upper right). The second-order derivatives are symmetric filters, thus they give weak responses exactly in the point where the signal change is most significant. Therefore, the maxima are localized at ridges and blobs for which the size of the Gaussian kernel σ_D matched by the size of the blob structure.

The Difference-of-Gaussians detector, or DoG for short, has been proposed in [86, 87, 88, 9, 10]. It is a scale-invariant detector which extracts blobs in the image by approximating the Laplacian $L_{xx}^2 + L_{yy}^2$. Gaussian blurring is often applied to generate images at various scales. Hence, the DoG images produce responses which approximate

the LoG. The computation of second-order derivatives in x and y directions is then avoided.

In the context of real-time face detection, Viola and Jones [89] have proposed to use integral images, which can be applied to obtain a fast approximation of the Hessian matrix, as used in Speeded-Up Robust Features (SURF) [90]. The Hessian matrix is roughly approximated, using a set of box-type filters, and no smoothing is applied when going from one scale to the next.

Gaussians are optimal for scale-space analysis, but in practice they have to be discretized. SURF pushes the approximation even further using the box filters. In spite of the rough approximations, the performance of the feature detector is comparable to the results obtained by the discretized Gaussian. SURF has been reported to be more than five times faster than DoG.

The Hessian-Laplace and Hessian-Affine detectors proposed by Mikolajczyk and Schmid [4] are similar as Harris-Laplace and Harris-Affine, except that they compute the determinant of the Hessian rather than the Harris corners, which turns the methods into viewpoint invariant blob-detectors. A large number of features can be extracted leading to a good coverage of the image. Furthermore, this detector also responds to some corner structures at fine scale (see Fig. 2.7). The locations of detected keypoints are more suitable for scale estimation than the Harris points due to the use of second-order Gaussian derivatives for both spatial and scale localization.

C. Region Detectors

In this part, a number of region detectors are discussed, which aim to extract distinctive image regions. First, an approach to extract intensity-based regions (IBR) is discussed, followed by maximally stable extremal regions (MSER). At last, the approach “superpixels” is considered.

Tuytelaars and Van Gool [6, 7] proposed a method to detect affine invariant regions, or intensity-based regions (IBR). The intensity function along rays which emanates from

Figure 2.8: Construction of intensity-based regions [3].

the extremum is studied, given a local extremum in intensity (see Fig. 2.8). Along each ray, the following function is computed:

$$f(t) = \frac{\text{abs}(I(t) - I_0)}{\max(\frac{\int_0^t \text{abs}(I(t) - I_0) dt}{t}, d)}, \quad (\text{Eq. 2.43})$$

where t is an arbitrary parameter along the ray, $I(t)$ is the intensity at position t , I_0 is the intensity value at the extremum, and d is a small number which has been added to prevent a division by zero. Typically, a maximum is achieved at locations where the intensity changes suddenly. Points are selected where this function reaches an extremum. Then an affine invariant region is enclosed by linking all points corresponding to maxima of $f(t)$ along rays from the same local extremum. An ellipse is used to replace this region if they have the same shape moments up to the second order.

Matas *et al.* [5] proposed an approach to extract Maximally Stable Extremal Regions (MSER). All pixels inside the MSER have either higher or lower intensity than all the pixels on the outer boundary. The enumeration of the set of extremal regions ξ is almost linear in the number of image pixels. The enumeration proceeds as below. First, pixels are sorted by intensity. Then, pixels in the image are marked and the list of growing and merging connected components and their areas is maintained. During the enumeration process, the area of each connected component as a function of intensity is stored. Since every extremal region is a connected component of a thresholded image, detection of

MSER is related to thresholding. However, the threshold is not optimal, all thresholds are tested and the stability of the connected components evaluated. The MSER features typically anchor on region boundaries, thus the resulting regions are accurately localized compared to other blob-detectors. The method works best for an image with uniform regions separated by strong intensity changes.

The two methods described above extract small regions whose intensity patterns are different from their immediate surroundings. However, image segments are too large to be used as local features. Mori *et al.* [91], Ren and Malik [92] proposed to group pixels into atomic regions, which refer to the resulting superpixels. In [91, 92], normalized cuts [93] is used to extract superpixels from the image. In contrast to traditional local features, superpixels cover the entire image and do not overlap. All superpixels extracted from an image have similar scale, so the method is not scale-invariant. An alternative construction method, based on Constrained Delauney Triangulation, has been proposed to obtain robustness against scale change [94].

The local features detected with the methods described above typically represent homogeneous regions, which incurs problems for the later description and matching. This problem is overcome by increasing the measurement region. In other words, a larger scale region is used to compute the descriptor, which also contains part of the surrounding image structures and captures the shape of the region boundary. Intensity-based regions and maximally stable extremal regions typically give very similar features. These methods are therefore not complementary. IBR breaks down when the region is non-convex, but it is more robust to small gaps in the region contour. In contrast, MSER has been shown to be relatively sensitive to image blur. Region detectors often detect blob-like structures, so they are less complementary to blobs than to corners. Region-based detectors are typically quite accurate in their localization. They work especially well for images with a well structured scene, clearly delineated regions. Even though superpixels share

some characteristics with the other region detectors, they are not the same. They are non-overlapping, and cover the entire image. Most importantly, the idea of superpixels is to speed up the image analysis by focusing on the superpixels only instead of analyzing all pixels. This is in contrast to local features which should be distinctive and, in the ideal case, uniquely identifiable. However, using region boundaries to build distinctive descriptors may overcome the occlusion problem from which traditional interest points suffer.

D. Summary on the Detectors

Feature Detector	Corner	Blob	Region	Rotation invariant	Scale invariant	Affine invariant	Localization			
							Repeatability	accuracy	Robustness	Efficiency
Harris	✓			✓			+++	+++	+++	++
Hessian		✓		✓			++	++	++	+
SUSAN	✓			✓			++	++	++	+++
Harris-Laplace	✓	(✓)		✓	✓		+++	+++	++	+
Hessian-Laplace	(✓)	✓		✓	✓		+++	+++	+++	+
DoG	(✓)	✓		✓	✓		++	++	++	++
SURF	(✓)	✓		✓	✓		++	++	++	+++
Harris-Affine	✓	(✓)		✓	✓	✓	+++	+++	++	++
Hessian-Affine	(✓)	✓		✓	✓	✓	+++	+++	+++	++
Salient Regions	(✓)	✓		✓	✓	(✓)	+	+	++	+
Edge-based	✓			✓	✓	✓	+++	+++	+	+
MSER			✓	✓	✓	✓	+++	+++	++	+++
Intensity-based			✓	✓	✓	✓	++	++	++	++
Superpixels			✓	✓	(✓)	(✓)	+	+	+	+

Figure 2.9: Overview of local feature detectors [3].

Tuytelaars and Mikolajczyk [3] provided a detailed survey about local feature detectors including their performance evaluation. Fig.2.9 gives an overview of the most important properties for the feature detectors. The feature detectors in this figure are organized in 4 groups according to their invariance: rotation, similarity, affine, and perspective. For rotation invariant features the highest repeatability and localization accuracy in many tests has been obtained by the Harris detector. The Hessian detector finds blobs which are not as well localized and requires second-order derivatives to be computed. The SUSAN detector avoids computation of derivatives and is known for its efficiency, however the absence of smoothing makes it more susceptible to noise.

In the scale-invariant group Harris-Laplace shows high repeatability and localization accuracy [72]. However, its scale estimation is less accurate due to the multi-scale nature of corners. Hessian-Laplace is more robust than its single scale version [73]. This is due to the fact that blob-like structures are better localized in scale than corners. DoG and SURF detectors were designed for efficiency and the other properties are slightly compromised. However, for most coverage of the image are crucial in recognition applications, where localization accuracy is less important. Thus, Hessian-Laplace detectors have been successful in various categorization tasks although there are detectors with higher repeatability rate. Although they result in far less compact representations than the interest points, random and dense sampling provide good results which confirms the coverage requirements of recognition methods [95]. DoG detector performs extremely well in matching [96] and image retrieval [9] probably due to a good balance between spatial localization and scale estimation accuracy.

Note that for the scale and affine invariant detectors, the difference between corner and blob detectors becomes small, with most detectors detecting both features, although they still show a preference for either type.

Salient regions require to compute a histogram and its entropy for each region candidate in scale or affine space, which results in large computational cost [73]. On the positive side the regions can be ranked according to their complexity or information content. Some applications exploit this and use only a small subset of the salient regions while still obtaining a good performance [97]. Originally, they were only scale-invariant, but later they have been extended to affine invariant. The edge based regions focus on corners formed by edge junctions which gives good localization accuracy and repeatability but the number of detected features is small.

The region detectors are based on the idea of segmenting boundaries of uniform regions. Intensity based regions use a heuristic method and find similar regions to MSER.

Superpixels are typically based on segmentation methods which are computationally expensive like normalized cuts. The level of invariance of superpixels depends mostly on the segmentation algorithm used. In contrast to superpixels, MSER selects only the most stable regions which results in high repeatability. In the case of category recognition variability of object appearance dominates over deformations due to viewpoint changes and affine invariance typically brings little improvement.

2.2.5.2 Local Feature Descriptors

Many different techniques for describing local image regions have been developed. The simplest descriptor is a vector of image pixels, which leads to high computational complexity.

A. Distribution-based Descriptors

In this category, histograms are used to represent different properties of shape or appearance, where a histogram representing the distribution of the pixel intensities is a simple descriptor. Johnson and Hebert [98] proposed spin image, which is a histogram of the point positions in the neighborhood of a 3D interest point, which is applied to images in [99].

Zabih and Woodfill [100] developed a descriptor robust to variations in illumination, which relies on histograms of ordering and reciprocal relations between pixel intensities. This descriptor is called local binary pattern which is suitable for texture representation. In [101], the author reported that to build a reliable descriptor, a large number of dimensions is required. Please refer to Section 2.2.4 for more detail.

Lowe [10, 9] proposed a scale invariant feature transform (SIFT), which consists of a scale invariant local feature detector (DoG) and a descriptor based on the gradient distribution around the area of the detected features. Ke and Sukthankar [102] further proposed an approach named PCA-SIFT, where PCA is applied to the image gradient

patch to reduce the dimension of the descriptor. Gradient location-orientation histogram (GLOH) [103] is an extension of the SIFT descriptor designed to increase its robustness and distinctiveness. The SIFT descriptor is computed in a log-polar location grid with three bins in radial direction and 8 in angular direction, which results in 17 location bins. The gradient orientations are quantized in 16 bins. This gives a 272 bin histogram. The size of this descriptor is reduced with PCA. The final descriptor is the 128 largest eigenvectors. Geometric histogram [104] and shape context [105] implement the same idea and are very similar to the SIFT descriptor. Both methods compute a histogram describing the edge distribution in a region.

B. Spatial-frequency Techniques

Many techniques describe the frequency content of an image. The Fourier transform decomposes the image content into the basis functions. However, this representation is difficult to adapt to a local approach because the spatial relations between points are not explicit and the basis functions are infinite. The Gabor transform [61] overcomes these problems, with a drawback that to capture small changes in frequency and orientation a large number of Gabor filters is required. Gabor filters and wavelets are frequently explored in the context of texture classification (see Section 2.2.2).

C. Differential Descriptors

In differential descriptors, a set of image derivatives computed up to a given order approximates a point neighborhood. Koenderink and Van Doorn [106] investigated the characteristics of local derivatives, or local jet. Florack et al. [107] derived differential invariants, which combine components of the local jet to obtain rotation invariance. Freeman and Adelson [108] developed steerable filters, which steer derivatives in a particular direction given the components of the local jet. Steering derivatives in the direction of the gradient makes them invariant to rotation. A stable estimation of the derivatives is obtained by convolution with Gaussian derivatives.

Baumberg [109], Schaffalitzky and Zisserman [110] proposed to use complex filters derived from the family $K(x, y, \theta) = f(x, y)\exp(i\theta)$, where θ is the orientation. For the function $f(x, y)$, Baumberg uses Gaussian derivatives and Schaffalitzky and Zisserman apply a polynomial. These filters are different from the Gaussian derivatives.

D. Other Techniques

Van Gool *et al.* [111] proposed generalized moment invariants to describe the multi-spectral nature of the image data. The invariants combine central moments defined by $M_{pq}^a = \int \int_{\Omega} x^p y^q [I(x, y)]^a dx dy$ of order $p + q$ and degree a . The moments give information of shape and intensity distribution in a region Ω . They can be easily computed for any order and degree because of the independency. However, the moments of high order and degree are sensitive to small geometric and photometric distortions. Computing the invariants reduces the number of dimensions. These descriptors are therefore more suitable for color images where the invariants can be computed for each color channel and between the channels.

E. Summary on the Descriptors

Mikolajczyk and Schmid [103] presented the performance comparisons of different interest region descriptors (SIFT, GLOH, shape context, PCA-SIFT, spin images, steerable filters, differential invariants, complex filters, moment invariants, and cross-correlation of sampled pixel values) in the presence of real geometric and photometric transformations. The evaluation was designed for matching and recognition of the same object or scene.

In most of the tests, GLOH performs best. And SIFT obtains the second best, which demonstrates that SIFT descriptor is robust and distinctive.

2.2.5.3 Scale Invariant Feature Transform

As mentioned previously, DoG detectors perform well in image matching [96]. The SIFT descriptor [10, 9] has shown good performances on various machine vision applications

Figure 2.10: (a) For each octave of scale space, the initial image is repeatedly convolved with Gaussian to produce the set of scale space images shown on the left. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images on the right. After each octave, the Gaussian image is down-sampled by a factor of 2, and the process repeated; (b) Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with cross) to its 26 neighbors in 3×3 regions at the current and adjacent scales (marked with circles).

[112, 113, 96]. Recently, some initial attempts apply the SIFT algorithm in the face recognition task. In [114] an overlapping sub-image matching strategy is used as the first attempt to explore the SIFT approach for face recognition. In [115], the SIFT descriptor is adopted to describe irregular local marks detected by a Hessian-Laplace detector. In [116], [117], and [118], a graph is built on SIFT features. The recognition problem is modeled as a graph matching process. In [119], the authors propose a method based on SIFT and Support Vector Machine. Fernandez and Vicente [120] combine Harris-Laplace and Difference-of-Gaussian detectors to detect both corner and blob structures in face images and use SIFT descriptors to represent them. In [121], [122] and [123], salient regions are firstly identified from face images, and then SIFT features are extracted in each region. Majumdar and Ward [124] rank the SIFT features according to their discriminative power and use the most discriminating ones for face recognition. In [125], multi-scale local features are extracted from a dense set of multi-scale image patches that deliver good generalization ability for face recognition. Experimental results reported in these papers show that SIFT approach can detect and describe distinct local features from face images effectively. Hence, I mainly introduce SIFT approach in this category.

The basic idea of SIFT algorithm is to detect distinctive invariant blob-like features in images. Then corresponding descriptors are computed based on the histogram of oriented gradients in the support area of the detected local features. The major steps of SIFT algorithm include:

- (i) Scale-space extrema detection. The first stage of keypoint detection is to identify locations and scales that can be repeatedly assigned under different views of the same object. The Difference of Gaussian function $D(x, y, \sigma)$ is used to approximate the scale-normalized Laplacian of Gaussian $\sigma^2 \nabla^2 G$ for scale invariance (see Fig. 2.10a). In order to detect the local maxima and minima of $D(x, y, \sigma)$, each sample point is compared to its eight neighbors in the current image and nine neighbors

in the scale above and below. It is selected only if it is larger than all of these neighbors or smaller than all of them (see Fig. 2.10b).

- (ii) Keypoint localization. In this step points that have low contrast (and are therefore sensitive to noise) or are poorly localized along an edge are to be rejected. Firstly, interpolation is done using the quadratic Taylor expansion of the Difference-of-Gaussian scale-space function $D(x, y, \sigma)$ with the candidate keypoint as the origin. This Taylor expansion is given by:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2} \mathbf{x}, \quad (\text{Eq. 2.44})$$

where D and its derivatives are evaluated at the sample point and $\mathbf{x} = (x, y, \sigma)$ is the offset from this point. The location of the extremum $\hat{\mathbf{x}}$ is determined by taking the derivative of this function with respect to $\mathbf{x}$ and setting it to zero, giving:

$$\hat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2} \frac{\partial D}{\partial \mathbf{x}}. \quad (\text{Eq. 2.45})$$

If the offset $\hat{\mathbf{x}}$ is larger than 0.5 in any dimension, then that's an indication that the extremum lies closer to another candidate keypoint. In this case, the candidate keypoint changes and the interpolation performed instead about that point. Otherwise, the offset is added to its candidate keypoint to get the interpolated estimate for the location of the extremum. It is known that Laplacian operator is very sensitive to noise, so keypoints with low contrast should be eliminated which are probably produced by noise rather than the inherent image structures. This can be obtained by computing the value of the second-order Taylor expansion $D(\mathbf{x})$ at the offset $\hat{\mathbf{x}}$, giving:

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2} \frac{\partial D^T}{\partial \mathbf{x}} \hat{\mathbf{x}}. \quad (\text{Eq. 2.46})$$

If this value is less than 0.03, the candidate keypoint is discarded. Otherwise it is kept, with final location $\mathbf{y} + \hat{\mathbf{x}}$ and scale σ , where $\mathbf{y}$ is the original location of

the keypoint at scale σ . In order to increase stability, Lowe proposed to further eliminate the keypoints that have poorly determined locations but have high edge responses. A poorly defined peak in the Difference-of-Gaussian function has a large principal curvature across the edge but a small one along it. The principal curvatures can be computed from the second-order Hessian matrix $\mathbf{H}$ computed at the location and scale of the keypoint:

$$\mathbf{H} = \begin{bmatrix} \mathbf{D}_{xx} & \mathbf{D}_{xy} \\ \mathbf{D}_{xy} & \mathbf{D}_{yy} \end{bmatrix} \quad (\text{Eq. 2.47})$$

The eigenvalues of $\mathbf{H}$ are proportional to the principal curvatures of $\mathbf{D}$. It turns out that the ratio of the two eigenvalues, say α is the larger one, and β the smaller one, with ratio $r = \frac{\alpha}{\beta}$, is sufficient for this purpose. Then, the sum of the eigenvalues can be computed from the trace of $\mathbf{H}$ and their product from the determinant:

$$\begin{aligned} Tr(\mathbf{H}) &= \mathbf{D}_{xx} + \mathbf{D}_{yy} \\ &= \alpha + \beta, \end{aligned} \quad (\text{Eq. 2.48})$$

$$\begin{aligned} Det(\mathbf{H}) &= \mathbf{D}_{xx}\mathbf{D}_{yy} - (\mathbf{D}_{xy})^2 \\ &= \alpha\beta. \end{aligned} \quad (\text{Eq. 2.49})$$

The ratio $\mathbf{R} = \frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})}$ can be shown to be equal to $\frac{(r+1)^2}{r}$, which depends only on the ratio of the eigenvalues rather than their individual values. $\mathbf{R}$ is minimum when the eigenvalues are equal to each other. Therefore the higher the absolute difference between the two eigenvalues, which is equivalent to a higher absolute difference between the two principal curvatures of $\mathbf{D}$, the higher the value of $\mathbf{R}$. It follows that, for some threshold eigenvalue ration r_{th} , if $\mathbf{R}$ for a candidate keypoint is larger than $\frac{(r_{th}+1)^2}{r_{th}}$, that keypoint is poorly localized and hence rejected. Lowe used a value of $r = 10$.

- (iii) Orientation assignment. The scale of the keypoint is used to select the Gaussian smoothed image L , with the closest scale. For each image sample $L(x, y)$ at this scale, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are computed using pixel differences:

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

$$\theta(x, y) = \arctan((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y))).$$

An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint. The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. The highest peak in the histogram is detected as the dominant direction of the keypoint. And any other local peak that is within 80% of the highest peak is used to create a keypoint with that orientation.

- (iv) Descriptor representation. To compute the keypoint descriptor, the image gradient magnitudes and orientations are sampled around the keypoint location, using the scale of the keypoint to choose the level of Gaussian blur for the image. To achieve, orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation. A Gaussian weighting function with σ equal to one half the width of the descriptor window is used to assign a weight to the magnitude of each sample point. Then the support area in the descriptor window is divided into 4×4 blocks. An 8-bin oriented gradient histogram is computed in each block. Thus, a descriptor vector for each keypoint has $4 \times 4 \times 8 = 128$ dimensions. To reduce the effect of illumination changes, first the histogram vector is normalized to the unit length. Then the values in the unit

length vector are thresholded to be no larger than 0.2, and then renormalized to unit length.

In the recognition stage, as described in [9], the initial correspondences of two keypoint sets are established by the nearest neighbor search. A more effective measure is to compare the distance of the closest neighbor to that of the second-closest neighbor. If the ratio of distance is greater than some threshold, the match is rejected. Then the Generalized Hough Transform is applied to group keypoints with similar poses. If the number of keypoints in the Generalized Hough Transform bin is larger than 3, an affine transformation with least-square solution is performed on the keypoints in this bin to further check for the geometric agreement. And the final decision to accept or reject a model hypothesis is based on a probabilistic model given in [112].

2.2.6 Face Recognition based on Multi-scale Local Features

Figure 2.11: Facial Components Modeled at Different Scales

The question is what types of local features should be used. One of the very fundamental problems that arises when analyzing face images originates from the fact that

faces appear in different ways depending upon the scale of observation. Intuitively, facial components are modeled at different levels of scales, ranging from skin textures and facial wrinkles at fine scale, through eyes and mouth represented at median scale, to the whole face region at larger scale, as shown in Figure 2.11. Notably, the types of facial component description are strongly dependent on the scale at which the face is modeled, and this is in clear contrast to certain idealized mathematical entities, such as “point” or “line”, which are independent of the scale of observation. In certain controlled situations, appropriate scales for analysis may be known a priori. Under other circumstance, however, it is not obvious at all how to determine in advance what are the proper scales. One such example is a vision system with the task of analyzing unknown face images. Besides the inherent multi-scale properties of those face images (which, in general, are unknown), such a system has to face the problems that the perspective mapping gives rise to size variations, that noise is introduced in the image formation process, and that the available data are two-dimensional data sets reflecting only indirect properties of a three-dimensional world. To be able to cope with these problems, an image representation that explicitly incorporates the notion of scale is a crucially important tool whenever interpreting sensory data, such as images, by automatic methods.

In computer vision and image processing, these insights have led to the construction of multi-scale representations of image data, obtained by embedding any given signal into a one-parameter family of derived signals. This family should be parameterized by a scale parameter and be generated in such a way that fine-scale structures are successively suppressed when the scale parameter is increased. A main intention behind this construction is to obtain a separation of the image structures in the original image, such that fine scale image structures only exist at the finest scales in the multi-scale representation. Thereby, the task of operating on the image data is simplified, provided that the operations are performed at sufficiently coarse scales where unnecessary and irrelevant fine-scale structures have been suppressed.

Gabor features and multi-scale Local Binary Patterns fall into this category. As mentioned in Section 2.2.2, Gabor wavelets of 5 scales and 8 orientations are used for face recognition. In EBGM [58], Gabor wavelet coefficients were calculated at fiducial points, which are nodes referring to specific facial landmarks, e.g., the pupils, the corners of the mouth, the tip of the nose, the top and bottom of the ears, etc.. Although the fiducial points refer to corresponding object locations, some may be occluded, and jets as well as distances vary due to rotation in depth. To be able to compare graphs of different poses, pointers are manually defined to associate corresponding nodes in the different graphs. Hence, the first set of graphs is generated manually. Nodes are located at fiducial points and edges between the nodes as well as correspondences between nodes of different poses are defined. Similarly, feature vectors are generated at the feature points as a composition of Gabor wavelet coefficients in [59]. However, the features are not limited to eyes, nose, etc., i.e. special facial features such as dimples are also extracted. Feature vectors are extracted at points with high information content on the face image done by searching the maximum valued pixel in a window W_0 of size $W \times W$ by the following procedure: A feature point is located at (x_0, y_0) , if

$$R_j(x_0, y_0) = \max_{(x,y) \in W_0} (R_j(x, y)), \quad (\text{Eq. 2.50})$$

and

$$R_j(x_0, y_0) > \frac{1}{N_1 N_2} \sum_{x=1}^{N_1} \sum_{y=1}^{N_2} R_j(x, y), j = 1, \dots, 40 \quad (\text{Eq. 2.51})$$

where R_j is the response of the face image to the j^{th} gabor filter. $N_1 N_2$ is the size of face image, the center of the window, W_0 is at (x_0, y_0) .

Compared with the methods extracting Gabor features at certain points, [126, 60, 127] represented face images by convolving the image with a family of Gabor kernels as defined by Eq. 2.34. The result Gabor feature set thus consists of the convolution results of an input image $I(x, y)$ with all of the 40 Gabor filters, which leads to a very high space

dimension. To select the most useful features for classification, [60, 127] performed Fisher linear discriminant analysis to project the high dimensional visual stimuli (face images) into a lower dimensional space with a high discrimination index. And in [126], the most discriminative Gabor features were selected by the Adaboost procedure, which were then subjected to the generalized discriminant analysis process for further class separability enhancement.

As mentioned in Section 2.2.4, LBP features can be extended to multi-scale. In [70], LBP features were generated from images after convolution with a set of Gabor kernels as in Eq. 2.34 instead of original gray value images. Face images were then modeled as a “histogram sequence” by concatenating the histograms of all the local regions of all the local Gabor magnitude binary pattern maps. For recognition, histogram intersection is used to measure the similarity of different Local Gabor Binary Pattern Histogram Sequences and the nearest neighborhood is exploited for final classification, but the computational complexity and the feature dimension are very high. In [71], an alternative but a much simple discriminative descriptor containing the information from a multi-scale analysis is proposed. By varying the sampling radius R , and combining the LBP images, a multiresolution representation based on LBP, called multi-scale local binary patterns can be obtained. Uniform local binary pattern operators at R scales are first applied to a face image. This generates a grey level code for each pixel at every resolution. The resulting LBP images, are cropped to the same size and divided into non-overlapping sub-regions, $M_0, M_1, \dots, M_{J-1}$. The regional pattern histogram for each scale is computed based on:

$$H_{P,r,j}^{u2}(i) = \sum_{(x',y') \in M_j} B(LBP_{P,r}^{u2}(x',y') = i) \quad (\text{Eq. 2.52})$$

where $i \in [0, (P-1)P+2]$, $r \in [1, R]$, $j \in [0, J]$, and $B(x) = \begin{cases} 1 & \text{when } x = 0 \\ 0 & \text{otherwise} \end{cases}$, which is a Boolean indicator. The set of histograms computed at different scales for region M_j

provides regional information. By concatenating these histograms into a single histogram, the final multiresolution regional face descriptor is presented as:

$$F_j = [H_{P,1,j}^{u2}, H_{P,2,j}^{u2}, \dots, H_{P,R,j}^{u2}] \quad (\text{Eq. 2.53})$$

This regional facial descriptor can be used to measure the face similarity by summing the similarities between all the regional histograms. However, by directly applying the similarity measurement to the multi-scale LBP histogram, the performance is compromised. The reason is that this histogram is of high dimensionality and contains redundant information. The dimension of the descriptor can be reduced by employing PCA before LDA. PCA is used to extract the statistically independent information as a basis for LDA to derive discriminative facial features. Thus a regional discriminative facial descriptor D_j , is defined by projecting the histogram information F_j into LDA space W_j^{lda} , i.e.:

$$D_j = (W_j^{lda})^T F_j \quad (\text{Eq. 2.54})$$

After the projection, the similarity measurement presented below is obtained by summing the similarity, i.e. normalized correlation, of regional discriminative descriptors.

$$Sim(I, I') = \sum_j \frac{D_j D_{j'}}{\|D_j\| \|D_{j'}\|} \quad (\text{Eq. 2.55})$$

This discriminative descriptor gives 4 different levels of locality:

- (i) The local binary patterns contributing to the histogram contain information at the pixel level;
- (ii) The patterns at each scale are summed over a small region to provide information at a regional level;
- (iii) The regional histograms at different scales are concatenated to produce multiresolution information;

- (iv) The global description of face is established by concatenating the regional discriminative facial descriptors.

However, a multi-scale representation by itself contains no explicit information about what image structures should be regarded as significant or what scales are appropriate for treating those. Hence, unless early judgments can be made about what image structures should be regarded as important, a substantial expansion of the amount of data is obtained to be interpreted by later stages. Thus, it is essential to complement traditional multi-scale process by explicit mechanisms for automatic scale selection, which is an intrinsic property of the SIFT algorithm as introduced in Section 2.2.5.3. Some initial attempts apply the SIFT algorithm in the face recognition task. Bicego et. al. [114] are the first to explore the SIFT approach for face verification. They investigated three matching methodologies which are more related to face images than Lowe's. The first is called the "minimum pair distance": computing the distance between all pairs of keypoint descriptors in the two images and the minimum distance is used as the matching score. This simple scheme didn't employ the location information of the keypoints. The second method is to match eyes and mouth. The authors assume the locations of the eyes and mouth are known, and only compare the SIFT features belonging to such image areas, neglecting other keypoints. The third scheme is named as "matching on a regular grid", which is applied on the well-aligned face images. The images are divided into different sub-images using a regular grid with overlapping. The matching between two images is then performed by computing distances between all pairs of corresponding sub-images, and finally averaging them. The final matching performances do not match the best face classifier tested on the BANCA database [128], due to the neglecting of the location information of keypoints.

In [115], Lin and Tang developed a multilayer framework for high resolution face recognition with information in multiple scales utilized. In the framework, each face is

decomposed into 4 layers: global appearance, facial organs (eyes, nose and mouth), skins (forehead skin and cheek skin), and irregular details. They proposed different feature description schemes for these layers, specially tailored to their different characteristics:

- Global appearance & facial organs: They first adopt multilevel PCA for dimension reduction. Then Regularized LDA is proposed to transform the features to a discriminative subspace.
- Skins: They propose the Discriminative Multiscale Texton features. Firstly a common skin texton dictionary is established, based on which, they describe the texture by texton distribution with soft histogram. R-LDA ensues to reduce the irrelevant variations.
- Irregular details: They propose SIFT-Activated Pictorial Structure, which unifies SIFT for detecting and describing local interest regions and an elastic graph for modeling their spatial configuration.

At last, an adaptive metric fusion algorithm is developed to combine the features in all layers. The performances on high resolution face images (720×576 or 1024×768) are promising.

In [116], [117], and [118], a graph is built on SIFT features. The verification problem is modeled as a graph matching process. In [119], the authors propose a method based on SIFT and Support Vector Machine. Fernandez and Vicente [120] combine Harris-Laplace and Difference-of-Gaussian detectors to detect both corner and blob structures in face images and use SIFT descriptors to represent them. They use both the Harris-Laplace detector and the DoG detector to obtain as many interest points as possible in each image. Then three measurements are performed to find the best match. The first is the distance measure. For each descriptor found in the first image, they compute the

ratio between the best and the second best matches for the second image, and count the number of matches whose ratio falls over a fixed threshold. Such a number is the first measure of distance between images. As the second measure of distance between images, they consider the number of matching points that are coherent in terms of scale and orientation. A Generalized Hough Transform is used to select the most coherent matches. The number of matches fulfilling this condition is the second measure of image similarity. Finally, an extra filtering is performed to discard possible false matches. By means of a second Generalized Hough Transform, only those matching points whose relative location is coherent are kept. The number of matching descriptors fulfilling this requirement is the third measure of image similarity. They further establish fixed weights for each measure to obtain a single similarity measure combining all the information.

$$S^{a-b} = M_D^{a-b} + 10 \cdot M_{SO}^{a-b} + 100 \cdot M_{RL}^{a-b}, \quad (\text{Eq. 2.56})$$

where S^{a-b} is the similarity between images a and b , M_D being the total number of matches between descriptors, M_{SO} being the matches fulfilling scale and orientation coherence, and M_{RL} being the matches fulfilling relative orientation coherence. The experimental results in the paper showed that a model-based approach using SIFT descriptors and two interest point detectors (Harris-Laplace and DoG) is valid for face recognition.

Wright and Hua [125] extract multi-scale local features from a dense set of multi-scale image patches. Dense features guarantee that most features in the test image have an (approximate) match in the corresponding gallery image, without having to rely on keypoint detection. They form a Gaussian pyramid of images (properly rectified and illumination-normalized) of size 32×32 , 31×31 , and 30×30 . Within each of these images, they compute feature descriptors at intervals of two pixels. The descriptors are computed from 8×8 patches, upsampled to 64×64 . They compute a $D = 400$ -dimensional feature descriptor $f \in \mathbb{R}_D$ for each patch, which is proposed in [129]. To incorporate loose

spatial information into the subsequent feature quantization process, they concatenate the pixel coordinates of the center of the patch into its feature descriptor, for a combined feature dimension of 402. The total number of feature vectors extracted from each image is 457. These features are then fed into a quantizer based on a set of randomized decision trees. The final representation of the face is just a sparse histogram of the quantized features. An IDF-weighted ℓ^1 norm is adopted as the final distance metric for the task of recognition. This method performs competitively with existing linear projection approaches.

Experimental results reported in these papers show that SIFT approach can detect and describe distinct local features from face images effectively.

2.3 Databases

Usually, after an algorithm for face recognition is proposed, people will test it on some public image databases [130] to show its performances. Some of the most popular face databases are as follows. The Facial Recognition Technology (FERET) database [131]² contains 1,564 sets of images for a total of 14,126 images that includes 1,199 individuals and 365 duplicate sets of images. AT&T “The Database of Faces” (formerly “The ORL Database of Faces”) [132]³, consists of ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The Georgia Tech Face Database [133]⁴ contains images of 50 people and is stored in JPEG format. Most of the images were taken in two different sessions to take into account the variations in

²Available at: <http://www.nist.gov/humanid/feret>

³Available at: <http://www.cl.cam.ac.uk/research/dtg/arrarchive/facedatabase.html>

⁴Available at: http://www.anefian.com/face_reco.htm

illumination conditions, facial expression and appearance. In addition to this, the faces were captured at different scales and orientations. The AR Face Database [134]⁵ contains 4,000 color images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions, illumination conditions and occlusions (sun glasses and scarf).

2.4 Summary

In this chapter I review the holistic approaches and local feature/component based approaches for face recognition applications. Although the holistic approaches are robust and efficient, they require a preprocessing procedure to normalize the face image variations in pose, scale and illumination. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. However, the recognition performance deteriorates considerably if the manual process is replaced by an automatic landmark detection algorithm. As a result, most approaches work on the normalized face images based on the manually identified landmarks, which makes these approaches semi-automatic. Moreover, global features are sensitive to variations in facial expressions, poses and occlusions. Another intrinsic problem of all holistic approaches is their dependence to the training databases because knowledge about the face discrimination is generalized by machine learning from the face samples. A representative training database is necessary, which, however, is not available in many applications.

In contrast to holistic methods, local feature/component based approaches have the potential of more robust to variations in pose, scale, expression and occlusion [11, 135]. EBGM [58], AAM [65] and SIFT fall into this category. However, the performances of both EBGM and AAM depend on a good selection of facial landmarks, which are often annotated manually. This makes the two approaches semi-automatic and labor

⁵Available at: http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html

consuming. The SIFT approach can detect distinct local structures from images and select appropriate scales to describe them automatically. It overcomes the limitations of holistic approaches and some local feature based approaches. However, there are still many outstanding issues and problems that need to be addressed and solved if I am to leverage the idea of SIFT and some of its good properties to solve the challenging face recognition problem.

Chapter 3

Multi-scale Local Feature Extraction

3.1 Introduction

One of the very fundamental problems arising when analyzing face images originates from the fact that face structures appear in different ways, depending upon the scales of observations. First, facial local structures are shown at different levels of scales, ranging from skin textures at fine scales, through eyes and mouth represented at median scales, to the shape of face contours at large scales. Second, the characteristic or the description of a local structure is strongly dependent on the scale at which the structure is modeled. Third, it is unknown in advance what the proper scales are to describe different local structures of unknown face images. A multi-scale representation by itself contains no explicit information about what image structures should be regarded as significant or what scales are appropriate for treating those. Hence, unless early judgments can be made about what image structures should be regarded as important, a substantial expansion of the amount of data is obtained to be interpreted by later stage processes. Thus, it is essential to complement traditional multi-scale processing by explicit mechanisms for automatic scale selection. In [80], the author studied the problem of identifying an appropriate and consistent scale for feature detection in depth. He showed that the normalization of the Laplacian with the factor σ^2 is required for true scale invariance. Lowe [10, 9] made use of this results and proposed a method describing image features

that have many properties that make them suitable for matching different images of an object or scene. The features are less vulnerable to change in illumination and 3D camera viewpoint. They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. Large number of features can be extracted from typical images with efficient algorithms. In addition, the features are highly distinctive, which allows a single feature to be correctly matched with high probability against a large database of features, providing a basis for object and scene recognition. Following are the major stages of computation used to generate the set of image features:

- (i) Scale-space extrema detection: The first stage of computation searches over all scales and image locations. It is implemented efficiently by using a Difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation.
- (ii) Keypoint localization: At each candidate location, a detailed model is fit to determine location and scale. Keypoints are selected based on measures of their stability.
- (iii) Orientation assignment: One or more orientations are assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.
- (iv) Keypoint descriptor: The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

This approach has been named the Scale Invariant Feature Transform (SIFT), as it transforms image data into scale-invariant coordinates relative to local feature. There are several advantages for us to implement this approach to face recognition:

- This approach detects blob type features which refer to points and/or regions in the image that are either brighter or darker than the surrounding. One merit of blob features is that they can be used as main primitives for texture analysis and texture recognition. Facial texture information can be captured by blob detection, which is never detected by corner or edge detectors.
- This approach chooses the most stable and distinct points in the scale space, indicating the optimal scale at which different blob structures should be represented, while corners or edges are not describable in scale space. As mentioned above, facial components/structures appear in different ways depending upon the scale of observation, so, it is very important to select the optimal scale which is close to the inherent scale of different components automatically in face image representations.
- The features are invariant to image scale and orientation, so I do not need to re-scale images to the same size or do alignment before recognition, which avoids errors occurring in those preprocessing steps.
- For image matching and recognition, SIFT features are first extracted from a set of reference images and stored in a database. A new image is matched by individually comparing each feature from the new image to this previous database and finding candidate matching features based on some distance measurements of their feature vectors. There is no training process in this approach, which means one reference image per person is enough to do the matching, so, the generalization problem in holistic methods is avoided.

- Mikolajczyk and Schmid [103] presented the experimental evaluation of interest region descriptors (SIFT, GLOH, shape context, PCA-SIFT, spin images, steerable filters, differential invariants, complex filters, moment invariants, and cross-correlation of sampled pixel values) in the presence of real geometric and photometric transformations. The goal was to compare descriptors computed on regions extracted with recently proposed scale and affine-invariant detection techniques. The evaluation was designed for matching and recognition of the same object or scene. In most of the tests, GLOH obtains the best results, closely followed by SIFT. This shows the robustness and the distinctive character of the region-based SIFT descriptor. The ranking of the descriptors is similar for different matching strategies.

It has also shown good performance on object detection and some other machine vision applications [112], [113], [96], [102], [103]. Recently, some initial attempts apply the SIFT algorithm in the face recognition task. In [114] an overlapping sub-image matching strategy is used as the first attempt to explore the SIFT approach for face recognition. In [115], the SIFT descriptor is adopted to describe irregular local marks detected by a Hessian-Laplace detector. In [116], [117], and [118], a graph is built on SIFT features. The recognition problem is modeled as a graph matching process. In [119], the authors propose a method based on SIFT and Support Vector Machine. Fernandez and Vicente [120] combine Harris-Laplace and Difference-of-Gaussian detectors to detect both corner and blob structures in face images and use SIFT descriptors to represent them. In [121], [122] and [123], salient regions are firstly identified from face images, and then SIFT features are extracted in each region. A modified keypoint descriptor and a redundant keypoint removal scheme are proposed for face recognition in [136, 137]. Majumdar and Ward [124] rank the SIFT features according to their discriminative power and use the most discriminating ones for face recognition.

Despite all efforts above, there are still many outstanding issues and problems that need to be addressed and solved if I am to leverage the idea of SIFT and some of its good properties to solve the challenging face recognition problem. In this chapter, I analyze the merits and deficiencies of SIFT and propose new strategies for feature extraction, which lead to a new framework that overcomes some limitations of SIFT in solving the face recognition problem. I propose a new approach to keypoint detection which can capture the information of many facial structures in the smooth area such as forehead, cheeks and chin. A partial descriptor is designed to represent the keypoints whose support areas exceed the face image. My proposed detection approach and my partial descriptor strategy produce a rich number of keypoints. As a significant keypoint should be distinct from others in terms of either its location or the image structures of its neighborhood, I further propose to remove keypoints based on their distinctiveness.

3.2 Keypoint Detection and Scale Selection

A local feature is represented by a keypoint that locates it and a descriptor that represents its intensity variations in a local support area. The scale of a local feature or of a keypoint determines the size of the support area. Thus, the keypoint detection and its scale determination are the most critical parts of finding the image local features.

A keypoint is a pixel or the center of a local area that shows some characteristics different from its neighbors. Obviously, blob-like and corner-like structures are candidates of keypoints. While a blob-like feature structure shows extreme gray value at its center in a properly smoothed image and hence can be easily detected, a corner-like structure does not. A Laplace operator ∇^2 applied to the image $I(x, y)$ produces extrema at both blob-like and corner-like structures. Therefore, the spatial extrema of the Laplace image $\nabla^2 I(x, y)$ are keypoint candidates. However, local structures have different scales so that the extrema appear in the images smoothed with different scales. To find the scale of

a possible keypoint I need to detect the extrema in the scale space, too. As studied by Lindeberg [80], the normalized Laplacian of Gaussian (LoG), $\sigma^2 \nabla^2 G$, provides scale invariance. Thus, automatic scale selection is done in the output images of the normalized LoG filter

$$O(x, y, \sigma^2) = \sigma^2 \nabla^2 G(x, y, \sigma) * I(x, y), \quad (\text{Eq. 3.1})$$

where $G(x, y, \sigma)$ is a Gaussian smooth filter with zero mean and diagonal covariance matrix $\sigma^2 \mathbf{I}$. As it is computationally intensive to search the extrema in the scale space σ , Lowe proposed to use Difference-of-Gaussian (DoG) of nearby scales at $k\sigma$ and σ to approximate the normalized Laplacian-of-Gaussian

$$\sigma^2 \nabla^2 G(x, y, \sigma) \approx \frac{1}{k-1} [G(x, y, k\sigma) - G(x, y, \sigma)], \quad (\text{Eq. 3.2})$$

where $k = \sqrt[3]{2}$ is selected in this work.

To detect the blob-like and corner-like structures and represent them at the optimal scales, points of the DoG images that are extrema in both spatial and scale spaces are selected. Lowe [10, 9] in his SIFT framework proposes to compare each sample point to its eight neighbors in the current DoG image (spatial space) and nine neighbors in the scale above and below (scale space). A SIFT keypoint is selected only if it is larger than all these 26 neighbor points or smaller than all of them. This keypoint detection method works well for rigid visual objects, which have sharp transitions between different sides of an object. In other words, there are distinct corner structures with high contrast in such objects. However, human faces are non-rigid, round and smooth. There are few obvious blobs and corners with high contrast, because the intensity changes in face images are gradual and slow in the most areas. On the other hand, the shape of the structures could be complex and some structures are close to each other or overlap. As a result, many local structures in the smooth area such as forehead, cheeks and chin cannot be detected due to the strict condition of the extrema in the 26 neighbors.

 (a) Input signal I

(b) DoG outputs at six different scales

(c) Local minima of DoG outputs

(d) Local minima of DoG outputs

Figure 3.1: Problem of the SIFT keypoint detection.

To show an example of how keypoints are missing in the detection process proposed by Lowe, I plot a one-dimensional signal I as in Fig.3.1(a), which is the sum of two Gaussian structures at scales $\sigma_i^1 = 3.0$ and $\sigma_i^2 = 4.1$, respectively. At least two keypoints near the two scales should be detected in the vicinity of the two Gaussian peaks. Fig.3.1(b) shows the DoG outputs at six successive scales in the SIFT framework that well cover the two scales of the signal I : $\sigma_1 = 1.6 \times 2^{1/3} \approx 2.02$, $\sigma_2 = 1.6 \times 2^{2/3} \approx 2.54$, $\sigma_3 = 1.6 \times 2^{3/3} = 3.20$, $\sigma_4 = 1.6 \times 2^{4/3} \approx 4.03$, $\sigma_5 = 1.6 \times 2^{5/3} \approx 5.08$ and $\sigma_6 = 1.6 \times 2^{6/3} = 6.40$. The DoG outputs in the vicinity of the two Gaussian peaks are marked and the details are shown in Fig.3.1(c) and 3.1(d). From Fig.3.1(c), I see that the sample point 1 (in red) at scale

Figure 3.2: Extrema of the DoG images are detected by comparing a pixel (marked with cross) to its 8 neighbors at the current scale and the corresponding pixels at the adjacent scales (marked with circle).

$\sigma_3 = 3.20$ is the local minimum at the current scale compared with the points 2 and 3. Although it is smaller than the three neighbors in blue (sample points 1', 2' and 3') in the scale below, it is not smaller than the neighbor point 2'' in the scale above (in black). Hence, the sample point 1 won't be detected as a local minimum. Similarly, in Fig.3.1(d) no keypoint is detected by Lowe's method. In a word, the original detection approach cannot detect any keypoint at the tested six scales in the vicinity of the two Gaussian peaks correctly. The missing detection is even worse for the complex two-dimensional structures. This decreases the number of keypoints detected and also deteriorate the following matching performance.

To fully detect the keypoints with distinct structures, I propose to compare a candidate point with its eight neighbors in the current scale and the corresponding one neighbor in the scale above and below. A keypoint is selected if it is larger than all of these neighbors or smaller than all of them. In other words, I compare a candidate point only to its 10 neighbors rather than 26 neighbors. Fig.3.2 visually shows the proposed approach. Obviously, keypoints detected by Lowe's approach form a subset of the keypoints detected by my approach. I can see from Fig.3.1 that the proposed approach

Figure 3.3: (a) Initial keypoints detected by the original method (in blue/big dots) and extra keypoints detected by the proposed approach (in red/small dots). (b) Remaining blue keypoints after keypoint removal by low contrast (as described in Section 2.2.5.3); (c) Remaining blue keypoints after further keypoint removal based on the ratio of principal curvatures (as described in Section 2.2.5.3).

successfully detects two keypoints (the sample point 1 in Fig.3.1(c) and the sample point 1 in Fig.3.1(d)) of signal I in the vicinity of the two Gaussian peaks.

Another detection example is shown in Fig.3.3, where the blue points denote the ones detected by Lowe’s approach and the red points denote some extra keypoints detected by the proposed approach. As I see, Lowe’s approach detects few keypoints from the forehead, cheeks, chin and facial contour which still provide useful information in telling different people although the change of these areas are subtle, while my approach can successfully detect these keypoints.

3.3 Descriptor Representation

There are several descriptors proposed in the literature to represent the local image structures [103]. Some efforts [138, 139] were made to produce discriminative and low-dimensional descriptors learned from large databases. In this work, I adopt Lowe’s descriptor, which is a set of histograms consisting of oriented gradients. In Lowe’s SIFT framework, the support area proportional to the scale of the keypoint is divided into 4×4 blocks. An 8-bin oriented gradient histogram is computed in each block. Thus,

a histogram vector $\mathbf{h}$ for each keypoint has $4 \times 4 \times 8 = 128$ dimensions. To reduce the effect of illumination changes, the histogram vector $\mathbf{h}$ is normalized to the unit length and serves as the descriptor of the keypoint.

Figure 3.4: An illustration of partial-descriptor: blocks marked with cross are considered as invalid.

In the application of face recognition, I aim to distinguish different face images. Keypoints near the face edge carry important information about the shape of the facial contour. However, these keypoints that represent the shapes of the facial contour are often of large scale and hence their support areas exceed the image area, if the image is cropped tightly to the face size. To make use of these important keypoints, I introduce a mask vector $\mathbf{m} = (m_1, m_2, \dots, m_{128})^T$ for each keypoint defined as

$$m_k = \begin{cases} 1, & \text{if the block is in image region;} \\ 0, & \text{if the block is outside image region.} \end{cases} \quad (\text{Eq. 3.3})$$

As face recognition is always followed by a face detection process, the rough position and size of a face in the image are known, thus I can have the mask vector $\mathbf{m}$, which enables a partial descriptor. The feature vector $\mathbf{f}$ is computed by

$$\mathbf{f} = \mathbf{M}\mathbf{h}, \quad (\text{Eq. 3.4})$$

where $\mathbf{M}$ is a diagonal matrix whose diagonal elements are the elements of the mask vector $\mathbf{m}$. To invalid the contribution of the possible non-image region out of the descriptor

matching, the similarity between two descriptors i and j is computed as

$$s_{ij} = \frac{\mathbf{f}_i^T \mathbf{f}_j}{\sqrt{\mathbf{f}_i^T \mathbf{M}_j \mathbf{f}_i \mathbf{f}_j^T \mathbf{M}_i \mathbf{f}_j}}. \quad (\text{Eq. 3.5})$$

It is easy to see that s_{ij} is a normalized similarity between two keypoints i and j in the common image region. It enables the participation of partial descriptors in the matching process. Fig. 3.4 gives an illustration of the proposed descriptor.

3.4 Insignificant Keypoint Removal

In this section, I propose two methods to remove unreliable keypoints. Firstly, I introduce a method to remove insignificant keypoints based on their volume information instead of their contrast. Secondly, to cater for the approaches introduced in the above two sections, I propose to remove insignificant keypoints based on their distinctiveness from others.

3.4.1 Volume-SIFT Approach

In the SIFT algorithm, scale-normalized Laplacian operators are used to select the scale for each blob structure. Consider an idealized model pattern: A symmetric Gaussian function

$$I(x, y) = G(x, y, \sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x^2+y^2)/(2\sigma_1^2)} \quad (\text{Eq. 3.6})$$

is a model of a two-dimensional blob with characteristic length σ_1 in either coordinate direction as the input image $I(x, y)$. The scale-space function is produced from the convolution of a variable-scale Gaussian $G(x, y, \sigma^2)$ with the input image $I(x, y)$:

$$L(x, y, \sigma^2) = G(x, y, \sigma^2) * I(x, y). \quad (\text{Eq. 3.7})$$

From the semigroup property of the Gaussian $G(x, y, \sigma_A^2) * G(x, y, \sigma_B^2) = G(x, y, \sigma_A^2 + \sigma_B^2)$ [140] and Eq. 3.7, it follows that the scale-space representation of I is:

$$L(x, y, \sigma^2) = G(x, y, \sigma_1^2 + \sigma^2). \quad (\text{Eq. 3.8})$$

Using Eq. 3.1, it can be shown that

$$\begin{aligned}
 O(x, y, \sigma^2) &= \sigma^2 \nabla^2 G(x, y, \sigma^2) * I(x, y) & (\text{Eq. 3.9}) \\
 &= \sigma^2 \left(\frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_1^2)}} \right)^2 \nabla^2 e^{-(x^2+y^2)/2(\sigma^2+\sigma_1^2)} \\
 &= \frac{\sigma^2}{2\pi(\sigma^2 + \sigma_1^2)} \nabla \left[-\frac{x}{\sigma_1^2 + \sigma^2} e^{-(x^2+y^2)/2(\sigma^2+\sigma_1^2)} - \frac{y}{\sigma_1^2 + \sigma^2} e^{-(x^2+y^2)/2(\sigma^2+\sigma_1^2)} \right] \\
 &= \frac{\sigma^2}{2\pi(\sigma^2 + \sigma_1^2)} \cdot \frac{(x+y)^2 - 2(\sigma^2 + \sigma_1^2)}{(\sigma^2 + \sigma_1^2)^2} e^{-(x^2+y^2)/2(\sigma^2+\sigma_1^2)}.
 \end{aligned}$$

For any $\sigma > 0$, there is a unique maximum over scales in

$$|O(0, 0, \sigma^2)| = \frac{\sigma^2}{\pi(\sigma_1^2 + \sigma^2)^2}. \quad (\text{Eq. 3.10})$$

And this maximum over scales is given by

$$\partial_\sigma(\sigma^2 \nabla^2 L(0, 0, \sigma^2)) = 0 \Leftrightarrow \sigma = \sigma_1 \quad (\text{Eq. 3.11})$$

Hence, the scale at which the scale-space maximum is assumed reflects a characteristic size of the blob. And this is the rationale behind the scale-space extrema detection of SIFT algorithm. Laplacian, being a second-order derivative operator, is very sensitive to noise. Thus, Lowe [9] proposed to remove keypoints based the contrast, as described in Section 2.2.5.3. Firstly, the value of $|O(x, y, \sigma^2)|$ at the location of each candidate keypoint is evaluated. If this value is below some threshold, the keypoint is removed. Consider two Gaussian blob structures $I_1(x, y, \sigma_1^2)$ and $I_2(x, y, \sigma_2^2)$ with different scales $\sigma_1 = 10$ and $\sigma_2 = 50$, which are shown intuitively in Fig.3.5. After convolution with scale-normalized Laplacian $\sigma^2 \nabla^2 G$, the one-dimensional outputs of I_1 and I_2 are shown in Fig.3.6. And the keypoints representing the both blobs are detected at the extrema of O_1 and O_2 . In the original SIFT algorithm, if the value of $|O(x, y, \sigma^2)|$ at the position of each candidate keypoint is below some threshold, the keypoint is removed because it represents structure with low contrast (and therefore is more sensitive to noise). From

Figure 3.5: Symmetric Gaussian blobs with different scales and their corresponding one dimensional illustrations.

Fig.3.6, I can see that the extrema of $|O_1(x, y, \sigma_1^2)|$ and $|O_2(x, y, \sigma_2^2)|$ are the same. Then, in the original SIFT algorithm, structures I_1 and I_2 are treated equally. They are either both removed or both kept. However, intuitively, from Fig.3.5, I can see that I_1 and I_2 are two significantly different structures with different scales. And they should not be considered equally useful nor equally useless. Moreover, the scale of I_2 is larger than that of I_1 , and therefore less sensitive to noise than I_1 . I should keep keypoints representing I_2 and remove keypoints representing I_1 . Thus, I propose Volume-SIFT (VSIFT) approach, which eliminates unreliable keypoints not based on the contrast (value of $|O(x, y, \sigma^2)|$ at the location of each candidate keypoint) but based on the volume of the structure:

$$V(x, y, \sigma^2) = \sigma^2 |O(x, y, \sigma^2)| = \sigma^4 |\nabla^2 G * I(x, y)|, \quad (\text{Eq. 3.12})$$

Figure 3.6: One-dimensional outputs of I_1 and I_2 .

where σ is the scale of the keypoint. From Equ. Eq. 3.12, I can derive

$$V_1(x, y, \sigma_1^2) = \sigma_1^2 |O_1(x, y, \sigma_1^2)|, \quad (\text{Eq. 3.13})$$

and

$$V_2(x, y, \sigma_2^2) = \sigma_2^2 |O_2(x, y, \sigma_2^2)|. \quad (\text{Eq. 3.14})$$

Though the extrema of $|O_1(x, y, \sigma_1^2)|$ and $|O_2(x, y, \sigma_2^2)|$ are the same, the extrema of $V_1(x, y, \sigma_1^2)$ and $V_2(x, y, \sigma_2^2)$ are different. And in this case,

$$V_1(x, y, \sigma_1^2)_{\text{extrema}} < V_2(x, y, \sigma_2^2)_{\text{extrema}}. \quad (\text{Eq. 3.15})$$

I evaluate the value of $V(x, y, \sigma^2)$ at the location of each candidate keypoint, if this value is below some threshold, the keypoint is removed.

3.4.2 Eliminating Similar Keypoints

As introduced in Section 2.2.5, scale-space extrema detection is performed by convolution an input image with scale-normalized Laplacian of Gaussian operators. Laplacian, being a second-order derivative operator, is very sensitive to noise. And it detects a lot of keypoints along edges. However, keypoints along edges are very unstable, because their locations are very sensitive to small changes of neighboring texture. In Lowe's SIFT

framework [9], keypoints with low contrast and high edge responses are removed, as described in Section 2.2.5.3. However, it is known that SIFT was designed for general object detection/recognition. General objects are rigid, and there are sharp transitions between different sides of an object. In other words, there are distinct structures with high contrast in general objects. Moreover, corners or straight lines are normal structures in general objects, which have high edge responses. Based on these characteristics of general objects, keypoints with low contrast (sensitive to noise) or along edges (unstable) should be removed. However, different from general objects, faces are non-rigid and smooth. There are few straight edges in face images. The intensity changes in face images are gradual and slow and hence the blob and corner structures are not significantly different from their neighboring pixels. Therefore, the keypoint removal scheme in the SIFT framework is inappropriate for face images. For instance, the blue points in Fig. 3.3a show the initially detected keypoints by Lowe's approach. Fig. 3.3b shows the remaining blue keypoints after applying a threshold on the minimum contrast of candidate keypoints. Fig. 3.3c shows the final remaining blue keypoints after further removing keypoints with high edge responses. The keypoints marked with ellipse and rectangle are removed, which do represent distinctive structures such as wrinkles and mouth corners.

To further prove my conclusions, I detect keypoints in a set of artificially generated images to approximate blob structures as shown in Fig.3.7. The structure in each image on the top left corner is 2D Gaussian with equal σ on either dimension. The structure on the top right corner is 2D Gaussian with different σ on either dimension. And the structures in the second row are truncated 2D Gaussian correspondingly. Fig.3.7(a) gives the location of original keypoints detected in the clean image, and Fig.3.7(b) gives the location of original keypoints detected in the image superposed by $\pm 4\%$ random noises. From Fig.3.7(a) and Fig.3.7(b), I can see that a lot of extra keypoints are detected in Fig.3.7(b) which are induced by noise. Fig.3.7(a') shows the keypoints remain

Figure 3.7: Keypoints detected by SIFT in artificially produced images: (a) Initial keypoints detected in the clean image; (b) Initial keypoints detected in the image imposed with $\pm 4\%$ random noises; (a') Keypoints remain in (a) after applying a threshold on the minimum contrast; (b') Keypoints remain in (b) after applying a threshold on the minimum contrast; (a'') The final keypoints remain in (a') after further applying a threshold on the ratio of principal curvatures; (b'') The final keypoints remain in (b') after further applying a threshold on the ratio of principal curvatures.

in Fig.3.7(a) after removal keypoints with low contrast. Keypoints marked with rectangle in Fig.3.7(a) are removed in Fig.3.7(a') due to low contrast. However, these keypoints are not induced by noise, therefore they should be kept. Fig.3.7(b') shows the location of keypoints remain in Fig.3.7(b) after applying a threshold on the minimum contrast. Keypoints in Fig.3.7(b') marked with ellipse are induced by noise, and they should be removed but remain. Fig.3.7(a'') gives the final location of keypoints remain in Fig.3.7(a') after further removal keypoints with high edge responses. Keypoints marked with rectangle in Fig.3.7(a') are eliminated in Fig.3.7(a''). But they represent structures as elongated ellipses, therefore they should be kept. Fig.3.7(b'') shows the final keypoints remain in Fig.3.7(b') after further applying a threshold on the ratio of principal curvatures. Keypoints marked with rectangle in Fig.3.7(b') represent useful structures as elongated ellipses, but they are removed in Fig.3.7(b''). Keypoints marked with ellipse in Fig.3.7(b'') are induced by noise. However, they still remain after the two-step removal process. From this analysis, I conclude that the algorithm to remove unreliable keypoints in SIFT approach is not reliable.

As the original insignificant keypoint removal approach is not appropriate for face images, I propose to remove keypoints based on their distinctiveness from others. A significant keypoint should be distinct from others in terms of either its location or the image structures of its neighborhood. Thus, an initially detected keypoint is removed if and only if the spatial Euclidean distances from it to any other keypoint is smaller than a threshold t_e and the similarity between their descriptors is higher than a threshold t_c . This process removes insignificant keypoints and hence retains distinctive ones.

3.5 Experimental Results

In this section, I validate each ingredient of the proposed feature extraction, feature description and insignificant keypoint removal framework by comparing it with the coun-

terpart of the SIFT framework. Firstly, I evaluate the repeatability and distinctiveness of the keypoints detected by my proposed approach following the protocols in [73]. Then I compare the performances of the local feature based approaches in face recognition with the matching procedure in [9]. I further add three representative holistic approaches: Fisherface (LDA) [19], the unified framework of subspace methods (UFS) [28] and the most successful holistic approach Eigenfeature Regularization and Extraction (ERE) [36, 37], and some multi-scale local feature-based approaches: Elastic Bunch Graph Matching (EBGM) [58, 71], Multi-scale Gabor feature (MG) [127], Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [70], Multi-scale Local Binary Pattern Histogram (MLBPH) [71], multiple keypoint detectors with SIFT descriptors (MKDS) [120], and dense multi-scale features (DMF) [125] into comparison to show the advantages of my proposed local feature based approaches.

Figure 3.8: (a) Sample images used in the holistic approaches; (b) Sample images used in the multi-scale local feature based approaches (SIFT and my proposed frameworks).

While images are preprocessed and normalized for the holistic approaches following the CSU Face Identification Evaluation System [141] with manually detected two eye coordinates, this preprocessing and normalization procedure is not applied for the multi-scale local feature based approaches. Fig.3.8 shows some image samples used in the holistic methods and image samples used in the SIFT and my proposed frameworks.

3.5.1 Validation of the Keypoint Detector

The images used to evaluate the performances of various detectors are shown in Fig. 3.9¹. In this data set, image changes in viewpoint, scale, image blur, JPEG compression and illumination are evaluated. The experimental setting is the same as that in [73]. Please refer to [73, 142] for the data description.

Two parameters decide the performance of a local feature detector: The repeatability and the distinctiveness. To measure the repeatability of a detector, the average number of repeated regions detected in images under different transformations are recorded. For the distinctiveness, the accuracy of region estimation and localization of a detector is measured.

To evaluate the accuracy of the detectors, the SIFT descriptor is adopted to match the corresponding regions. The matching score is computed between a reference image and the remaining images in a set in two steps:

- (i) A match is assumed to be correct if the overlap error is minimal and less than 40%.
- (ii) The matching score is computed as the ratio between the number of correct matches and the smaller number of detected regions in the pair of images. A match is the nearest neighbor (smallest Euclidean distance) in the descriptor space.

Fig. 3.10 to 3.17 give the evaluation results of different keypoint detectors under various image conditions. My proposed keypoint detection scheme achieves similar performances under all these changes. For the evaluation of the repeatability, the absolute number of correspondences obtained by my detection scheme is the largest among different keypoint detectors when the geometric or photometric transformations are not obvious. Similarly, for the distinctiveness, my detection scheme achieves the largest number of correct matches when the image changes are subtle. The relative values (repeatability%

¹Available at: <http://www.robots.ox.ac.uk/vgg/research/affine/index.html>

Figure 3.9: Database. (a) Viewpoint change, (b) Viewpoint change, (c) Zoom+rotation, (d) Zoom+rotation, (e) Image blur, (f) Image blur, (g) JPEG compression, (h) Light change. The left most image of each set is used as the reference image in the experimental comparisons.

Figure 3.10: Performance evaluation of various keypoint detectors (Harris-Affine (haraff) [4], Hessian-Affine (hesaff) [4], maximally stable extremal regions (mseraf) [5], intensity-based regions (ibraff) [6, 7], edge-based detectors (ebraff) [8, 7], Difference-of-Gaussians (DoG) [9, 10], and the proposed) under viewpoint change. (a) Repeatability score (overlap 40%, normalized size = 30 pixels). (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.11: Performance evaluation of various keypoint detectors under viewpoint change for the textured scene. (a) Repeatability score. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.12: Performance evaluation of various keypoint detectors under scale change for the structured scene. (a) Repeatability score for scale change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.13: Performance evaluation of various keypoint detectors under scale change. (a) Repeatability score for scale change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.14: Performance evaluation of various keypoint detectors under blur change for the structured scene. (a) Repeatability score for blur change. (b) Absolute number of corresponding regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.15: Performance evaluation of various keypoint detectors under blur change for the textured scene. (a) Repeatability score for blur change. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.16: JPEG compression. (a) Repeatability score for different JPEG compression. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

Figure 3.17: Illumination change. (a) Repeatability score. (b) Absolute number of repeated regions. (c) Matching score. (d) Absolute number of correct matches.

and matching score%) are not so high as absolute ones, because my keypoint detection approach detects a great number of keypoints, which leads to a large number of absolute matches and a low relative matching score. However, the repeatability of my detected keypoints is comparable to others. In many face recognition tasks, images are in small size, where the feature quantity (coverage) is crucial to the final performances. With large number of correspondences and correct matches, my keypoint detection scheme is more appropriate for face recognition applications.

3.5.2 Validation of Feature Detection and Description in Face Recognition

Firstly, I compare my proposed feature detection approach P_dec and feature description approach P_des introduced in Section 3.2 and Section 3.3 with the original SIFT detection approach S_dec and the feature description approach S_des on FERET, ORL, Georgia Tech (GT) and AR databases. For a fair and comprehensive comparison, I give the performances of S_dec+S_des , S_dec+P_des , P_dec+S_des and P_dec+P_des , respectively. And I do not remove any keypoint in the following experiments.

3.5.2.1 Results on FERET Database I

There are 2388 images comprising 1194 persons (two images Fa/Fb per person) selected from the FERET database. Images are cropped into the size of 65×75 pixels. In the first experiment, images of 250 people are randomly selected for training, and the remaining images of 944 people are used for testing. The best recognition performances of the holistic approaches LDA [19], UFS [28] and ERE [36] with default parameters over all possible numbers of features are recorded. As in the local feature based approaches, there are no training process, so I only use the images of 944 people for the experiments. In the second experiment, more training samples (497 people) are randomly selected, and the remaining images of 697 people are used for testing.

Table 3.1: Recognition rate on FERET database I

training / testing	MLBPH [71]	LGBPHS [70]	MG [127]	EBGM [58, 71]	LDA	
500 / 1888	-	-	-	-	89.72%	
994 / 1394	98.5%	98%	90.5%	89.8%	96.13%	
training / testing	UFS	ERE	S_dec+S_des	S_dec+P_des	P_dec+S_des	P_dec+P_des
500 / 1888	93.75%	95.23%	93.33%	95.66%	96.50%	97.67%
994 / 1394	96.70%	97.13%	95.41%	96.56%	97.56%	98.42%

Table 3.1 gives the recognition rates of the holistic approaches, multi-scale local feature-based approaches together with the SIFT feature detection and description approach (S_dec+S_des) in [9], S_dec with my feature description P_des , my feature detection with SIFT feature description (P_dec+S_des), and my feature detection and description framework (P_dec+P_des). Comparing the recognition performances of S_dec+S_des and S_dec+P_des , I can see that my feature description approach can increase the recognition rates on both experimental settings, which can also be validated by comparing the performances of P_dec+S_des and P_dec+P_des . The superiority of my feature detection approach over the original SIFT counterpart is proved by either comparing the performances of S_dec+S_des and P_dec+S_des or S_dec+P_des and P_dec+P_des , where the recognition rates increase either from 93.33% and 95.41% to 96.50% and 97.56% or from 95.66% and 96.56% to 97.67% and 98.42%. Comparing the performances with holistic approaches, my feature detection and description framework (P_dec+P_des) achieves the highest recognition rate on both experimental settings. And the proposed method performs competitively with multi-scale local feature-based approaches.

3.5.2.2 Results on FERET Database II

This database is constructed, same to one data set used in [36], by choosing 256 subjects with at least four images per subject. And I use the same number of images (four) per subject for all subjects. 512 images of the first 128 subjects are used for training, and the remaining 512 images serve as testing images. In the SIFT framework and my proposed

feature detection and description approaches, I only use the last 512 images for testing as there are no training process. The size of the image is 130×150 pixels, same as that in [36]. The i^{th} ($i = 1, 2, 3, 4$) images of all the testing subjects are chosen to form a gallery set, and the remaining three images per subject serve as the probe images to be identified from the gallery set.

Table 3.2: Average recognition rate and standard deviation on FERET Database II

	LDA	UFS	ERE	S_dec+S_des	S_dec+P_des	P_dec+S_des	P_dec+P_des
ARR	80.79%	82.23%	83.07%	81.58%	87.17%	85.68%	90.17%
Std	1.54%	1.55%	1.43%	0.89%	2.46%	2.16%	0.75%

Table 3.2 shows the average recognition rates (ARR) and corresponding standard deviations (Std) over the four probe sets, each of which has a distinct gallery set, obtained by holistic approaches LDA, UFS and ERE, together with local feature based approaches SIFT feature detection and description approach (S_dec+S_des), S_dec with my feature description P_des , my feature detection with SIFT feature description (P_dec+S_des), and my feature detection and description framework (P_dec+P_des). From this table, I can see that local feature based approaches can achieve better recognition performance than holistic approaches. For instance, the average recognition rate of my local feature detection and description approach (P_dec+P_des) can reach 90.17% with 0.75% standard deviation, which is significantly higher than the average recognition rate obtained by the ERE approach relative to its standard deviation. Comparing different local feature detection approaches, I can conclude that my feature detection approach P_dec can improve the performances from 81.58% (S_dec+S_des) to 85.68% (P_dec+S_des) and from 87.17% (S_dec+P_des) to 90.17% (P_dec+P_des), respectively. Moreover, the average recognition rates increase from 81.58% (S_dec+S_des) to 87.17% (S_dec+P_des), and from 85.68% (P_dec+S_des) to 90.17% (P_dec+P_des) due to my local feature description approach P_des . In all, the experimental results demonstrate that my local

feature detection and description approaches significantly outperform the counterparts in the original SIFT framework. My local feature detection and description approaches can capture more distinctive information in face images than the original SIFT approach.

3.5.2.3 Results on ORL, GT and AR Databases

The ORL database contains 400 images from 40 subjects taken at different times, varying the lighting, facial expressions and facial details. Images of the ORL database are resized into 50×57 pixels. The first 5 samples per subject form the training and gallery sets, and the remaining 5 samples serve as the probe images. The Georgia Tech (GT) face database contains 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray scale and resized into 60×80 pixels. Similarly, the GT database is partitioned into the training set consisting of the first 8 samples per subject and the probe set consisting of the remaining 7 samples. The gallery set is the same as the training set in the experiments. The color images in AR database are converted to gray scale and cropped into the size of 60×85 , same as the image size used in [36]. There were 75 subjects with 14 nonoccluded images per subject used in [36]. In my experiment, I follow the settings in [36]. The first 7 images of all subjects are used in the training and gallery sets, and the remaining 7 images serve as probe set. In local feature based approaches, there is no training process. Hence, I only use the gallery and probe sets.

Table 3.3 gives the recognition rates of the holistic approaches PCA, UFS and ERE, together with the local feature based approaches MKDA [120], DMF [125], S_dec+S_des , S_dec+P_des , P_dec+S_des and P_dec+P_des , on ORL, GT and AR databases. On ORL and AR databases, my local feature detection and description approach P_dec+P_des outperforms both holistic approaches and all the other local feature based approaches listed in the table. On GT database, P_dec+P_des can achieve the same recognition

Table 3.3: Recognition rates on ORL, GT and AR databases

ORL database								
MKDS [120]	DMF [125]	LDA	UFS	ERE	S_dec+S_des	S_dec+P_des	P_dec+S_des	P_dec+P_des
97.0%	96.5%	92.5%	83.5%	97.0%	90.0%	93.5%	94.0%	98.0%
GT database								
		LDA	UFS	ERE	S_dec+S_des	S_dec+P_des	P_dec+S_des	P_dec+P_des
		92.00%	91.43%	92.86%	80.57%	87.14%	87.71%	92.86%
AR database								
		LDA	UFS	ERE	S_dec+S_des	S_dec+P_des	P_dec+S_des	P_dec+P_des
		93.90%	95.05%	95.05%	93.71%	95.62%	96.19%	96.76%

rate as the holistic approach ERE. The large differences between the recognition rates of S_dec+S_des and S_dec+P_des , or P_dec+S_des and P_dec+P_des on these three databases, show the efficacy of my local feature description approach P_des compared to the original SIFT feature descriptor S_des . The superiority of my local feature detection approach P_dec over the original SIFT feature detection strategy S_dec is shown by comparing the performances of S_dec+S_des and P_dec+S_des , or S_dec+P_des and P_dec+P_des . The advantages of my local feature detection approach P_dec and local feature description strategy P_des are validated over all the three databases.

3.5.3 Validation of Insignificant Keypoint Removal

In this section, I validate two insignificant keypoint removal approaches proposed in Section 3.4 on ORL database, by comparing them with the counterpart of the original SIFT framework. Each subject in the ORL database has ten images index i from 1 to 10. In the first experiment, I adopt the same experimental settings used in Section 3.5.2.3, and I use the original SIFT keypoints. In the second experiment, the images from each subject with the same index i are picked out to form the gallery set and the remaining 360 images serve as the probe set. The face recognition error rate is computed by the number of misidentified probe images over 360. Ten runs of experiments are performed where each run has distinct gallery images. Different from the first experiment, I use my local

feature extraction and description approach to detect keypoints. For a fair comparison, in these two experiments, I adjust the thresholds in the original SIFT removal approach (SIFTR), my Volume-SIFT approach (VSIFTR) and my similar keypoints elimination approach (SIMR) to make the number of keypoints similar.

Figure 3.18: Recognition rate of different keypoint removal approaches obtained in (a) the first experimental setting (b) the second experimental setting.

Fig. 3.18(a) shows the recognition rate against the number of remaining keypoints after these three keypoint removal approaches, respectively. From this figure I can see that the keypoint removal method SIMR outperforms the SIFTR when around 20% keypoints are eliminated. Removing more keypoints, VSIFTR achieves the better performance. To show the efficacy of SIMR, more runs of experiments are performed in the second setting, where I apply my proposed keypoint extraction and description scheme. Fig. 3.18(b) gives the results. This figure demonstrates again that by SIMR, I achieve better recognition performance than SIFTR. However, the gain in recognition accuracy by the keypoint removal is insignificant. Removing the keypoints more than 20% of the initially detected reduces the recognition accuracy. Therefore, the keypoint removal should not target at improving the recognition accuracy but at reducing the computational complexity of the recognition process.

3.6 Summary

Face recognition based on the multi-scale local features has the potential to be more robust to variations in pose, scale, expression and occlusion than the holistic approaches. The SIFT framework is designed and works well for general object recognition. However, it is not optimal for the face recognition because face is a non-rigid, round and smooth object. Its intensity has slow and gradual change but the local structure is complex due to the spatial overlap. The proposed keypoint detection, partial descriptor and insignificant keypoint removal extract and retain more useful local features for the face recognition compared with the counterparts in the SIFT framework. While some basic tools such as DoG filter and HoG descriptor are inherited from the SIFT framework, this work investigates and contributes to all major steps in the feature extraction and description for the challenging face recognition task.

In the proposed local feature extraction approach, I compare one sample point to its 10 neighbors instead of 26 in the original SIFT algorithm, which leads to a large number of keypoints detected in face images. For example, 65 keypoints are detected averagely from one face image with size 60×85 by the original SIFT approach, while 191 keypoints are extracted by my approach. The average extraction time per image is 8.7295s on a laptop with an Intel(R) Core(TM) *i7-2640M* CPU @ 2.8GHz processor and 4GB RAM. The code is mainly written in matlab without any optimization.

Chapter 4

Image Matching Approach and Its Application

In this chapter, I investigate on the local feature matching approaches following the feature extraction and description processes described in the previous chapter. To determine the identity of a probe face image based on a set of gallery images, local structures of the probe image represented by the keypoints and their descriptors are compared with those in the gallery. The gallery image whose local structures have the maximum similarity to the probe image establishes the identity of the probe image. In the identification tasks, there are many similar gallery images. As a result, the nearest keypoints to the probe often disperse to many candidates in the gallery, and hence the probability that the largest number of the nearest keypoints fall into the right candidate is low. This problem becomes severe if the gallery contains a large number of subjects. Moreover, multiple templates per subject in the gallery make it even worse. In this chapter, firstly I describe our keypoint matching strategies which circumvent the problem that local structures matched with those in probe disperse into many different gallery images. Secondly, a training procedure is developed for multiple samples per subject. It contains template selection, template synthesis and unstable keypoint removal to meet different requirements in face recognition applications.

4.1 Image Matching Approach

In this section, a strategy of keypoint search for the nearest subjects and a two-stage image matching scheme are developed for the face identification task.

4.1.1 Introduction

The general steps of local feature matching are as below:

- (i) Putative matching of keypoints over images using nearest neighbor search based on the descriptor similarities.
- (ii) Detect and remove mismatches (outliers) by robust methods, which search for subsets of matches consistent with a global constraint (inliers).
- (iii) Make final decisions based on the inliers and outliers.

Although each step is important to make the final decision correctly, the second step is the most critical. Various approaches are proposed to remove mismatches in the tentative correspondences and retain the reliable ones which are consistent with some projective transformation, by estimating transformation parameters. An early and most popular example of a robust estimation algorithm is the random sample consensus (RANSAC) introduced by Fishler and Bolles in 1981 [143], which has become a standard choice for outlier removal in many computer vision applications. RANSAC is a hypothesize-and-verify approach, which generates candidate solutions by using the minimum number observations (data points) required to estimate the underlying model parameters. Unlike conventional sampling techniques that use as much of the data as possible to obtain an initial solution and then proceed to prune outliers, RANSAC uses the smallest set possible and proceeds to enlarge this set with consistent data points. Recently, a variety of modifications to the basic RANSAC algorithm have been proposed to improve its

efficiency. Torr and Zisserman [144] established the Maximum Likelihood Estimation Sample Consensus (MLESC) as a generalization of the RANSAC method to estimate image geometry. It uses the same sampling strategy to generate putative solutions, but maximizes the likelihood rather than the number of inliers. Torr [145] extended MELSAC to the Bayesian case with Maximum A Posteriori Sample Consensus (MAPSAC) by replacing the likelihood with posterior probability. Tordoff and Murray [146] proposed a Guided-MLESAC estimation by adding prior information. They achieved faster image transform estimations by use of prior probabilities from the matching process. Chum and Matas [147] demonstrate the Progressive Sample Consensus (PROSAC) on the wide-baseline matching problem. Instead of random guessing, they exploit the linear ordering defined on the set of correspondences by a similarity function to achieve computational savings.

However, when the percent of inliers falls much below 50%, the performances of RANSAC based approaches tend to be poor [9]. Lowe in [9] proposed to cluster the putative keypoint matching pairs with the Generalized Hough Transform, as the parameters: 2D location, scale and orientation of each keypoint are known. The image matching algorithm in Lowe's framework contains two stages. First, the nearest neighbor in terms of the descriptor of every keypoint in the probe image is search from all the gallery images. [9]Gallery images that have at least three nearest neighbors are picked out as candidate images. Second, all nearest neighbors in a candidate image undergo a geometrical verification based on a set of affine transform parameters estimated by the clustered nearest neighbors in the Generalized Hough Transform. The image matching score is then computed based on the number of nearest neighbors that coincide with the affine transform. This image matching algorithm works well for visual object detection where gallery contains few and quite dissimilar objects to be detected from a probe image that contains several objects against a cluttered background. It is problematic if

this algorithm is applied to an identification problem where the gallery contains a lot of similar objects.

Figure 4.1: The process of keypoints matching strategy proposed by Lowe: the number of keypoints matched in the right image is 4 and the number of keypoints matched in the wrong image is 5.

Fig. 4.1 shows a wrong case when applying the matching algorithm of the SIFT framework [9] to the face identification. Fig. 4.1a shows the candidate images that contain at least three nearest neighbors to the probe keypoints. Geometric verification is then performed between the probe keypoints and their nearest neighbors in each candidate image, as shown in Fig. 4.1b. Only those matches (linked by straight lines) with coherent relative locations are kept. The final decision is based on the number of the matched keypoints, which is wrong apparently since the third candidate image has more matched keypoints than the correct (the first) candidate.

The problem occurs at the very first stage. From Fig. 4.1, I can see that only the keypoints that fulfill the nearest neighbor condition are considered for further processing. In the identification tasks, there are many similar gallery images. As a result, the nearest keypoints to the probe often disperse to many candidates in the gallery, and hence the probability that the largest number of the nearest keypoints fall into the right candidate

is low. This problem becomes severe if the gallery contains a large number of subjects. Moreover, multiple templates per subject in the gallery make it even worse. To alleviate this problem, I propose a new matching framework for face recognition as shown in Fig. 4.2, which has the following new features:

Figure 4.2: The process of keypoints matching strategy proposed by us: the number of keypoints matched in the right image is 16 and the number of keypoints matched in the wrong image is 7.

- (i) To find a set of candidate gallery images, I search the *k-nearest* neighbors of the *nearest subject*, where *k* is determined by the second nearest subject.
- (ii) A second matching process is introduced into the framework that *re-matches* the probe keypoints and keypoints in *each individual candidate gallery image* to augment the matched keypoint set.
- (iii) Compute the similarity scores between the probe image and the candidate gallery images based on the *accumulated keypoint similarities*.

4.1.2 Search the k -Nearest Neighbors of the Nearest Subject and Affine Transform Estimation

The best candidate match of a keypoint in the probe image is found by identifying its first nearest neighbor in the keypoint set of all gallery images. The first nearest neighbor is defined as the gallery keypoint whose descriptor has the maximum similarity based on Eq. 3.5 to that of the probe keypoint. The subject ID of the first nearest neighbor is recorded. Then, I further identify the k -nearest neighbors that have the same subject ID as the first nearest neighbor so that the $(k + 1)^{th}$ -nearest neighbor has a different subject ID. If two or more such nearest neighbors fall into a same gallery image, only the one with highest similarity is chosen from them. A candidate image is identified if at least three such k -nearest neighbors are found from it. I often obtain multiple candidate images. The minimum similarity s_m of the entire probe keypoints to their k -nearest neighbors is recorded for the second stage of the image matching. Note that I do not adopt the match rejection mechanism in the SIFT framework that is based on the similarity ratio between the first and the second nearest neighbors because it is not appropriate for the identification task. Fig. 4.2a shows the k -nearest neighbors (marked by circle) found in the gallery.

Based on the correspondence between the keypoints in the probe image and those in a candidate gallery image found in the k -nearest neighbor search, I can compute the affine transform parameters between the two images. I follow Lowe's approach here. Firstly, Generalized Hough Transform is used to cluster keypoints with similar pose transform by using each feature to vote for all object poses that are consistent with the feature. When clusters of features are found to vote for the same pose of an object, the probability of the interpretation being correct is much higher than for any single feature. Each of the keypoints specifies 4 parameters: 2D location, scale, and orientation, so does each matched keypoint. Therefore, a Generalized Hough Transform entry can be

created to predict the model location, orientation, and scale from the match hypothesis. This prediction has large error bounds, as the similarity transform implied by these 4 parameters is only an approximation to the full 6 degree-of-freedom pose space for a 3D object and also does not account for any non-rigid deformations. Therefore, Lowe used broad bin sizes of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times the maximum projected training image dimension for location. Soft voting is applied to avoid the problem of boundary effects in bin assignment.

Figure 4.3: The geometric relationship between points p, p' , and points q, q' .

Suppose keypoint p is the candidate match of the keypoint q . $[x, y], s, \theta$ are the location, scale and orientation of the keypoint p , while $[x_n, y_n], s_n, \theta_n$ are the location, scale and orientation of the keypoint q . As seen in Fig. 4.3, the location of point p' is:

$$x_1 = x + s \cdot \cos \theta, \quad y_1 = y + s \cdot \sin \theta. \quad (\text{Eq. 4.1})$$

Similarly, the location of point q' is:

$$x_{n1} = x_n + s_n \cdot \cos \theta_n, \quad y_{n1} = y_n + s_n \cdot \sin \theta_n. \quad (\text{Eq. 4.2})$$

If keypoint p is the candidate match of the keypoint q , then point p' is the candidate match of the point q' . Based on these two pairs of candidate matches, I can compute the

similarity transformation parameters s_0, θ_0, t_x, t_y between these two structures as:

$$A = \begin{bmatrix} x & -y & 1 & 0 \\ y & x & 0 & 1 \\ x_1 & -y_1 & 1 & 0 \\ y_1 & x_1 & 0 & 1 \end{bmatrix} \quad (\text{Eq. 4.3})$$

$$b = \begin{bmatrix} x_n \\ y_n \\ x_{n1} \\ y_{n1} \end{bmatrix} \quad (\text{Eq. 4.4})$$

$$c = A^{-1} * b \quad (\text{Eq. 4.5})$$

$$\theta_0 = \arctan(c(2), c(1)) \quad (\text{Eq. 4.6})$$

$$s_0 = c(1) / \cos(\theta_0) \quad (\text{Eq. 4.7})$$

$$t_x = c(3) \quad (\text{Eq. 4.8})$$

$$t_y = c(4) \quad (\text{Eq. 4.9})$$

I compute the similarity transformation parameters between each of the candidate keypoint pairs. Then all the parameters are assigned to the nearest multi-dimensional Generalized Hough Transform bins as illustrated in Fig. 4.4. Lowe proposed to assign each

Figure 4.4: Generalized Hough Transform bins with different values.

keypoint to the 2 closest bin in each dimension, giving a total 16 entries for each hypothesis. Each cluster/bin with at least 3 keypoints is used to estimate the affine projection parameters that project keypoints from the probe image to the candidate gallery image.

Six affine transformation parameters m_1, m_2, m_3, m_4, t_x and t_y are obtained, where m_i represent the affine rotation, scaling and stretch and $[t_x, t_y]$ is the translation vector. More details can be found in Lowe's paper [9]. Fig. 4.2b shows the keypoint pairs used to calculate the affine transformation parameters. Note that some k -nearest neighbors are rejected by this process due to their geometric inconsistency.

4.1.3 Further Matching between Probe and Each Candidate Gallery Image

Although the proposed method that searches the k -nearest neighbors of the nearest subject to a probe keypoint circumvents the problem of multiple templates per subject, the k -nearest neighbors often disperse to many different subjects if the gallery contains a large number of subjects. In general, the more subjects the gallery contains, the smaller the number of the k -nearest neighbors can be found in a candidate gallery image. This decreases the probability that the largest number of the nearest keypoints fall into the gallery image with the correct ID. This problem can be very severe if the gallery contains a large number of subjects. Thus, in an identification problem, the k -nearest neighbors of the probe keypoints found in a gallery image are often only a small portion of the keypoints that can be well matched with those in the probe image. Only considering the k -nearest neighbors in the whole database as the matched keypoints in a gallery image greatly weakens the discriminative power of the local structures of an image. Therefore, I propose to further search the keypoints in each single candidate gallery image that can well match with those in the probe.

I have obtained the six affine transform parameters m_1, m_2, m_3, m_4, t_x and t_y based on the matched keypoint pairs in the k -nearest neighbor search. I project the location of a probe keypoint $[x_p, y_p]$ to the gallery image $[x'_p, y'_p]$ by the affine transform as

$$\begin{bmatrix} x'_p \\ y'_p \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_p \\ y_p \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (\text{Eq. 4.10})$$

The geometric distance d between the location $[x_g, y_g]$ of a gallery keypoint and that of a transformed probe keypoint is computed as

$$d = \sqrt{(x_g - x'_p)^2 + (y_g - y'_p)^2}. \quad (\text{Eq. 4.11})$$

If a gallery keypoint i is geometrically close to a transformed probe keypoint j , $d_{ij} < d_t$, and their descriptor is similar, $s_{ij} > s_t$, keypoint i is identified as a candidate matched with keypoint j . The thresholds d_t and s_t are chosen respectively to be the one fourth of the translation bin width used in the Generalized Hough Transform and the minimum similarity s_m of the probe keypoints to their k -nearest neighbors. If multiple gallery keypoints satisfy the above conditions, the one with the maximum descriptor similarity is chosen as the matched keypoint. If there is no gallery keypoint satisfies the above conditions, the probe keypoint is not matched. The matched keypoint pairs are linked by straight lines in Fig. 4.2c. It shows a significant increase of the matched keypoints of the correct gallery image from Fig. 4.1b or Fig. 4.2b while the number of the matched keypoints of the incorrect gallery image remains almost the same. I see that the proposed matching algorithm greatly increases the discriminative power of the image local structures.

The thresholds d_t and s_t affect the number of matched keypoints of two images. It is difficult to find the optimal thresholds for all applications. To reduce the sensitivity of the image matching to the thresholds d_t and s_t , instead of the number of matched keypoints, I proposed to use the accumulated similarities over all probe keypoints. The similarity of a probe keypoint j to a candidate gallery image is defined as

$$s_j = \begin{cases} \max_{i \in \mathcal{I}_j} (s_{ij}), & \text{if } \mathcal{I}_j \neq \emptyset \\ 0, & \text{if } \mathcal{I}_j = \emptyset, \end{cases} \quad (\text{Eq. 4.12})$$

where $\mathcal{I}_j = \{i \mid d_{ij} < d_t \ \& \ s_{ij} > s_t\}$ and i is the index of the keypoint in the candidate gallery image. The similarity score of the probe image to the candidate gallery image

S_{pg} is then the accumulated similarities of all probe keypoints:

$$S_{pg} = \sum_{j=1}^q s_j, \quad (\text{Eq. 4.13})$$

where q is the number of keypoints in the probe image. The identity of the probe image is established as that of the gallery image that has the highest similarity score S_{pg} .

4.2 Training for Multiple Image Samples per Subject

While multiple templates per subject in general increase the recognition rate as shown in the experiments later, they bring a great computation burden in the recognition process. It is possible to greatly reduce the matching computation complexity by removing redundant information resides in the multiple training images. To meet different requirements for the matching complexity, I propose three schemes: template selection, unstable keypoint removal and template synthesis.

4.2.1 Template Selection

From a set of training images of a subject, I want to select a subset of images serving as templates that best represent all training images of this subject in terms of differentiating it from others in the training database. Suppose in the training data set $\mathcal{D}$, each subject has a training set $\mathcal{S}$ with N images, $N > 1$. To select q templates from them, $1 \leq q < N$, I pick a subset $\mathcal{P}_k$ that contains $N - q$ images of this subject out of the training data set $\mathcal{D}$. The remaining training images form a data set $\mathcal{G}_k$, $\mathcal{G}_k = \mathcal{D} - \mathcal{P}_k$. The candidate template set selected for this subject is then $\mathcal{T}_k = \mathcal{S} - \mathcal{P}_k$.

The face recognition framework proposed in the last two sections is applied to the probe set $\mathcal{P}_k$ and gallery set $\mathcal{G}_k$. The number of probe images that are correctly identified as the identity of $\mathcal{T}_k$ is recorded as n_k . Their similarity scores are accumulated and recorded as sc_k . In addition, the similarity scores of all the probe images to the most

similar gallery images in $\mathcal{G}_k - \mathcal{T}_k$ are accumulated and recorded as sf_k . This process is applied to all subsets of the q -combination of the elements in set $\mathcal{S}$, denoted by $\mathcal{T}_k$, $k = 1, 2, \dots, \binom{N}{q}$.

The subset $\mathcal{T}_k$ with the maximum value of n_k is selected as the templates of this subject. If there are multiple subsets $\mathcal{T}_k$ having the same maximum value of n_k , the subset $\mathcal{T}_k$ with the maximum value of sc_k among them is selected. If there are still multiple subsets having the same maximum values of n_k and sc_k , the subset $\mathcal{T}_k$ with the maximum value of sf_k among them is selected. Although the probability that multiple subsets $\mathcal{T}_k$ have the same maximum values of n_k and sc_k is zero for $N > 2$, this event likely occurs for $N = 2$. Larger value of sf_k indicates that the subset $\mathcal{T}_k$ is more dissimilar to the images of other subjects than the other subsets $\mathcal{T}_i, i \neq k$. The number of templates q can be determined by the event that the maximum value of n_k for $q + 1$ is not smaller than that for q .

4.2.2 Unstable Keypoint Removal

In Section 3.4.2, I propose an approach to remove insignificant keypoints based on their geometric location and descriptor similarity. If I have multiple training images per subject, I can check the repeatability of a keypoint in different images of the same subject. A keypoint with low repeatability is unstable and hence can be removed. Take a training image I_t from the training image set $\mathcal{S}$ of a subject and call it probe image and call all its keypoints probe keypoints in this section. The descriptor of a probe keypoint is compared with all the keypoints of the other subjects in the training set $\mathcal{D} - \mathcal{S}$ and the similarity of its nearest neighbor is denoted by s^b . The minimum similarity of the entire probe keypoints to their nearest neighbors is denoted by s_m^b . Then, the descriptor of this probe keypoint is compared with all the keypoints of the other images of the same subject. Its similarity to its nearest neighbor in the k^{th} image I_k is denoted by s_k^w , $I_k \in \mathcal{S}, k \neq t$. The

repeatability γ of this keypoint is initialized as zero and accumulated over all images I_k , $I_k \in \mathcal{S}, k \neq t$, by

$$\begin{cases} \gamma \leftarrow \gamma + 2, & \text{if } s_k^w > s^b \\ \gamma \leftarrow \gamma + 1. & \text{if } s_m^b < s_k^w \leq s^b, \end{cases} \quad \forall k : I_k \in \mathcal{S}, k \neq t \quad (\text{Eq. 4.14})$$

A keypoint of the image I_t with a larger value of γ has a higher repeatability. The rationale behind the two conditions and two different values added to γ can be seen in my two-stage image matching process in Section 4.1. I can set a threshold T_1 to select keypoints with high repeatability. If the value of γ of a keypoint is smaller than T_1 , it is removed.

As many keypoints have the same value of γ , I cannot well control the keypoint removal to some desirable number. To make the selection process more flexible, I further propose to distinguish keypoints with the same value of γ by the discriminative value δ defined as

$$\delta = \frac{\sum_{I_k \in \mathcal{S}, k \neq t} s_k^w}{(N-1)s^b}. \quad (\text{Eq. 4.15})$$

If the number of the keypoints satisfying $\gamma \geq T_1$ is larger than a desirable n but that satisfying $\gamma \geq T_1 + 1$ is smaller than n , I remove the keypoints satisfying $\gamma < T_1$ and the keypoints satisfying $(T_1 \leq \gamma < T_1 + 1) \& (\delta < T_2)$. In this way, I can keep a desirable number of keypoints by varying T_1 and T_2 . This process is visually shown in Fig. 4.5.

4.2.3 Template Synthesis

In general, multiple templates per subject lead to better recognition accuracy because they can represent different expressions, poses and illumination conditions of a subject. The proposed matching algorithm circumvents the problem caused by the multiple templates per subject. However, multiple templates greatly increase the computation complexity of the recognition process. And in some applications with limited computation power, single template per subject is required. Although the template selection

Figure 4.5: An illustration of the process to remove keypoints with low repeatability.

algorithm proposed in Section 4.2.1 can select the most representative template, The representation power of a single image is limited. A solution is template synthesis. Fig. 4.6 visually shows the process of template synthesis.

Figure 4.6: An illustration of the process of template synthesis.

First, the most representative template of a subject denoted by I_t is selected by my proposed matching algorithm. Second, this template is used as a probe image and another image I_k from the same subject and all images of the other subjects serve as gallery images. If the probe image is correctly identified by the algorithm proposed in Section 4.1, the unmatched keypoints in I_k are candidates to be integrated into the

template. Third, the repeatability and the discriminative value of each candidate is computed by the algorithm proposed in Section 4.2.2 if there are more than two training images of this subject. The stable candidate keypoints are integrated into the template I_t . If there are only two training images of this subject, all unmatched keypoints in I_k are integrated into the template I_t . Last, the geometrical affine transform between I_k and I_t established in the image match process based on Eq. 4.10 is used to transfer the stable candidate keypoints in I_k to the template I_t . This process is repeated for all images I_k , $k \neq t$ of the same subject.

4.3 Experimental Results

I shall first validate my image matching approach by comparing it with the counterpart of the SIFT framework, together with some multi-scale local feature-based approaches: Elastic Bunch Graph Matching (EBGM) [58, 71], Multi-scale Gabor feature (MG) [127], Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [70], Multi-scale Local Binary Pattern Histogram (MLBPH) [71] on FERET, ORL and GT databases. Then, I report the performances of my training process for multiple samples per subject on ORL, Georgia Tech (GT) and AR databases.

4.3.1 Validation of Image Matching Approach

4.3.1.1 Results on FERET Database I

There are 2388 images comprising 1194 persons (two images Fa/Fb per person) selected from the FERET database. Images are cropped into the size of 65×75 pixels. In the experiment, Fa images of each person form the gallery set and the remaining Fb images form the probe set. Table 4.1 records recognition performances of the original SIFT feature detection and description approach ($S_{dec}+S_{des}$) with the original SIFT matching framework (S_{mat}) [9], $S_{dec}+S_{des}$ with my proposed image matching framework

(P_mat), my feature detection and description approach (P_dec+P_des) with S_mat , and P_dec+P_des with P_mat .

Table 4.1: Recognition rate with single template per subject on FERET database I

MLBPH [71]	LGBPHS [70]	MG [127]	EBGM [58, 71]
98.5%	98%	90.5%	89.8%
$S_dec+S_des+S_mat$	$S_dec+S_des+P_mat$	$P_dec+P_des+S_mat$	$P_dec+P_des+P_mat$
92.96%	92.29%	97.15%	97.32%

In this experimental setting, there is only one template per subject in the gallery. I do not encounter the problem that the nearest keypoints to the probe disperse to many candidates of the same subject when there are multiple templates per subject. In other words, the first step of my matching process introduced in Section 4.1.2 is less important in this experiment. Hence, the recognition performance achieved by my matching approach is comparable to the original SIFT matching framework, which can be seen from Table 4.1. And my proposed approach achieves comparable recognition performances with other multi-scale local feature-based approaches.

4.3.1.2 Results on FERET Database II

This database is constructed, same to one data set used in [36], by choosing 256 subjects with at least four images per subject. And I use the same number of images (four) per subject for all subjects. The size of the image is 130×150 pixels. The i^{th} images of all the 256 subjects are chosen to form a gallery set, and the remaining three images per subject serve as the probe images to be identified from the gallery set. Table 4.2 shows the average recognition rates (ARR) and corresponding standard deviations (Std) over the four probe sets, each of which has a distinct gallery set ($i = 1, 2, 3, 4$).

In this experiment, there is only one image per subject in the gallery set. And the recognition performances are worse than those obtained in Section 4.3.1.1 due to the

Table 4.2: Average recognition rate and its standard deviation with single template per subject on FERET database II

	$S_dec+S_des+S_mat$	$S_dec+S_des+P_mat$	$P_dec+P_des+S_mat$	$P_dec+P_des+P_mat$
ARR	79.56%	80.89%	87.79%	89.68%
Std	0.88%	0.60%	1.44%	1.30%

increase of variations of probe images. However, I achieve higher recognition rates by my matching approach than the original SIFT matching method. Although the first step of my matching process has no effect in this experimental setting, the second matching step, which rematches the probe keypoints and keypoints in each individual candidate gallery image to augment the matched keypoint set, can relieve the problems caused by the variations of probe images.

4.3.1.3 Results on ORL database

The ORL database contains 400 images from 40 subjects taken at different times, varying the lighting, facial expressions and facial details. Images of the ORL database are resized into 50×57 pixels. Each subject has 10 images with index i from 1 to 10. The images from each subject with the same index i are picked out to form the gallery set and the remaining 360 images serve as the probe set. The face recognition rate is computed by the number of correctly identified probe images over 360. 10 runs of experiments are performed where each run has distinct gallery images. The average recognition rate (ARR) and its standard deviation (Std) over the 10 runs are recorded as indications of the recognition performance.

Table 4.3: Average recognition rate and its standard deviation with single template per subject on ORL database

	$S_dec+S_des+S_mat$	$S_dec+S_des+P_mat$	$P_dec+P_des+S_mat$	$P_dec+P_des+P_mat$
ARR	57.84%	62.47%	77.22%	81.56%
Std	2.52%	2.55%	1.64%	1.85%

Table 4.3 gives the ARR of $S_dec+S_des+S_mat$ [9], $S_dec+S_des+P_mat$, $P_dec+P_des+S_mat$ and $P_dec+P_des+P_mat$. The recognition rate is low because the large variations of the face pose and expression in this database cannot be well represented by a single template per subject. From Table 4.3, I can see that the recognition performance increases from 57.84% to 62.47% and from 77.22% to 81.56%, which demonstrates that the proposed matching approach significantly outperforms the counterpart in the original SIFT framework. In this experiment, one image per subject form the gallery set and nine images per subject form the probe set. Similarly, the first step of my matching process plays no role in this experiment. However, the rematching process between probe and each putative candidate gallery image alleviates the problem that nearest neighbors disperse into various gallery images of different subjects when there are multiple subjects in the database. This is the rationale behind the performance improvement. Note that the performance improvements in this experiment are more significant than those obtained in Section 4.3.1.2, which demonstrates that my matching approach can relieve the problems caused by the variations of probe images. The difference between the average recognition rates relative to their standard deviations shows the statistical significance of the experimental results.

4.3.1.4 Results on GT database

The GT face database contains 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray scale and resized into 60×80 pixels. I partition the GT database into the gallery set consisting of the first 8 samples per subject and the probe set consisting of the remaining 7 samples. Table 4.4 records recognition performances of the original SIFT feature detection and description approach (S_dec+S_des) with the original SIFT matching framework (S_mat) [9], S_dec+S_des with my pro-

posed image matching framework (P_mat), my feature detection and description approach (P_dec+P_des) with S_mat , and P_dec+P_des with P_mat .

Table 4.4: Recognition rate with multiple templates per subject on GT database

$S_dec+S_des+S_mat$	$S_dec+S_des+P_mat$	$P_dec+P_des+S_mat$	$P_dec+P_des+P_mat$
80.57%	82.86%	92.86%	95.43%

In this experimental setting, 8 images per subject form the gallery set. From Table 4.4, I can see that my matching approach can improve the recognition performance from 80.57% to 82.86% and from 92.86% to 95.43%. By the original SIFT matching approach, the nearest neighbors of probe keypoints disperse into many gallery images, which lowers the number of putative matches in the correct gallery image. However, my strategy of searching the k-nearest neighbors of the nearest subject solves the problem that nearest neighbors disperse into many gallery images of the same subject when there are multiple samples per subject. And the rematching process between probe and each putative candidate gallery image alleviates the problem that nearest neighbors disperse into various gallery images of different subjects when there are multiple subjects in the database. This is the rationale behind the performance improvement.

4.3.2 Validation of Template Selection and Synthesis

4.3.2.1 Validation of Template Selection for Multiple Samples per Subject

Three databases, ORL, GT and AR are applied to test the face recognition performances with multiple templates selected per subject. For ORL database, the first 5 samples per subject form the training set, and the remaining 5 samples serve as the probe images. The GT database contains 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray scale and resized into 60×80 pixels. The GT database is partitioned into the training set consisting of the first 8 samples per subject and the probe

set consisting of the remaining 7 samples. The color images in AR database are converted to gray-scale and cropped into the size of 120×170 pixels. In the experiments, 75 persons with 14 non-occluded images per person are selected, which makes the database containing 1050 images. The first 7 images per subject serve as training images and the remaining 7 images as probe images.

The SIFT framework ($S_dec+S_des+S_mat$) with default parameter in [9] and the proposed feature extraction and matching framework ($P_dec+P_des+P_mat$) are applied on the three databases with and without the proposed template selection approach described in Section 4.2.1. Table 5.5 shows the recognition rates on ORL, GT and AR databases. The first row under the name of the database shows the number of templates per subject used or selected in the gallery.

Table 4.5: Recognition rate on ORL, GT and AR databases

ORL database			
5		4	
$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$	$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$
90%	98.5%	89%	98%
GT database			
8		7	
$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$	$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$
80.57%	95.43%	78.6%	94.57%
AR database			
7		6	
$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$	$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$
97.14%	99.81%	96.77%	99.43%

On the ORL database, the proposed approach ($P_dec+P_des+P_mat$) achieves 98.5% recognition rate with 5 templates per subject, which is significantly better than the original SIFT algorithm ($S_dec+S_des+S_mat$). The template selection of 4 out of 5 only slightly decreases the recognition rate. On the GT database, the proposed approach ($P_dec+P_des+P_mat$) achieves a remarkably high recognition rate of 95.43% comparing to 80.57% of SIFT on 8 templates per subject. Reducing the number of the templates per subject from 8 to 7 by the proposed template selection algorithm worsens

the recognition performances a little. On the AR database, the proposed approach ($P_{dec}+P_{des}+P_{mat}$) achieves higher recognition rates than SIFT ($S_{dec}+S_{des}+S_{mat}$) on both 7 and 6 templates per subject. The template selection of 6 out of 7 by the proposed template selection approach slightly decreases the recognition rates. Table 5.5 shows that the deployment of multiple templates of a subject is an effective way to circumvent the problems caused by various intraclass variations of pose, expression and illumination. It seems that more templates of a subject than necessary cause no harm to the recognition accuracy. This is attributed to the proposed scheme that searches the *k-nearest neighbors* of the *nearest subject*.

4.3.2.2 Validation of Unstable Keypoint Removal

In Section 4.2.2, I proposed an approach to remove unstable keypoints if multiple training images per subject are available. To test the effect of this approach, I choose the first 5 samples per person of the ORL database as the training set, and the last 5 samples per person as the probe set. One template per subject is selected from the training images using the template selection method proposed in Section 4.2.1 and the keypoints of the selected template are removed based on the method proposed in Section 4.2.2 with the help of the other training images.

Table 4.6: Average number of keypoints per template and recognition rate under different threshold T

T	0	1	2	2.5	3	3.5	4	5
ANo.	101	93	86	81	75	69	63	36
RR %	85	85.5	86	86	85.5	84.5	83	80

Table 4.6 gives the average number ($ANo.$) of the remaining keypoints per template and the recognition rate (RR) under different threshold T , $T = T_1 + T_2$. It shows that the gain in recognition accuracy by the keypoint removal is insignificant. Removing the

keypoints by this method more than 30% reduces the recognition accuracy. Therefore, the keypoint removal by this method should not target at improving the recognition accuracy but at reducing the computational complexity of the recognition process.

4.3.2.3 Validation of Template Synthesis

To test the effect of the proposed template selection approach against the random template and the effect of the template synthesis, I choose the first 5 samples per person of the ORL database as the training images, and the last 5 samples per person as the probe images. I first pick out one template per subject randomly from the training set to form a gallery set and use it together with the probe set to test the recognition performance. This random template picking-up is repeated twenty times and the average recognition rate (ARR) and its standard deviation (Std) are computed. Then, one template per subject is selected from the training images using the template selection method proposed in Section 4.2.1. I test the recognition rate on the gallery formed by the selected templates and the probe set. Finally, some stable keypoints of the remaining training images, which are not selected, are integrated into the selected template based on the template synthesis approach proposed in Section 4.2.3. The synthesized template is basically a single template, which has 20% more keypoints on average than that before synthesis.

Table 4.7: Recognition rate of single template on ORL database

Template	Random	Random	Selected	Synthesized
Algorithm	$S_dec+S_des+S_mat$	$P_dec+P_des+P_mat$	$P_dec+P_des+P_mat$	$P_dec+P_des+P_mat$
ARR/RR	59.9%	80.6%	85%	89%
Std	4.60%	2.48%	N.A.	N.A.

Table 4.7 shows the recognition rates (ARR or RR). Firstly, $P_dec+P_des+P_mat$ achieves significantly higher recognition rate (80.6%) than the original SIFT approach (59.9%), which demonstrates again that my proposed feature extraction and matching approaches are more appropriate for face recognition tasks. Secondly, Table 4.7

demonstrates the contribution of the template selection and synthesis to the recognition accuracy. The recognition rate increases from 80.6% to 85% by my template selection approach and further increases to 89% by my template synthesis strategy. However, the single template selected and synthesized cannot remarkably increase the recognition rate because the representation power of a single template for lighting, pose and expression variations is limited.

4.4 Summary

The nearest neighbor search of the local structures suffers the problem that the nearest neighbors of different probe keypoints disperse into different templates of the correct subject. The proposed search of the k -nearest neighbors of the nearest subject solves this problem. Due to the inevitable intraclass variation, I cannot expect that a probe local structure is always nearest to the corresponding one of the same subject among all structures of all subjects in the gallery. The proposed second matching stage matches the local structures of the probe image individually with each candidate gallery image. This solve the problem caused by the large amount of subjects in the gallery. As mentioned in Section 3.6, the average keypoint detection time for one face image with size 60×85 is around 8.7295s. Given a gallery set with 525 face images, the average matching time per image is around 2.2975s. Suppose that n keypoints are detected in each image, and there are m images in the database. Then the complexity of my matching algorithm is $O(n^2m)$, because of the nearest neighbor search. Currently the algorithm is mainly written in matlab without any optimization. For real-world face recognition, other fast algorithms (like kd-trees [148]) approximate to nearest neighbor search should be adopted.

The deployment of multiple templates per subject is a solution to the large variation of the face pose and expression. This, however, imposes great computational burden on the recognition process, and hence I cannot arbitrarily increase the number of templates

of a subject in the gallery. In addition, a face database collected in practice often contains some very similar or even identical images, which need be trimmed off. Different training schemes including template selection, unstable keypoint removal and template synthesis are proposed to meet different requirements in the face recognition applications.

Chapter 5

Fully Automatic Face Recognition Framework Based on the Multi-scale Local Features

This chapter introduces a fully automatic face recognition framework based on both the local and global features. This framework includes automatic face alignment and identification.

5.1 Introduction

Generally, face recognition algorithms are divided into two categories based on the property of the features: holistic approaches [149] and local feature based approaches. PCA [150], LDA [19] and its variants [151, 22, 152, 153, 28, 154], Marginal Fisher Analysis (MFA) [155], and Eigenfeature Regularization and Extraction (ERE) [36] are popular holistic approaches due to their good performance and low computational complexity. These holistic approaches require a preprocessing procedure to normalize the face image variations in pose and scale, which is not an easy task because it depends on the accurate detection of at least two landmarks from the face image [156]. Some algorithms for eye localization have been proposed based on the eye ball [157, 158, 159, 160, 161, 162, 163, 164, 165]. However, in many real applications the appearances of eyeball are not distinct

or missing due to expressions, occlusions, illuminations or image noise. Hence, some algorithms localize multiple facial features like corners of eyes, nostrils, the tip of nose, corners of mouth, etc. Face alignment is performed based on these semantic features [65, 166, 167, 168, 169, 170]. The same problem encountered in the detection of eyes remains. Moreover, in the training process, these semantic features are hand-annotated, which is very labor-consuming. In [171], an unsupervised holistic approach is proposed for face alignment, which is not based on the localizations of semantic facial features. As the performance of the face alignment algorithm influences the final recognition performance, many research papers on the holistic approaches report the recognition performance on the pre-normalized faces based on the manually detected eyes. The recognition performance deteriorates considerably if the manual process is replaced by an automatic alignment algorithm.

In contrast to holistic methods, some local feature based approaches for face recognition are more robust to image variations in pose and scale [10, 9, 172]. Furthermore, unlike the holistic approaches, the face normalization is an integrated part of the local approaches. Recently, some initial attempts apply the Scale Invariant Feature Transform (SIFT) [10, 9] in the face recognition tasks [114, 173, 117, 120, 121, 123, 136, 137]. Experimental results in the previous two chapters show that the performance of the local feature based approach is significantly better than the popular holistic approaches. However, its computational time is much longer. For instance, to fulfill the face recognition task, one must search all the images in the database and compare each local feature in every image, which causes very heavy computational burden.

Although the computational complexity of holistic approaches is much lower than local feature based methods, they need an accurate face alignment. On the contrary, the local feature based approach proposed in the previous two chapters, which achieves better recognition performance than holistic approaches, is performed on unaligned face

images. However, its computational burden is much heavier. To solve these two critical problems in holistic and local feature based approaches, I propose a fully automatic face recognition framework based on both the local and global features. To speed up the local feature based approach, I propose to use the holistic approach as a filter to retrieve some candidate face images from the whole gallery set. The selected face images have higher probabilities matching to the probe, than the remains in the gallery. They form a new gallery set with reduced size, on which I perform the local feature based approach for face recognition. The reduction in the size of the gallery set speeds up the recognition process of local feature based approaches. To solve the alignment problem in holistic approaches, I design a face alignment scheme based on multi-scale local features instead of relying on the semantic facial parts. Generally, face alignment is performed based on the localizations of eyes, corners of mouth, nostrils and so on. However, in many real applications the appearances of facial parts are not distinct or missing due to expressions, occlusions, illuminations or image noise, which makes the alignment results unreliable. In face images, non-semantic facial features also hold distinct information, which can be utilized in the alignment process. Hence, I propose to align face images based on non-semantic multi-scale local features. The performance of my face alignment strategy is validated by face recognition tasks using local binary patterns (LBP) [174] and holistic approaches: LDA [19], UFS [28] and ERE [36]. Experimental results show that my alignment approach outperforms those based on localization of eyes [161, 162, 165], the localization of facial parts [175, 176] and the holistic alignment approach [171]. In my fully automatic face recognition framework, I adopt the holistic approach ERE [36] to narrow down the database as it is one of the best performed holistic approaches [37]. For the local feature based approach, I choose to use the multi-scale local feature extraction and matching framework (LFEM) described in Chapters 3 and 4, as it achieves better performance than many many other face recognition approaches. I firstly align face

images based on the multi-scale local features. Then ERE is performed on those well-aligned images to retrieve candidate faces from the whole gallery set. Finally, I perform LFEM on the narrowed gallery set. My fully automatic face recognition framework not only speeds up the local feature based approach, but also improves the recognition performance. My main contributions include:

- (i) I propose a face alignment approach based on multi-scale local features. Given an unaligned face image resulting from a face detector and a set of aligned face images in the data set, I build an automatic transformation mechanism, under which the unaligned face image can be precisely aligned for the recognition process.
- (ii) A fully automatic face recognition framework (FAFF) integrates the local feature based approach LFEM and the holistic approach ERE in a cascaded way. I firstly use ERE as a filter to retrieve some candidate images which form a gallery set with reduced size, where I perform LFEM for face recognition. This framework not only speeds up the LFEM approach, but also achieves better recognition performance than LFEM, ERE, and their parallel combination.

5.2 Framework Outline

Although some local feature based approaches achieve better recognition performance than holistic approaches [172], their computational burden is much heavier. To speed up the recognition process of local feature based approaches, in my fully automatic face recognition framework, I firstly perform holistic approach to retrieve candidate images from the whole gallery set. The selected images have higher probabilities matching to the probe than the remains in the whole gallery. Then I perform local feature based approach for face recognition on the narrowed gallery set, composed of the retrieved candidate images. The reduction in the size of the gallery set speeds up the recognition

Figure 5.1: Fully automatic face recognition framework based on both the local and global features.

process of local feature based approaches. Before I use the holistic approach as a filter, I must align face images into the same canonical pose, which is the premise of most holistic approaches. Usually, the alignment is done based on the localization of semantic facial parts like corners of eyes, eyeballs, corners of mouth, nostrils, etc.. However, it is difficult to precisely detect these structures in many real applications. Besides semantic features, non-semantic facial features also hold distinct information. Hence, I propose to perform face alignment based on the non-semantic multi-scale local features.

Fig. 5.1 visually gives the outline of my fully automatic face recognition framework based on local and global features. It consists of offline training and online recognition. In the offline training process, I firstly extract local and global features from the face database. The images in the database are aligned manually, where I learn a common face template using multi-scale local features. In the online recognition process, I build an

automatic transformation mechanism, under which the probe image is aligned with the common face template learned during the offline training. Then I perform the holistic approach on the well-aligned probe image and the face database to reduce the size of the gallery set. At last, the multi-scale local feature based approach is applied on the narrowed gallery set and outputs the final image ID. In the following sections, I will introduce each ingredient of my fully automatic face recognition framework in detail.

5.3 Face Alignment

The objective of my face alignment algorithm is not to localize facial feature points such as eye-brows, eyes, nose, mouth and contour of chin. The purpose of my alignment is to rectify face images into the same canonical pose for subsequent holistic recognition tasks. As mentioned in Section 5.1, face alignment algorithms based on localization of facial parts are not reliable as the appearances of semantic facial features vary with expressions, illuminations, occlusions or image noise. Hence, I propose an approach for face alignment not just relying on the semantic features. In the field of face recognition, there is one interesting phenomenon: the variations between the images of the same identity due to expression, illumination and viewing direction are almost always larger than image variations due to change of face identity. This is because the appearance of faces is highly constrained, for example, any frontal view of a face has eyes on the sides, nose in the middle, mouth in the lower part. Moreover, the appearances of facial components of different identities may be similar. My face alignment approach is inspired by this phenomenon.

5.3.1 Generate the Common Face Template

Given a set of face images $\mathcal{O}$ in the training database, I align them in pose and scale with manually detected two eye coordinates. My goal in this step is to learn a common face

Figure 5.2: Mean face $\mathbf{m}$ with (a) SIFT keypoints; (b) Symmetric keypoints $\mathcal{K}$.

template based on these aligned face images $\mathcal{I}$. As mentioned above, the similarities are high, among the facial componential appearances of different subjects. The mean face $\mathbf{m}$ of $\mathcal{I}$ captures the common information of various identities and removes noises. The SIFT keypoints detected in $\mathbf{m}$ tell us the locations of the common and stable features in $\mathcal{I}$. Fig. 5.2(a) shows the mean face $\mathbf{m}$ computed from $\mathcal{I}$ with SIFT keypoints. I further add extra keypoints to meet the symmetry property of face images, as shown in Fig. 5.2(b). The keypoints in Fig. 5.2(b) are the anchor points in the common face template $\mathbf{m}$.

The anchor points in $\mathbf{m}$ tell us the possible locations of common features in $\mathcal{I}$. However, their descriptors provide little information, as $\mathbf{m}$ is the mean face image. The support area of SIFT descriptors in $\mathbf{m}$ is smoothed by the mean. I need to compute the descriptors directly from the individual images in the original training set $\mathcal{O}$. There are pose variations in the image set $\mathcal{O}$. The locations of the detected keypoints in $\mathcal{O}$ cannot be used in the alignment process. I should project their locations into the coordinates of the well-aligned image set $\mathcal{I}$. Let $\mathcal{P}_q = \{p_i^q\}$, $q = 1, \dots, Q$, where Q is the number of images in the training set $\mathcal{O}$, represent the keypoint set detected in the q^{th} image of $\mathcal{O}$,

where p_i^q is the i^{th} keypoint in $\mathcal{P}_q$. Suppose that the two eye coordinates of the q^{th} face image in the set $\mathcal{O}$ are $[e_{1x}, e_{1y}; e_{2x}, e_{2y}]$, and the two eye coordinates in the corresponding well-aligned face image in the set $\mathcal{I}$ are $[a_{1x}, a_{1y}; a_{2x}, a_{2y}]$. Based on these two pairs of corresponding points, I can compute the similarity transformation parameters $[s, \theta, t_x, t_y]$ between image sets $\mathcal{O}$ and $\mathcal{I}$ as below:

$$\begin{bmatrix} s \cos \theta \\ s \sin \theta \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} e_{1x} & -e_{1y} & 1 & 0 \\ e_{1y} & e_{1x} & 0 & 1 \\ e_{2x} & -e_{2y} & 1 & 0 \\ e_{2y} & e_{2x} & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} a_{1x} \\ a_{1y} \\ a_{2x} \\ a_{2y} \end{bmatrix} \quad (\text{Eq. 5.1})$$

Once I get the transformation parameters, I can project the location of the keypoint p_i^q , $[x_p, y_p]$, to the corresponding well-aligned coordinates $[x'_p, y'_p]$ by

$$\begin{bmatrix} x'_p \\ y'_p \end{bmatrix} = \begin{bmatrix} s \cos \theta & -s \sin \theta \\ s \sin \theta & s \cos \theta \end{bmatrix} \begin{bmatrix} x_p \\ y_p \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (\text{Eq. 5.2})$$

I perform this projection on the locations of each keypoint in the set $\mathcal{P}$. Thus the keypoint descriptors of $\mathcal{P}$ capture various pose information, and their locations are well-aligned.

Let $\mathcal{K} = \{k_i\}$, $i = 1, \dots, t$, denote the anchor point set in $\mathbf{m}$, where t is the number of anchor points in the common face template. Image set $\mathcal{O}$ is the original images with pose variations. The keypoint descriptors of $\mathcal{P}$ capture various pose information. To enhance the representative power of the template image $\mathbf{m}$, I embed the descriptors of $\mathcal{P}$ into the anchor keypoint set $\mathcal{K}$. In a region R around the location of k_i , I search its neighbors in $\{\mathcal{P}_1, \dots, \mathcal{P}_Q\}$. R is set to 1/6 times maximum of the image dimension in the experiments. If there are multiple keypoints in one face image falling into R , I select the one which is nearest to the location of k_i . In this way, around each anchor point k_i , I can find a series of keypoints $\mathcal{N}_i$ from different face images. Comparing with the scheme of one anchor point with one descriptor from one face image, variance of keypoints coming from the same semantic region of different faces enrich feature representation and are less subject to pose variations. Note that I only use the location information of anchor point k_i to

locate candidate keypoints nearby in $\mathcal{P}$. I do not use the location or the descriptor of the anchor point k_i during the alignment process.

Now around each anchor point k_i , there is a series of keypoints $\mathcal{N}_i$. To make the number of keypoints in $\mathcal{N}_i$ less dependent to the number of images Q in the training database $\mathcal{O}$, I adopt hierarchical clustering [177] to group the descriptors of each keypoint set $\mathcal{N}_i$ into h clusters. The cluster center c_j^i , where $j = 1, \dots, h$, is selected as the descriptor who has the largest accumulated cosine similarities among all the other descriptors in the same cluster. If the number of keypoints in $\mathcal{N}_i$ is smaller than h , I keep all the keypoints in $\mathcal{N}_i$. Hence, in the common face template $\mathbf{m}$, the final number of keypoints is smaller than or equal to $t \times h$. And I denote these keypoints in the template image $\mathbf{m}$ as final anchor point set $\mathcal{T}$.

5.3.2 Establish the Feature Correspondences

Now in the template image $\mathbf{m}$, there are at most $t \times h$ anchor points extracted from various face images. Suppose that image I is the output of some face detector, which should be aligned into the same canonical pose as the template image $\mathbf{m}$ for the subsequent holistic recognition process. SIFT keypoint set $\mathcal{B}$ is extracted from the image I . The best candidate match of a probe keypoint in $\mathcal{B}$ is found by identifying its nearest neighbor in the anchor point set $\mathcal{T}$. The nearest neighbor is defined as the anchor point whose descriptor has the maximum similarity to that of the probe keypoint.

The nearest-neighbor search can only establish putative correspondences between keypoint sets $\mathcal{B}$ and $\mathcal{T}$. To eliminate spurious keypoint pairs, I further check their geometric consistencies. There are 4 parameters for each SIFT keypoint: 2D location, scale and orientation. I use Generalized Hough Transform to cluster keypoint pairs with similar poses. The orientation bin size used in the Generalized Hough Transform is 30 degrees, the scale bin size is 2, and the location bin size is 0.25 times the image size.

Figure 5.3: (a) Sample images before alignment; (b) Sample images after alignment.

Each keypoint pair votes for the 2 closest bins in each dimension. Each bin with at least 3 keypoints is used to estimate the affine projection parameters that project the keypoint set $\mathcal{B}$ from the probe image I to the anchor point set $\mathcal{T}$. An affine transformation correctly accounts for 3D rotation of a planar surface under orthographic projection, but the approximation can be poor for 3D rotation of non-planar faces. A more general solution would be to solve for the fundamental matrix. However, a fundamental matrix solution requires at least 7 keypoint pairs as compared to only 3 for the affine solution and in practice requires even more matches for good stability. Note that the affine transformation here is used to select keypoint pairs which are geometrically consistent. I can account for errors in the affine approximation by allowing large residual errors. When the number of keypoint pairs in the bin is larger than 3, least-squares solution is adopted to compute the affine transformation parameters. After the geometric verification, I can obtain keypoint pairs $\mathcal{B}_{sub}$ and $\mathcal{T}_{sub}$, which are subsets of $\mathcal{B}$ and $\mathcal{T}$ respectively. Note that some putative keypoint pairs are rejected by this process due to their geometric inconsistency. The anchor point set $\mathcal{T}_{sub}$ contains the final anchor points for the keypoint set $\mathcal{B}_{sub}$ to align to.

5.3.3 Face Alignment by Similarity Transformation

The purpose of my face alignment is to rotate, resize and crop the output face images of face detectors automatically, and transform them into canonical pose for the subsequent

holistic recognition tasks. I do not want to change their structures. Hence, I adopt similarity transformation in the final alignment step. The similarity transformation gives the mapping of a model point $[x, y]$ to an image point $[u, v]$ in terms of an image scaling s , rotation θ , and translation $[t_x, t_y]$. I project the location of a probe keypoint $[x_p, y_p]$ in $\mathcal{B}_{sub}$ to its corresponding anchor keypoint $[x_p', y_p']$ in $\mathcal{T}_{sub}$ by the similarity transform as

$$\begin{bmatrix} x_p & -y_p & 1 & 0 \\ y_p & x_p & 0 & 1 \\ & \dots & & \\ & \dots & & \end{bmatrix} \begin{bmatrix} s \cos \theta \\ s \sin \theta \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} x_p' \\ y_p' \\ \cdot \\ \cdot \end{bmatrix} \quad (\text{Eq. 5.3})$$

Each matched keypoint pair contributes two rows to the first and last matrices in Eq. 5.3.

At least 2 matches are needed to provide a solution. I can write this linear system as

$$\mathbf{Ax} = \mathbf{b} \quad (\text{Eq. 5.4})$$

The least-squares solution for the parameters $\mathbf{x}$ can be determined by solving the corresponding normal equations

$$\mathbf{x} = [\mathbf{A}^T \mathbf{A}]^{-1} \mathbf{A}^T \mathbf{b}, \quad (\text{Eq. 5.5})$$

which minimizes the sum of the squares of the distances from the projected model locations to the corresponding image locations.

Once I obtain the transformation parameters $[s, \theta, t_x, t_y]$, I can transform the probe image I according to the 2-D spatial similarity transformation. Fig. 5.3(a) shows some sample images before alignment. Fig. 5.3(b) shows the corresponding images aligned by my approach.

5.4 Fully Automatic Face Recognition Framework based on Local and Global Features

5.4.1 Eigenfeature Regularization and Extraction

Jiang [36] proposed a holistic approach named eigenfeature regularization and extraction (ERE), which facilitates a discriminative and a stable low-dimensional feature represen-

tation of the face image. Experiments comparing ERE with some other popular subspace methods on various databases show that ERE consistently outperforms others [36, 37].

Given a set of properly normalized w -by- h face images, I can form a training set of column image vectors X_{ij} , where $X_{ij} \in \mathbb{R}^{n=wh}$. Let the training set contain p persons and q_i sample images for person i . The number of total training sample is $l = \sum_{i=1}^p q_i$. For face recognition, each person is a class with prior probability of c_i . The within-class scatter matrix is defined by $\mathbf{S}^w = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$, where $\bar{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. By solving the eigenvalue problem, I have

$$\mathbf{\Lambda} = \mathbf{\Phi}^{wT} \mathbf{S}^w \mathbf{\Phi}^w, \quad (\text{Eq. 5.6})$$

where $\mathbf{\Phi}^w = [\phi_1^w, \dots, \phi_n^w]$ is the eigenvector matrix of $\mathbf{S}^w$, and $\mathbf{\Lambda}^w$ is the diagonal matrix of eigenvalues $\lambda_1^w, \dots, \lambda_n^w$ corresponding to the eigenvectors.

The eigenspace $\mathbb{R}^n$ spanned by eigenvectors $\{\phi_k^w\}_{k=1}^n$ is decomposed into three subspaces: a reliable face variation dominating subspace (or simply face space) $\mathbf{F} = \{\phi_k^w\}_{k=1}^m$, an unstable noise variation dominating subspace (or simply noise space) $\mathbf{N} = \{\phi_k^w\}_{k=m+1}^r$ and a null space $\emptyset = \{\phi_k^w\}_{k=r+1}^n$, where r is the rank of $\mathbf{S}^w$. To determine the start point of the noise dominant region $m+1$, I first find the median of nonzero eigenvalues $\lambda_{med}^w = \text{median}\{\forall \lambda_k^w | k \leq r\}$. The distance between λ_{med}^w and the smallest nonzero eigenvalue is $d_{m,r} = \lambda_{med}^w - \lambda_r^w$. The start point of the noise region $m+1$ is estimated by

$$\lambda_{m+1}^w = \max\{\forall \lambda_k^w | \lambda_k^w < (\lambda_{med}^w + \mu(\lambda_{med}^w - \lambda_r^w))\}, \quad (\text{Eq. 5.7})$$

where μ is a constant. The optimal value of μ may be slightly larger or smaller than 1 for different applications.

The eigenspectrum is regularized by replacing the noise dominating λ_k^w in $\mathbf{N}$ with a model $\alpha/(k + \beta)$ and replacing the zero λ_k^w in the null space $\emptyset$ with a constant. The regularized eigenspectrum $\tilde{\lambda}_k^w$ is given by

$$\tilde{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m; \\ \frac{\alpha}{k+\beta}, & m \leq k \leq r; \\ \frac{\alpha}{r+1+\beta}, & r < k \leq n, \end{cases} \quad (\text{Eq. 5.8})$$

where $\alpha = \frac{\lambda_1^w \lambda_m^w (m-1)}{\lambda_1^w - \lambda_m^w}$, and $\beta = \frac{m\lambda_m^w - \lambda_1^w}{\lambda_1^w - \lambda_m^w}$. The training data are transformed to $\tilde{Y}_{ij} = \tilde{\Phi}_n^{wT} X_{ij}$, where

$$\tilde{\Phi}_n^w = [\phi_1^w / \sqrt{\tilde{\lambda}_1^w}, \dots, \phi_n^w / \sqrt{\tilde{\lambda}_n^w}]. \quad (\text{Eq. 5.9})$$

After the feature regularization, a new total scatter matrix is formed by vectors $\tilde{Y}_{ij}$ of the training data as $\tilde{\mathbf{S}}^t = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (\tilde{Y}_{ij} - \bar{Y})(\tilde{Y}_{ij} - \bar{Y})^T$, where $\bar{Y} = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} \tilde{Y}_{ij}$.

The regularized features $\tilde{Y}_{ij}$ are decorrelated for $\tilde{\mathbf{S}}^t$ by solving the eigenvalue problem similar to Eq. 5.6. The eigenvectors in the eigenvector matrix $\tilde{\Phi}_n^t = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_n^t]$ are sorted in a descending order of the corresponding eigenvalues. The dimensionality reduction is performed by keeping the eigenvectors with the d largest eigenvalues $\tilde{\Phi}_d^t = [\tilde{\phi}_k^t]_{k=1}^d = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_d^t]$, where d is the number of features usually selected by a specific application. The feature regularization and extraction matrix $\mathbf{U}$ is given by $\mathbf{U} = \tilde{\Phi}_n^w \tilde{\Phi}_d^t$, which transforms a face image vector X , $X \in \mathbb{R}^n$, into a feature vector F , $F \in \mathbb{R}^d$, by $F = \mathbf{U}^T X$.

5.4.2 Fully Automatic Face Recognition Framework based on both the Local and Global Features

The local feature based approach introduced in the previous two chapters achieves better recognition performance than holistic approaches, but its computational burden is much heavier. To speed up the local feature based approach, I propose to use the holistic approach as a filter to retrieve candidate images from the whole gallery set. The retrieved images form a narrowed gallery set, on which I perform the multi-scale local feature based approach for face recognition. The reduction in the size of the gallery set significantly relieves its computational complexity.

Image I is the output of a face detector. Firstly, I align image I by the approach proposed in Section 5.3. Then the aligned image I' can be used as the input of the holistic approach ERE [36]. To calculate the similarities between the probe image I' and all the

gallery images, each n -D face image vector is transformed into d -D feature vector F by using the feature regularization and extraction matrix $\mathbf{U}$ obtained in the training stage. Then the Cosine similarity measure between a probe feature vector F_p and a gallery feature vector F_g is applied,

$$S(F_p, F_g) = \frac{F_p^T F_g}{\|F_p\|_2 \|F_g\|_2}, \quad (\text{Eq. 5.10})$$

where $\|\cdot\|_2$ is the norm 2 operator.

There are m similarity scores $\mathcal{S} = \{S_1, S_2, \dots, S_m\}$ between image I' and m gallery images. The scores S_i , $i = 1, \dots, m$ are ordered as

$$S_{(1)} \geq S_{(2)} \geq \dots \geq S_{(m)}, \quad (\text{Eq. 5.11})$$

and the gallery images with the top n highest similarity scores are selected as the input of the following multi-scale local feature extraction and matching framework (LFEM). The final image ID is the one with the highest similarity score of the n candidate gallery images computed by the LFEM approach. The rationale behind this cascaded process are as follows:

- (i) ERE is one of the best performed holistic approaches.
- (ii) The probability that the true image ID is included in the top n gallery images is high.
- (iii) The computational complexity of the ERE approach at the recognition stage is significantly lower than the LFEM approach.
- (iv) The recognition performance of LFEM is better than the ERE approach.
- (v) For the LFEM approach, comparing image I' with n gallery images is much faster than comparing it with all the m gallery images ($n < m$).

5.5 Experimental Results

In this section, I test my proposed approaches on AR, Georgia Tech (GT) and ORL databases. Firstly, I validate my face alignment strategy through LBP [69] and three holistic face recognition approaches: LDA [19], UFS [28] and ERE [36]. Secondly, I report the results of my fully automatic face recognition framework based on both the local and global features (FAFF).

5.5.1 Validation of My Face Alignment Strategy

Eight face alignment approaches are compared: Adaboost eye detector [175, 176], eye localization by pixel differences (PD) [161], eye localization by rank order filter (ROF) [162], eye localization by cascaded asymmetric principal and discriminative analysis (C-APCDA) [178, 165], localization of semantic facial features (SF) [167], Congealing [171], my proposed approach (Prop.) and manual alignment (MA).

5.5.1.1 Results on AR Database

The color images in AR database are converted to gray-scale and cropped into the size of 60×85 pixels. I conduct two sets of experiments on this database.

In the first experiment, 75 persons with 14 non-occluded images per person are selected, which makes the database containing 1050 images. At the alignment stage, for the C-APCDA eye detector [165], the first 7 images per subject serve as training images and are aligned manually by two eye coordinates. The remaining 7 images per subject serve as the output of the face detector, which should be aligned. For my proposed alignment approach, the first 7 images per subject are aligned manually to generate the face template. At the recognition stage, images are divided into 5×5 windows for the LBP [69] approach. For the three holistic approaches LDA [19], UFS [28] and ERE [36], the first 7 images of all subjects are used in the training and gallery sets, which are normalized

manually. The remaining 7 images of all subjects serve as probe, which are aligned by different approaches. The best recognition performances of the holistic approaches LDA [19], UFS [28] and ERE [36] over all possible numbers of features are recorded.

Figure 5.4: Sample images of occluded face.

To verify the performance of my alignment approach on occluded face images, in the second set of experiments, I select 55 persons with 12 occluded images per person (some sample images are shown in Fig. 5.4). The first 6 images of all 55 subjects are used in the training and gallery sets and the remaining 6 images of all subjects serve as probe. Obviously, alignment approaches that are only based on eye localizations fail in this case because of the occlusion. Therefore, I compare the performances of three alignment approaches: localization of semantic facial features (SF) [167], my proposed approach and the manual alignment.

Table 5.1 shows the rank one recognition rates on AR database of four recognition algorithms based on seven different alignment approaches. From this table, I can see that in the first set of experiments, the recognition performance achieved by my alignment approach, though slightly worse than manual alignment, outperforms all other automatic alignment approaches. Because in my alignment approach I do not just depend on the positions of semantic facial parts, my approach can get more reliable alignment results. Fig. 5.5(a) gives the cumulative matching curves obtained from seven different alignment

approaches based on the ERE approach. The cumulative recognition performance obtained by my alignment approach is significantly better than other automatic alignment approaches. The second set of experiments is conducted on the occluded face images. The results in Table 5.1 show that my proposed alignment approach significantly outperforms the alignment approach based on localizations of semantic facial features [167]. The recognition rates are quite low because all face images in the training/gallery and probe sets are occluded.

Table 5.1: Recognition rate on AR database

<i>SET 1</i>	LBP	LDA	UFS	ERE
MA	98.10%	93.90%	95.05%	95.05%
Adaboost [175, 176]	95.81%	89.71%	91.05%	90.67%
PD [161]	87.24%	80.38%	82.86%	82.29%
ROF [162]	73.52%	69.71%	70.48%	71.24%
C-APCDA [165]	94.67%	89.52%	90.10%	89.14%
SF [167]	91.05%	84.76%	87.05%	86.48%
Prop.	95.43%	92.95%	92.95%	94.10%
<i>SET 2</i>	LBP	LDA	UFS	ERE
MA	86.67%	72.12%	72.12%	80.30%
SF [167]	50.91%	23.64%	27.58%	32.12%
Prop.	76.06%	58.79%	60.00%	62.73%

5.5.1.2 Results on GT Database

The GT database consists 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray scale and cropped into the size of 60×80 . At the alignment

stage, for the C-APCDA approach [165], the first 8 images per subject serve as training images and are aligned manually based on the two eye positions. The remaining 7 images per subject serve as the output of the face detector, which should be aligned. For my proposed approach, the first 8 images per subject are aligned manually to generate the template. At the recognition stage, images are divided into 5×5 windows for the LBP approach [69]. For the three holistic approaches LDA [19], UFS [28] and ERE [36], the first 8 images of all subjects are used in the training and gallery sets, which are normalized manually. The remaining 7 images of all subjects serve as probe images, which are aligned by different approaches. The best recognition performances of the holistic approaches LDA [19], UFS [28] and ERE [36] over all possible numbers of features are recorded.

Table 5.2: Recognition rate on GT database

	LBP	LDA	UFS	ERE
MA	90.29%	92.00%	91.43%	92.86%
Adaboost [175, 176]	85.43%	82.29%	81.71%	83.71%
PD [161]	72.00%	64.29%	69.14%	71.43%
ROF [162]	84.29%	60.29%	64.00%	67.43%
C-APCDA [165]	85.43%	83.43%	83.71%	86.29%
Congealing [171]	83.43%	84.29%	82.29%	85.71%
SF [167]	85.43%	83.43%	84.57%	86.00%
Prop.	84.29%	92.57%	90.57%	92.00%

Table 5.2 shows the rank one recognition rates on GT database of four recognition algorithms based on eight different alignment approaches. Comparing different automatic face alignment approaches, the performance of my face alignment approach is significantly better than eye-based approaches: Adaboost [175, 176], PD [161], ROF[162] and C-APCDA [165], semantic feature based approach [167] and the holistic unsupervised

Figure 5.5: Cumulative matching curves of different alignment approaches obtained by ERE on (a) AR database; (b) GT database; (c) ORL database.

approach: Congealing [171]. Note that in the experiments, manual alignment is to align the face images based on the manually detected eye coordinates, which only ensures the positions of the eyes in all the images are the same. The positions of other facial parts may be different. Manual alignment is not perfect, which is shown by the fact that my proposed alignment approach for the LDA method achieves better performance than the manual alignment. Fig. 5.5(b) gives the cumulative matching curves obtained from eight different alignment approaches based on the ERE approach. The cumulative recognition performance obtained by my alignment approach is significantly better than other automatic alignment approaches.

5.5.1.3 Results on ORL Database

Images of ORL database are cropped into the size of 50×57 . The ORL database contains 400 images of 40 people (10 images per person). At the alignment stage, for the C-APCDA approach [165], the first 5 images per subject serve as training images and are aligned manually by the two eye coordinates. The remaining 5 images per subject serve as the output of the face detector, which should be aligned. For my proposed alignment approach, the first 5 images per subject are aligned manually to generate the template. At

the recognition stage, for the LBP approach [69], images are divided into 3×3 windows. For the three holistic approaches LDA [19], UFS [28] and ERE [36], the first 5 images of all subjects are used in the training and gallery sets, which are normalized manually. The remaining 5 images of all subjects serve as probe images, which are aligned by different approaches. The best recognition performances of the holistic approaches LDA [19], UFS [28] and ERE [36] over all possible numbers of features are recorded.

Table 5.3: Recognition rate on ORL database

	LBP	LDA	UFS	ERE
MA	95.0%	92.5%	83.5%	97.0%
Adaboost [175, 176]	87.5%	88.5%	77.0%	91.0%
PD [161]	85.5%	80.5%	69.5%	84.5%
ROF [162]	86.5%	74.5%	65.5%	80.0%
C-APCDA [165]	85.5%	83.0%	62.5%	88.0%
SF [167]	93.5%	93.0%	81.5%	96.5%
Prop.	94.0%	93.0%	83.0%	95.5%

Table 5.3 shows the rank one recognition rates on ORL database of four recognition algorithms based on seven different alignment approaches. Comparing different automatic alignment approaches, my proposed approach achieves comparable performance as the semantic feature based approach SF [167], but consistently significantly outperforms others over different face recognition algorithms. Note that in the experiments, manual alignment is to align the face images based on the manually detected eye coordinates, which only ensures the positions of the eyes in all the images are the same. The positions of other facial parts may be different. Manual alignment is not perfect, which is shown by the fact that my proposed alignment approach for the LDA method performs

slightly better than the manual alignment. Fig. 5.5(c) gives the cumulative matching curves obtained from seven different alignment approaches based on the ERE approach. The cumulative recognition performance obtained by my alignment approach is ,though slightly worse than the SF [167] approach, significantly better than other automatic alignment approaches.

5.5.2 Validation of My Fully Automatic Face Recognition Framework

In this section, I validate my fully automatic face recognition framework (FAFF) based on the holistic approach eigenfeature regularization and extraction (ERE) [36] and the multi-scale local feature extraction and matching framework (LFEM) [172], on three databases AR, GT and ORL. My fully automatic face recognition framework utilizes the classification results of both the ERE and the LFEM approaches. From some points

Table 5.4: Percentage of the gallery images selected by ERE where FAFF achieves the same recognition performance as LFEM on the original database at r_1 and the best performance at r_2

AR database		GT database		ORL database	
r_1	r_2	r_1	r_2	r_1	r_2
40%	40%	15%	80%	20%	40%

of view, it is a kind of score fusion [179] for multiple algorithms. And from the other point of view, it is similar to image retrieval. To validate the efficacy of my proposed framework, I compare it with ERE, LFEM, two score fusion approaches: sum rule with equal weights (SR_{ew}) [179] and sum rule with weights learned by LIBLINEAR (SR_{lw}) [180], LBP, LDA, UFS, multiple keypoint detectors with SIFT descriptors (MKDS) [120], dense multi-scale features (DMF) [125], the attribute face service (AFS) [53, 54], and the vocabulary tree (VT) [181]. The experimental settings in this section are similar to those

in Section 5.5.1. The only difference is that in this section, the images used in the holistic approaches are aligned automatically by my proposed face alignment strategy.

Table 5.4 gives the percentage of the gallery images retrieved by the holistic approach ERE at two different levels r_1 and r_2 . My FAFF approach achieves the same recognition performance as LFEM by comparing one probe with r_1 images in the gallery. And by comparing with r_2 images, the recognition performance of my framework reaches the best.

Table 5.5: Recognition rate on AR, GT and ORL databases

	AR	GT	ORL
ERE	94.10%	92.00%	95.5%
LFEM	99.05%	95.43%	98.0%
SR_{ew}	98.48%	96.86%	98.5%
SR_{lw}	99.05%	96.29%	98.0%
FAFF	99.05%	97.43%	99.0%
LBP	95.43%	84.29%	94.0%
LDA	92.57%	92.57%	93.0%
UFS	92.95%	90.57%	83.0%
MKDS [120]	-	-	97.0%
DMF [125]	-	-	96.5%
AFS [53, 54]	97.14%	94.00%	-
VT [181]	93.14%	91.14%	99.0%

Table 5.5 gives the rank one recognition rate of the holistic approach ERE [36], local feature based approach LFEM, score fusion of ERE and LFEM by sum rule with equal weights (SR_{ew}) [179] and learned weights (SR_{lw}) [180], and the best recognition performance of FAFF by selecting r_2 images of all the gallery. From this table, I see that on AR

database, the best recognition rate is achieved by LFEM, SR_{lw} and FAFF. However, the proposed FAFF is about twice as faster as LFEM and SR_{lw} because the number of gallery images is reduced to 40% by ERE (from Table 5.4) and the time consumption of ERE is negligible comparing to LFEM. On GT database, the FAFF approach outperforms all other approaches in Table 5.5 and is yet less time consuming than LFEM, SR_{ew} and SR_{lw} because the number of gallery images is reduced to 80% by ERE. SR_{ew} and SR_{lw} outperform LFEM at a price of greater computational complexity as they parallel combine LFEM with ERE. In contrast, the proposed FAFF not only speeds up the LFEM, but also enhances its accuracy more than the two parallel combination approaches SR_{ew} and SR_{lw} . According to Table 5.4, the proposed FAFF can achieve the same recognition rate as LFEM by comparing the probe with only 15% gallery images. Similarly, on ORL database, by comparing the probe with 20% gallery images, the proposed FAFF reaches the same performance as LFEM. Increasing the number of images to 40% of the whole gallery set, the recognition rate of the proposed FAFF is the highest among all approaches in Table 5.5.

5.6 Summary

Face recognition algorithms can be divided into two categories based on the types of features: holistic and local feature based approaches. Holistic approaches become popular due to their efficacy and efficiency. However, they depend on careful positioning of the face images into the same canonical pose. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. Besides eye detection, some face alignment algorithms rely on other semantic facial parts. However, in real applications the appearances of these semantic features may not be distinct or missing due to expressions, occlusions, illuminations or noise. Local feature based approaches to face recognition [10, 9, 172] are more robust to image variations in pose and scale than

holistic ones. The multi-scale local feature extraction and matching framework proposed in [172] even achieves better recognition accuracy than many popular holistic methods. However, its computational burden is much heavier.

In this chapter, I propose a fully automatic face recognition framework based on both the local and global features. To relieve the computational burden of local feature based approaches, I firstly apply the holistic method to retrieve candidate images from the whole gallery set. The selected gallery images have higher probabilities matching to the probe than the remains, which form a new gallery set with reduced size. Then the local feature based approach is performed on the narrowed gallery set. The reduction in the size of the gallery set relieves the computational burden of local feature based approaches. Furthermore, the recognition accuracy is better than the holistic approach, local feature based approach and their parallel combination. As the local features can be used to align images automatically, I propose an alignment strategy based on non-semantic multi-scale local features. Given a set of aligned images in the training data set, a common face template with common keypoints is trained. Putative correspondences are established between the keypoint sets from the unaligned probe face image and the learnt face template. Geometric verifications are performed to eliminate spurious matches with inconsistent poses.

Experimental results demonstrate that the proposed fully automatic face recognition framework not only speeds up the local feature based approach for face recognition, but also improves the recognition accuracy over the holistic approach, local feature based approach, and their parallel combination.

Chapter 6

Conclusions

The topic of this PhD dissertation is face recognition, and my focus is the development of multi-scale componential features for face identification. In this conclusion chapter, I first summarize my research works on the area of face identification. I then conclude the achieved principle contributions in Section 6.1. Finally Section 6.2 overviews the possible future directions of this research work.

In the beginning of this dissertation, I review the background of various face recognition algorithms in Chapter 2. From the literature review, I find that although the holistic approaches are robust and efficient, they require a preprocessing procedure to normalize the face image variations in pose, scale and illumination. This is not an easy task because it depends on the accurate detection of at least two landmarks from the face image. And the recognition performance deteriorates considerably if the manual process is replaced by an automatic landmark detection algorithm. As a result, most approaches work on the normalized face images based on the manually identified landmarks, which makes these approaches semi-automatic. Moreover, global features are sensitive to variations in facial expressions, poses and occlusions. Another intrinsic problem of all holistic approaches is their dependence to the training databases because knowledge about the face discrimination is generalized by machine learning from the face samples. A representative training database is necessary, which, however, is not available in many applications.

In contrast to holistic methods, local feature/component based approaches have the potential of more robust to variations in pose, scale, expression and occlusion. EBGM, AAM and SIFT fall into this category. However, the performances of both EBGM and AAM depend on a good selection of facial landmarks, which are often annotated manually. This makes the two approaches semi-automatic and labor consuming. The SIFT approach can detect distinct local structures from images and select appropriate scales to describe them automatically. It overcomes the limitations of holistic approaches and some local feature based approaches. However, there are still many outstanding issues and problems that need to be addressed and circumvented if I are to leverage the idea of SIFT and some of its good properties to solve the challenging face recognition problem.

To overcome the limitations in the original SIFT algorithm, I first analyze the deficiencies of the feature extraction part of the SIFT algorithm in deep. Chapter 3 presents new approaches to keypoint detection, partial descriptor description and insignificant keypoint removal specifically for human face images, a type of non-rigid and smooth visual objects. There are few obvious blobs and corners with high contrast, because the intensity changes in face images are gradual and slow in the most areas. On the other hand, the shape of the structures could be complex and some structures are close to each other or overlap. As a result, many local structures in the smooth area such as forehead, cheeks and chin cannot be detected, which makes the number of detected keypoints not sufficient for the subsequent recognition tasks. To solve this problem, I propose a new keypoint detection approach which can capture the information of many facial structures in the smooth area such as forehead, cheeks and chin. A partial descriptor is designed to represent the keypoints whose support areas exceed the face images. My proposed detection approach and my partial descriptor strategy produce a rich number of keypoints. I further propose some strategies to remove insignificant keypoints.

Chapter 4 introduces the design of my keypoint matching strategy. To determine the identity of a probe face image based on a set of gallery images, local structures of the

probe image represented by the keypoints and their descriptors are compared with those in the gallery. The gallery image whose local structures have the maximum similarity to the probe image establishes the identity of the probe image. In the identification tasks, there are many similar gallery images. As a result, the nearest keypoints to the probe often disperse to many candidates in the gallery, and hence the probability that the largest number of the nearest keypoints fall into the right candidate is low. This problem becomes severe if the gallery contains a large number of subjects. Moreover, multiple templates per subject in the gallery make it even worse. In this chapter, firstly I describe my keypoint matching strategies which circumvent the problem that local structures matched with those in probe disperse into many different gallery images. Secondly, a training procedure is developed for multiple samples per subject. It contains template selection, template synthesis and unstable keypoint removal to meet different requirements in face recognition applications.

I use the multi-scale local feature extraction and matching framework (LFEM) introduced in Chapters 3 and 4 to develop a fully automatic face recognition framework based on both the local and global features (FAFF) in Chapter 5. Firstly, I propose a face alignment approach based on the multi-scale local features. Many face alignment approaches are dependent on the localization of semantic facial features such as eyes, corners of mouth, nostrils and so on, which is not an easy task. In real applications, the appearances of these semantic facial components are not distinct or missing. To solve this problem, I propose to align face images based on non-semantic multi-scale local features instead of semantic features. Given an unaligned face image resulting from a face detector and a set of aligned face images in the data set, I build an automatic transformation mechanism, under which the unaligned face image can be precisely aligned for the recognition process. Then the face images after alignment are used as the input of the holistic approach ERE, which retrieves some candidate gallery images from the

whole gallery set. Finally, I perform the local approach LFEM on the narrowed gallery set composed of the selected gallery images. The reduction in the size of the gallery set relieves the computational burden of the local feature based method. My FAFF approach not only speeds up the LFEM approach, but also achieve better recognition performance than LFEM, ERE, and their parallel combination.

On AR database (image size 60×85 , the average computational time for the ERE approach (one of the best holistic approaches) is $0.4052s$, and its recognition rate is 95.05% on well-aligned images. The average computational time for the LFEM approach is around $11.027s = 8.7295s$ (keypoint detection)+ $2.2975s$ (matching) with a 99.05% recognition rate on images without alignment, on a laptop with an Intel(R) Core(TM) *i7* – 2640M CPU @ 2.8GHz processor and 4GB RAM. The advantages of LFEM over holistic approaches are:

- (i) There is no training process in the LFEM approach, which avoids the generalization problem in the holist approaches;
- (ii) The image alignment procedure is not needed in the LFEM approach;
- (iii) The recognition performance is better than the holistic approach.

The only concern of the LFEM approach is the computational time. The LFEM approach is superior to the ERE approach, if the recognition performance is more emphasized.

6.1 Principle Contributions

To summarize, the main contributions of this dissertation are listed as follows.

- Multi-scale local feature extraction and description. I propose approaches to detect and describe multi-scale local features specifically in face images. To extract missing

structures in the original SIFT approach, I propose to compare each candidate point to its 8 neighbors in the current scale and corresponding one in the scale above and below. In other words, I compare it to 10 neighbors instead of 26. The partial-descriptor is proposed to describe keypoints at large scales or near image boundaries, whose support areas exceed the image size. In Chapter 3, I give the details and experimental results of my approaches.

- A two-stage image matching scheme and a strategy of keypoint search for the nearest subject are developed in Chapter 4 to cater for the identification task. In the first stage, I search the k -nearest neighbors of the nearest subject, where k is determined by the second nearest subject to find a set of candidate gallery image. In the second step, I re-match the probe keypoints and keypoints in each individual candidate gallery image to augment the matched keypoint set. The final similarity scores are computed between the probe image and the candidate gallery images based on the accumulated keypoint similarities. It circumvents the problem that the most similar local structures to the probe disperse to many different gallery images.
- A fully automatic face recognition framework is introduced in Chapter 5. It includes automatic face alignment, and a cascaded face recognition framework combining the holistic approach ERE and my proposed multi-scale local feature extraction and matching approaches.

6.2 Future Directions

The computational burden of holistic approaches for face recognition is less than the local feature based approaches. However, the local feature based approaches can overcome some limitations in the holistic approaches. The future directions of my research can be categorized as below:

- Discriminative local feature extraction and selection. The number of keypoints extracted by my approach is sufficient for the recognition, but some of them could be redundant. How to extract discriminative and representative local features is still an open question.
- More sophisticated classification methods. In the future, I could apply or develop more sophisticated classification methods which utilized machine learning.
- Real-world face recognition. I could extend my approaches to solve real-world face recognition problems.

References

- [1] G. Shakhnarovich and B. Moghaddam, “Face recognition in subspaces,” in *Handbook of Face Recognition*, S. Li and A.K. Jain, Eds., pp. 141–168. Springer, 2004.
- [2] T. Ahonen, A. Hadid, and M. Pietikainen, “Face recognition with local binary patterns,” *European Conference on Computer Vision*, pp. 469–481, 2004.
- [3] T. Tuytelaars and K. Mikolajczyk, *Local invariant feature detectors: a survey*, Now Publishers Inc., Hanover, MA, USA, 2008.
- [4] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 1, no. 60, pp. 63–86, 2004.
- [5] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *British Machine Vision Conference*, pp. 384–393, 2002.
- [6] T. Tuytelaars and L. Van Gool, “Wide baseline stereo matching based on local, affinely invariant regions,” *British Machine Vision Conference*, pp. 412–425, 2000.
- [7] T. Tuytelaars and L. Van Gool, “Matching widely separated views based on affine invariant regions,” *International Journal of Computer Vision*, vol. 1, no. 59, pp. 61–85, 2004.
- [8] T. Tuytelaars and L. Van Gool, “Content-based image retrieval based on local affinely invariant regions,” *International Conference on Visual Information Systems*, pp. 493–500, 1999.
- [9] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

- [10] D. G. Lowe, “Object recognition from local scale-invariant features,” *IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, 1999.
- [11] W. Zhan, R. Chellappa, P.J Phillips, and A. Rosenfeld, “Face recognition: a literature survey,” *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.
- [12] M.D. Kelly, *Visual identification of people by computer*, Ph.D. thesis, Stanford, CA, USA, 1971.
- [13] T. Kanade, “Computer recognition of human faces,” *Interdisciplinary Systems Research*, vol. 47, 1977.
- [14] B.S. Manjunath, R. Chellappa, and C. von der Malsburg, “A feature based approach to face recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 373–378, 1992.
- [15] A.K. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, pp. 4–20, 2004.
- [16] M.A. Turk and A.P. Pentland, “Face recognition using eigenfaces,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.
- [17] I.T. Jolliffe, *Principal component analysis*, Springer, second edition, October 2002.
- [18] M.M loève, *Probability theory*, Princeton, Van Nostrand, 1955.
- [19] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.
- [20] D.Q. Dai and P.C. Yuen, “Regularized discriminant analysis and its application to face recognition,” *Pattern Recognition*, vol. 36, pp. 845–847, 2003.
- [21] X.G. Wang and X.O. Tang, “Dual-space linear discriminant analysis for face recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 564–569, 2004.

- [22] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, Jan. 2005.
- [23] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.
- [24] J.W. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 195–200, Jan. 2003.
- [25] W. Liu, Y.H. Wang, S.Z. Li, and T.N. Tan, "Null space approach of fisher discriminant analysis for face recognition," in *In Proc. European Conference on Computer Vision, Biometric Authentication Workshop*, 2004, pp. 32–44.
- [26] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.
- [27] J. Yang and J.Y. Yang, "Why can lda be performed in pca transformed space?," *Pattern Recognition*, vol. 36, pp. 563–566, 2003.
- [28] X.G. Wang and X.O. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.
- [29] X.F. He, S.C. Yan, Y.X. Hu, P. Niyogi, and H.J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.
- [30] X.F. He and P. Niyogi, "Locality preserving projections," *Proc. Conf. Advances in Neural Information Processing Systems*, 2003.
- [31] S. Noushath, A. Rao, and G.H. Kumar, "Mixture-of-laplacian faces and its application to face recognition," in *Proceedings of the 2nd International Conference on Pattern Recognition and Machine Intelligence*, 2007, pp. 568–575.

- [32] N. Wang and X. Li, “Semi-supervised laplacianfaces from pairwise constraints for face recognition,” *International Conference on Intelligent Systems Design and Applications*, vol. 2, pp. 517–522, 2008.
- [33] D. Cai, X.F. He, J.W. Han, and H.J. Zhang, “Orthogonal laplacianfaces for face recognition,” *IEEE Trans. on Image Processing*, pp. 3608–3614, 2006.
- [34] B. Niu, Q. Yang, Shiu S.C.K, and S.K. Pal, “Two-dimensional laplacianfaces method for face recognition,” *Pattern Recognition*, vol. 41, pp. 3237–3243, October 2008.
- [35] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: a general framework for dimensionality reduction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
- [36] X. D. Jiang, B. Mandal, and A. Kot, “Eigenfeature regularization and extraction in face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, March 2008.
- [37] J.Y. Choi, W.D. Neve, K.N. Plataniotis, and Y.M Ro, “Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks,” *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 14–28, Feb. 2011.
- [38] B. Moghaddam and A. Pentland, “Probabilistic visual learning for object representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, Jul. 1997.
- [39] B. Moghaddam, T. Jebara, and A. Pentland, “Bayesian face recognition,” *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, Nov. 2000.
- [40] B. Moghaddam, “Principal manifolds and bayesian subspace for visual recognition,” *IEEE International Conference on Computer Vision*, vol. 2, no. 2, pp. 1131–1136, 1999.
- [41] B. Moghaddam, “Principal manifolds and probabilistic subspace for visual recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 780–788, Jun. 2002.

- [42] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electronics Letters*, vol. 42, no. 19, pp. 1089–1090, September 2006.
- [43] P.J. Phillips, "Support vector machines applied to face recognition," in *Advances in Neural Information Processing Systems*, 1999, pp. 803–809.
- [44] K. Jonsson, J. Matas, J. Kittler, and Y.P. Li, "Learning support vectors for face verification and recognition," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 208–213, 2000.
- [45] G.D. Guo, S.Z. Li, and K. Chan, "Face recognition by support vector machines," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 196–201, 2000.
- [46] G. Cui and W. Gao, "Support vector machines for face recognition with two-layer generated virtual data," *IEEE International Conference on Pattern Recognition*, vol. 3, pp. 570–573, Aug. 2004.
- [47] V.N. Vapnik, *Statistical learning theory*, Wiley-Interscience, Sep. 1998.
- [48] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
- [49] Q. Yang and X.O. Tang, "Recent advances in subspace analysis for face recognition," *Advances in Biometric Person Authentication*, pp. 275–287, 2005.
- [50] X.Y. Tan, S.C.Chen, Z.H. Zhou, and F.Y. Zhang, "Face recognition from a single image per person: a survey," *Pattern Recognition*, vol. 39, pp. 1725–1745, 2006.
- [51] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Transaction on Image Processing*, vol. 16, no. 10, pp. 2617–2628, Oct. 2007.
- [52] W.W. Bledsoe, "The model method in facial recognition," Tech. Rep., Panoramic Research Inc., Palo Alto, California, 1964.

- [53] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree. K. Nayar, “Describable visual attributes for face verification and image search,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue on Real-World Face Recognition*, Oct 2011.
- [54] N. Kumar, “Attribute face service,” in <http://afs.automaticfacesystems.com/>.
- [55] C.C. Chang and C.J. Lin, “Libsvm: a library for support vector machines,” in <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.
- [56] S. Marčelja, “Mathematical description of the responses of simple cortical cells,” *Journal of the Optical Society of America*, vol. 70, no. 11, pp. 1297–1300, 1980.
- [57] B.S. Manjunath, R. Chellappa, and C. von der Malsburg, “A feature based approach to face recognition,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 373–378, Jun. 1992.
- [58] L. Wiskott, J.M. Fellous, N. Kuiger, and C. von der Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, Jul. 1997.
- [59] B. Kepenekci, F.B. Tek, and G.B. Akar, “Occluded face recognition based on gabor wavelets,” *IEEE International Conference on Image Processing*, vol. 1, pp. 293–296, 2002.
- [60] C.J. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
- [61] D. Gabor, “Theory of communication,” *J. IEE*, vol. 3, no. 93, pp. 429–457, 1946.
- [62] D. Bolme, “Elastic bunch graph matching,” M.S. thesis, Colorado State University, May 2003.

- [63] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, “Active shape models - their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.
- [64] A. Witkin, M. Kass, and D. Terzopoulos, “Snakes: active contour models,” *International Conference on Computer Vision*, pp. 259–268, 1987.
- [65] T.F. Cootes, G.J. Edwards, and C.J. Taylor, “Active appearance models,” *ECCV*, 1998.
- [66] T. Ojala, M. Pietikainen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996.
- [67] G.C. Zhang, X.S. Huang, S.Z. Li, Y.S. Wang, and X.H. Wu, “Boosting local binary pattern (lbp)-based face recognition,” *Advances in Biometric Person Authentication, Lecture Notes in Computer Science*, vol. 3338, pp. 179–186, 2005.
- [68] T. Ahonen, M. Pietikainen, A. Hadid, and T. Maenpaa, “Face recognition based on the appearance of local regions,” *International Conference on Pattern Recognition*, vol. 3, pp. 23–26, Aug. 2004.
- [69] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: application to face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Oct. 2006.
- [70] W.C. Zhang, S.G. Shan, W. Gao, X.L. Chen, and H.M. Zhang, “Local aabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition,” *International Conference on Computer Vision*, vol. 1, pp. 786–791, Oct. 2005.
- [71] C.H. Chan, J. Kittler, and K. Messer, “Multi-scale local binary pattern histograms for face recognition,” *Advances in Biometrics, Lecture Notes in Computer Science*, vol. 4642, pp. 809–818, 2007.

- [72] C. Schmid, R. Mohr, and C. Bauckhage, “Evaluation of interest point detectors,” *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.
- [73] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *International Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.
- [74] S.M. Smith and J.M. Brady, “Susan - a new approach to low level image processing,” *International Journal of Computer Vision*, vol. 23, no. 34, pp. 45–78, 1997.
- [75] E. Rosten and T. Drummond, “Fusing points and lines for high performance tracking,” in *International Conference on Computer Vision*, 2005, pp. 1508–1511.
- [76] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision*, 2006, pp. 430–443.
- [77] J.R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, pp. 81–106, 1986.
- [78] V. Lepetit and P. Fua, “Keypoint recognition using randomized trees,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1465–1479, 2006.
- [79] C. Harris and M. Stephens, “A combined corner and edge detector,” *Alvey Vision Conference*, pp. 147–151, 1988.
- [80] T. Lindeberg, “Feature detection with automatic scale selection,” *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.
- [81] T. Lindeberg, “Direct estimation of affine image deformation using visual front-end operations with automatic scale selection,” *International Conference on Computer Vision*, pp. 134–141, 1995.
- [82] T. Lindeberg and J. Garding, “Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure,” *Image and Vision Computing*, vol. 15, no. 6, pp. 415–434, 1997.

- [83] J. Canny, “A computational approach to edge detection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.
- [84] R.M. Haralick and L.G. Shapiro, “Computer and robot vision,” *Addison-Wesley*, pp. 453–507, 1993.
- [85] K. Mikolajczyk, *Scale and affine invariant interest point detectors*, Phd thesis, INRIA Grenoble, 2002.
- [86] J.L. Crowley and A.C. Parker, “A representation for shape based on peaks and ridges in the difference of low pass transform,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 2, pp. 156–170, 1984.
- [87] P. Gaussier and J.P. Cocquerez, “Neural networks for complex scene recognition: Simulation of a visual system with several cortical areas,” *International Joint Conference on Neural Networks*, vol. 3, no. 10-11, pp. 233–259, 1992.
- [88] S. Grossberg, E. Mingolla, and D. Todorovic, “A neural network architecture for preattentive vision,” *IEEE Transaction on Biomedical Engineering*, vol. 36, pp. 65–84, 1989.
- [89] P. Vilola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, 2001.
- [90] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: speeded up robust features,” *European Conference on Computer Vision*, pp. 404–417, 2006.
- [91] G. Mori, X. Ren, A. Eefros, and J. Malik, “Recovering human body configurations: Combining segmentation and recognition,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 326–333.
- [92] X. Ren and J. Malik, “Learning a classification model for segmentation,” *International Conference on Computer Vision*, pp. 10–17, 2003.

- [93] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [94] X. Ren, C. Fowlkes, and J. Malik, “Scale-invariant contour completion using conditional random fields,” *International Conference on Computer Vision*, vol. 2, pp. 1214–1221, 2005.
- [95] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” *European Conference on Computer Vision*, pp. 490–503, 2006.
- [96] M. Brown and D. G. Lowe, “Recognising panoramas,” *IEEE International Conference on Computer Vision*, vol. 2, pp. 1218 –1225, Oct. 2003.
- [97] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 264–271, 2003.
- [98] A. Johnson and M. Hebert, “Object recognition by matching oriented points,” *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 684–689, 1997.
- [99] S. Lazebnik, C. Schmid, and J. Ponce, “Sparse texture representation using affine-invariant neighborhoods,” *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 319–324, 2003.
- [100] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” *Proc. Third European Conf. Computer Vision*, pp. 151–158, 1994.
- [101] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [102] Y. Ke and R. Sukthankar, “Pca-sift: a more distinctive representation for local image descriptors,” *International Conference on Computer Vision and Pattern Recognition*, pp. 511–517, 2004.

- [103] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [104] A. Ashbrook, N. Thacker, P. Rockett, and C. Brown, “Robust recognition of scaled shapes using pairwise geometric histograms,” *Proc. Sixth British Machine Vision Conf.*, pp. 503–512, 1995.
- [105] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 4, pp. 509–522, Apr. 2002.
- [106] J. Koenderink and A. Van Doorn, “Representation of local geometry in the visual system,” *Biological Cybernetics*, vol. 55, pp. 367–375, 1987.
- [107] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever, “General intensity transformations and second order invariants,” *Proc. Seventh Scandinavian Conf. Image Analysis*, pp. 338–345, 1991.
- [108] W. Freeman and E. Adelson, “The design and use of steerable filters,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.
- [109] A. Baumberg, “Reliable feature matching across widely separated views,” *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 774–781, 2000.
- [110] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets,” *Proc. Seventy European Conf. Computer Vision*, pp. 414–431, 2002.
- [111] L. Van Gool, T. Moons, and D. Ungureanu, “Affine/photometric invariants for planar intensity patterns,” *Proc. Fourth European Conf. Computer Vision*, pp. 642–651, 1996.
- [112] D. G. Lowe, “Local feature view clustering for 3d object recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 682–688, 2001.
- [113] M. Brown and D. G. Lowe, “Invariant features from interest point groups,” in *British Machine Vision Conference*, 2002, pp. 656–665.

- [114] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of sift features for face authentication," *Workshop on Computer Vision and Pattern Recognition*, pp. 35–40, Jun. 2006.
- [115] D. H. Lin and X. O. Tang, "Recognize high resolution faces: from macrocosm to microcosm," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1355–1362, 2006.
- [116] D. R. Kisku, A. Rattani, E. Grosso, and M. Tistarelli, "Face identification by sift-based complete graph topology," *IEEE Workshop on Automatic Identification Advanced Technologies*, pp. 63–68, Jun. 2007.
- [117] C. Rosenberger and L. Brun, "Similarity-based matching for face authentication," *IEEE International Conference on Pattern Recognition*, pp. 1–4, Dec. 2008.
- [118] D. R. Kisku, A. Rattani, M. Tistarelli, and P. Gupta, "Graph application on face for personal authentication and recognition," *International Conference on Control, Automation, Robotics and Vision*, pp. 1150–1155, Dec. 2008.
- [119] L. C. Zhang, J. Chen, Y. Lu, and P. Wang, "Face recognition using scale invariant feature transform and support vector machine," *International Conference for Young Computer Scientists*, pp. 1766–1770, Nov. 2008.
- [120] C. Fernandez and M. A. Vicente, "Face recognition using multiple interest point detectors and sift descriptors," *IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–7, Sept. 2008.
- [121] C. Cruz, L. E. Sucar, and E. F. Morales, "Real-time face recognition for human-robot interaction," *IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6, Sept. 2008.
- [122] Y. B. Han, J. Q. Yin, and J. P. Li, "Human face feature extraction and recognition base on sift," *International Symposium on Computer Science and Computational Technology*, vol. 1, pp. 719–722, Dec. 2008.

- [123] D. R. Kisku, M. Tistarelli, J. K. Sing, and P. Gupta, "Face recognition by fusion of local and global matching scores using ds theory: an evaluation with uni-classifier and multi-classifier paradigm," *IEEE Workshop on Computer Vision and Pattern Recognition*, pp. 60–65, Jun. 2009.
- [124] A. Majumdar and R. K. Ward, "Discriminative sift features for face recognition," *Canadian Conference on Electrical and Computer Engineering*, pp. 27–30, May 2009.
- [125] J. Wright and G. Hua, "Implicit elastic matching with random projections for pose-variant face recognition," *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2009.
- [126] L. Bai and L. Shen, "A fast and robust gabor feature based method for face recognition," *The IEEE International Symposium on Imaging for Crime Detection and Prevention*, pp. 95–98, Jun. 2005.
- [127] B.C. Zhang, "Gabor-kernel fisher analysis for face recognition," *Advances in Multimedia Information Processing*, pp. 802–809, 2005.
- [128] E. Bailly-baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethhoz, J. Matas, K. Messer, F. Poree, and B. Ruiz, "The banca database and evaluation protocol," in *In Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA03)*. 2003, pp. 625–638, Springer-Verlag.
- [129] S. Winder and M. Brown, "Learning local image descriptors," *International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [130] R. Gross, "Face databases," in *Handbook of Face Recognition*, S. Li and A.K. Jain, Eds. Springer, New York, February 2005.
- [131] P.J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The feret evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

- [132] F. Samaria and A. Harter, “Parameterisation of a stochastic model for human face identification,” *2nd IEEE Workshop on Applications of Comp. Vision*, pp. 138–142, Dec. 1994.
- [133] A.V. Nefian, “Embedded bayesian networks for face recognition,” *ICME*, pp. 133–136, Aug. 2002.
- [134] A.M. Martinez and R. Benavente, “Cvc technical report #24,” Tech. Rep., Jun. 1998.
- [135] X. Tan, S. Chen, Z. Zhou, and F. Zhang, “Face recognition from a single image per person: a survey,” *Pattern Recognition*, vol. 39, pp. 1725–1745, 2006.
- [136] C. Geng and X. D. Jiang, “Face recognition using sift features,” *IEEE International Conference on Image Processing*, pp. 3313–3316, Nov. 2009.
- [137] C. Geng and X. D. Jiang, “Sift features for face recognition,” *IEEE International Conference on Computer Science and Information Technology*, pp. 598–602, Aug. 2009.
- [138] G. Hua, M. Brown, and S. Winder, “Discriminant embedding for local image descriptors,” *IEEE International Conference on Computer Vision*, pp. 1–8, Dec. 2007.
- [139] M. Brown, G. Hua, and S. Winder, “Discriminative learning of local image descriptors,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 43–57, 2011.
- [140] T. Lindeberg, “Scale-space for discrete signals,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 234–254, 1990.
- [141] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, “The csu face identification evaluation system user’s guide: version 5.0,” technical report, 2003.
- [142] “Data description,” in http://www.robots.ox.ac.uk/~vgg/research/affine/det_eval_files/DataREADME.

- [143] M.A. Fischler and R.C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [144] P.H.S. Torr and A. Zisserman, “Mlesac - a new robust estimator with application to estimating image geometry,” *Computer Vision and Image Understanding*, vol. 78, pp. 138–156, 2000.
- [145] P.H.S. Torr and C. Davidson, “Impsac: synthesis of importance sampling and random sample consensus,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 354–364, 2003.
- [146] B.J. Tordoff and D.W. Murray, “Guided-mlesac: faster image transform estimation by using matching priors,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1523–1535, 2005.
- [147] O. Chum and J. Matas, “Matching with prosac - progressive sample consensus,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 220–226, 2005.
- [148] J.L. Bentley, “Multidimensional binary search trees used for associative searching,” *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.
- [149] X.D. Jiang, “Linear subspace learning-based dimensionality reduction,” *Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, March 2011.
- [150] M. Kirby and L. Sirovich, “Application of karhunen-loeve procedure for the characterization of human faces,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, January 1990.
- [151] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, “Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition,” *Pattern Recognition Letters*, vol. 26, pp. 181–191, 2005.

- [152] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, and Z. Jin, “Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, February 2005.
- [153] W. Zheng and X. Tang, “Fast algorithm for updating the discriminant vectors of dual-space lda,” *IEEE Trans. Information Forensics And Security*, vol. 4, no. 3, pp. 418–427, September 2009.
- [154] J. Ye, R. Janardan, C. Park, and H. Park, “An optimization criterion for generalized discriminant analysis on undersampled problems,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982–994, August 2004.
- [155] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, “Graph embedding and extensions: a general framework for dimensionality reduction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, January 2007.
- [156] T. Riopka and T. Boulton, “The eyes have it,” in *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, 2003, pp. 9–16.
- [157] D.T. Lin and C.M. Yang, “Real-time eye detection using face circle fitting and dark-pixel filtering,” *IEEE International Conference on Multimedia and Expo*, pp. 1167–1170, 2004.
- [158] G. Li, X. Cai, X. Li, and Y. Liu, “An efficient face normalization algorithm based on eye detection,” *IEEE International Conference on Intelligent Robots and Systems*, pp. 3843–3848, 2006.
- [159] Y.W. Chen and K. Kubo, “A robust eye detection and tracking technique using gabor filters,” *IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, vol. 1, pp. 109–112, 2007.
- [160] C.W. Park, J.M. Kwak, H. Park, and Y.S. Moon, “An effective method for eye detection based on texture information,” *IEEE International Conference on Convergence Information Technology*, pp. 586–589, 2007.

- [161] J.F. Ren and X.D. Jiang, “Fast eye localization based on pixel differences,” *IEEE International Conference on Image Processing*, pp. 2733–2736, 2009.
- [162] J.F. Ren and X.D. Jiang, “Eye detection based on rank order filter,” *IEEE International Conference on Information, Communication and Signal Processing*, pp. 1–4, Dec 2009.
- [163] T. D’Orazio, M. Leo, G. Cicirelli, and A. Distante, “An algorithm for real time eye detection in face images,” *International Conference on Pattern Recognition*, vol. 3, pp. 278–281, 2004.
- [164] J.W. Wu and M.M. Trivedi, “A binary tree for probability learning in eye detection,” *Workshop on Computer Vision and Pattern Recognition*, pp. 170–177, 2005.
- [165] X.D. Jiang J.F. Ren and J. Yuan, “A fast and accurate cascade subspace face/eye detector on mobile devices,” *13th International Conference on Computer Vision, IEEE International Workshop on Mobile Vision*, pp. 84–91, Nov. 2011.
- [166] T.L. Berg, A.C. Berg, M. Maire, R. White, Y.W. Teh, E. Learned-Miller, and D.A. Forsyth, “Names and faces in the news,” *CVPR*, 2004.
- [167] M. Everingham, J. Sivic, and A. Zisserman, ““Hello! My name is... Buffy” – automatic naming of characters in TV video,” *BMVC*, pp. 889–908, Sep. 2006.
- [168] S.Z. Li, S.C. Yan, H. Zhang, and Q. Cheng, “Multi-view face alignment using direct appearance models,” *IEEE Int. Conf. on Aut. Face and Gesture Recog.*, 2002.
- [169] Y. Zhou, L. Gu, and H.J. Zhang, “Bayesian tangent shape model: estimating shape and pose parameters via bayesian inference,” *CVPR*, 2003.
- [170] Y. Zhou, W. Zhang, X. Tang, and H. Shum, “A bayesian mixture model for multi-view face alignment,” *CVPR*, 2005.
- [171] G.B. Huang, V. Jain, and E. Learned-Miller, “Unsupervised joint alignment of complex images,” *ICCV*, pp. 1–8, Dec. 2007.

- [172] C. Geng and X.D. Jiang, “Face recognition based on the multi-scale local image structures,” *Pattern Recognition*, vol. 44, no. 10-11, pp. 2565–2575, Oct.-Nov. 2011.
- [173] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B. L. Lu, “Person-specific sift features for face recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 593–596, Apr. 2007.
- [174] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: application to face recognition,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, pp. 2037–2041, 2006.
- [175] A. Hadid, J.Y. Heikkilä, O. Silver, and M. Pietikäinen, “Face and eye detection for person authentication in mobile phones,” *IEEE Int. Conf. Distributed Smart Cameras*, pp. 101–108, 2007.
- [176] S.U. Jung, Y.S. Chung, J.H. Yoo, and K.Y. Moon, “Real-time face verification for mobile platforms,” *Advances in visual computing*, pp. 823–832, 2008.
- [177] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2009.
- [178] X. D. Jiang, “Asymmetric principal component and discriminant analyses for pattern classification,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 5, pp. 931–937, May 2009.
- [179] A. Ross and A. Jain, “Information fusion in biometrics,” *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, September 2003.
- [180] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, “Liblinear: a library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, August 2008.
- [181] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2161–2168, 2006.

List of Publications

Journal

1. Cong Geng and Xudong Jiang. Face Recognition based on the Multi-scale Local Image Structures. *Pattern Recognition*, vol. 44, no. 10-11, pp. 2565-2575, Oct.-Nov., 2011.
2. Cong Geng and Xudong Jiang. Fully Automatic Face Recognition Framework based on Local and Global Features. *Machine Vision and Applications*, Online published, DOI: 10.1007/s00138-012-0423-7, 2012.

Conference

1. Cong Geng and Xudong Jiang. SIFT Features for Face Recognition. *In Proceeding of IEEE International Conference on Computer Science and Information Technology (ICCSIT'09)*, Aug., 2009.
2. Cong Geng and Xudong Jiang. Face Recognition Using SIFT Features. *In Proceeding of IEEE International Conference on Image Processing (ICIP'09)*, Nov., 2009.
3. Cong Geng and Xudong Jiang. Face Alignment Based on The Multi-Scale Local Features. *In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*, Mar., 2012.