Illumination Invariant Face Recognition Based on Active Near-Infrared Differential Imaging

X. ZOU

Submitted for the Degree of Doctor of Philosophy from the University of Surrey

Centre for Vision, Speech and Signal Processing School of Electronics and Physical Sciences University of Surrey Guildford, Surrey GU2 7XH, U.K.

March 2007

© X. ZOU 2007
Summary

Changes in the illumination condition cause dramatic variation in face appearance and seriously affect the performance of face recognition systems. This problem is addressed in the thesis by introducing an approach based on active Near-Infrared differential imaging.

Assuming a static scene, a linear response of the sensor to the scene radiation and no saturation, it is shown theoretically and empirically in this thesis that the active differential imaging technique yields a face image independent of the change in ambient illumination. By taking the difference of two face images, one captured with the active illumination on and one with it off, the resulting image contains the face illuminated only by the active illumination source. This technique is compared with several representative illumination invariant face recognition techniques on a database containing faces captured under different illuminations and at different time. The results in face identification and verification experiments demonstrate the significant advantage of this Near-Infrared differential imaging technique over the other techniques.

This thesis also presents a multistage approach to automatic face localisation for the Near-Infrared face images. This multistage approach is a combination of a novel pupil detection approach based on edge following and chaincode representation, and an approach based on FloatBoost learning. Accurate face localisation results are achieved by the proposed multistage approach, and this leads to the excellent face recognition performance in fully automatic scenario.

A subject appears at two different locations before and after the active illumination is turned on if he/she is moving. This causes motion artifact in the difference image from the active differential imaging system and degrades the performance of the face recognition system. The thesis presents an approach based on motion compensation to deal with this problem. It is shown from the experimental results that the proposed approach successfully removes the motion artifacts and improves the face recognition performance significantly.

Key words: Biometrics, Face Recognition, Illumination, Near-Infrared, Active Differential Imaging, Face Localisation, Motion Compensation.
Acknowledgements

I would like to take this opportunity to express my deep gratitude to my supervisor Professor Josef Kittler for offering me the opportunity to conduct this research and providing invaluable guidance and support. I am also in debt to Dr. Kieron Messer for all the helpful suggestions and support.

I am grateful to my colleagues who have been helping me a lot with my research. Especially I want to say thanks to Dr. Bill Christmas, Dr. Ong Eng Jon, Dr. Miroslav Hamouz, Dr. James Short, Dr. JeanYves Guillemaut, Chiho Chan, Ning Jin, Fei Yan, Mathew Prior. And thanks to my flatmates and friends.

I would like to thank my wife Mrs Wei CAO, my parents and all other family members for all your love, understanding, support and encouragement.

The support from Overseas Research Students Awards Scheme is gratefully acknowledged.
## Contents

1 Introduction ................................................. 1
   1.1 Motivation ............................................... 1
      1.1.1 Biometrics ....................................... 1
      1.1.2 Face Recognition ........................... 2
      1.1.3 The Illumination Problem ............... 4
   1.2 Contributions ........................................ 7
   1.3 Overview of Thesis ................................. 8

2 Face Recognition ............................................. 11
   2.1 Face Recognition System ......................... 11
   2.2 Subspace Approaches .............................. 14
      2.2.1 Principle Component Analysis .......... 14
      2.2.2 Linear Discriminant Analysis ........... 15
      2.2.3 Others .......................................... 17
   2.3 Classifiers ............................................. 17
      2.3.1 Nearest Neighbor ......................... 17
      2.3.2 Support Vector Machine ................. 17
   2.4 Performance Evaluation of Face Recognition 18
      2.4.1 Verification Test and Performance Evaluation 19
      2.4.2 Identification Test and Performance Evaluation 20
   2.5 Conclusion ............................................ 21
### 3 Literature Review on Illumination-Invariant Face Recognition

3.1 Introduction ................................................. 23

3.2 Passive Approaches ........................................ 23
   3.2.1 Illumination Variation Modelling ................. 24
   3.2.2 Illumination Invariant Features .................. 29
   3.2.3 Photometric Normalization ......................... 35
   3.2.4 3D Morphable Model ............................... 38

3.3 Active Approaches ......................................... 38
   3.3.1 3D information ..................................... 39
   3.3.2 Infrared ........................................... 40

3.4 Conclusion .................................................. 45

### 4 Near-IR Active Differential Imaging for Face Recognition

4.1 Introduction .................................................. 47

4.2 Near-IR Active Differential Imaging ..................... 48

4.3 Image Formation Analysis for Active Differential Imaging ................. 49
   4.3.1 Radiometry Analysis ............................... 50
   4.3.2 Sensor Model ...................................... 52
   4.3.3 Near-IR Difference Image ....................... 52

4.4 Active Differential Imaging for Face Recognition ....................... 53

4.5 Capture System and Face Database Acquisition ..................... 56

4.6 Face Recognition Experiments on Manually Registered Data .......... 62
   4.6.1 Face Identification Experiment ................... 62
   4.6.2 Face Verification Experiments .................... 66

4.7 Conclusion .................................................. 76

### 5 Literature Review on Face Localisation

5.1 Introduction ................................................. 77

5.2 Passive Approaches ........................................ 78
   5.2.1 Top-down Approaches ............................... 78
   5.2.2 Bottom-up Approaches ............................. 79
   5.2.3 Eye Detection/Localisation ......................... 80

5.3 Active Approaches ......................................... 81

5.4 Localisation Performance Measures ............................. 82

5.5 Conclusion .................................................. 83
List of Figures

1.1 Diagram of face identification. ............................................................ 3
1.2 Diagram of face verification. ............................................................... 4
1.3 Examples of face images of the same subject while under different illuminations in Yale-B database. ...................................................... 5
1.4 Examples of face images of the same subject while under different illuminations in PIE database. .......................................................... 6

2.1 Diagram of face recognition system. ................................................... 12
2.2 Geometric normalisation. ................................................................. 13
2.3 ROC curve. ....................................................................................... 20

3.1 The electromagnetic spectrum(reprinted from [30] with modification). .. 40

4.1 Active differential Imaging: ambient illumination is removed by taking the difference between an actively illuminated frame and an ambient frame. ................................................................. 49
4.2 Image formation. ............................................................................... 50
4.3 The spectral response of KP-M3RP camera[46] .................................. 56
4.4 A picture of face capture system for our experiment ....................... 57
4.5 Ambient faces (the left column), combined illumination faces (the middle column) and LED illuminated faces (the right column) under 4 different illumination configurations. The changes in ambient illumination caused significant differences in the appearance of the whole face. All important facial features look very different in different illumination conditions. Ambient faces and LED faces are relatively dark because the aperture of the lens is adjusted to avoid the saturation of the combined illuminated faces. .......................................................... 59
4.6 Ambient face images of two subjects for the two different sessions. ... 60
4.7 LED face images of two subjects for the two different sessions. ........ 61
4.8 Resulting images after the histogram equalization is performed for manually registered ambient faces ((a)(b)) and for corresponding LED faces ((c)(d)). It is obvious that data from LED faces exhibits much less variation as compared to the data from ambient faces. Bright pupils are prominent in LED faces. ................................. 62

4.9 Images after the photometric normalisation based on anisotropic smoothing. (a) ambient face images with lighting from four directions. (b) corresponding LED faces. ................................. 65

4.10 Face regions after cropping and histogram equalisation for images normalised photometrically by the method based on anisotropic smoothing. (a) two sessions of ambient face images of a subject with lighting from four directions, (b) corresponding LED faces ................................................ 66

4.11 The distributions of the true claim and the false claim scores for ambient faces and LED faces in the Cross Session test (a)(b), the Cross Illumination test (c)(d) and the Combined test (e)(f) based on Fisherface representation. ............................................................... 71

4.12 ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Cross Session test. .............. 73

4.13 ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Cross Illumination test. .... 74

4.14 ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Combined test. .............. 75

6.1 Near-IR illuminated face images. ................................................................. 86

6.2 Diagram of the face localisation method based on pupil edge detection. 88

6.3 (a) direction chain code definition. (b)(c) valid segments with chain code string 001232334 and 11122334. ................................................................. 90

6.4 Face images (a)(d), edge images (b)(e), detected candidate pupil edge segments (c)(f), and zoomed images (g)-(l) of eye regions for the corresponding images (a)-(f). ................................................................. 90

6.5 Face image (a), edge image (b), detected candidate pupil edge segments (c) after scaling (X1.5), rotation (25°) and translation have been performed on the face image in Figure 6.4(a). (d-f) are zoomed images of the eye regions for the corresponding images (a-c). True pupil edges are still successfully detected. ................................................................. 91

6.6 Positive training samples (a)(c) and negative samples (b)(d) for left eye SVM and right eye SVM, respectively. ........................................ 93

6.7 Positive training samples (a) and negative samples (b) for face SVM. ... 93

6.8 Diagram of the multistage face localisation method. ............................... 95

6.9 Localisation performance curves of the three approaches. ....................... 96
List of Figures

6.10 Localisation results. ................................................................. 97
6.11 Images with false detection using the first approach. ............... 97

7.1 Face images under combined illumination(a), ambient illumination(b) and their difference image(c), without any motion. ............ 102
7.2 Face images under combined illumination(a), ambient illumination(b) and their difference image(c), when the face is moving. ........ 102
7.3 Combined illumination frame (a) and ambient illumination frame (b) . 104
7.4 original face template (a), and resulting difference image when A-frame was shifted (-10, -7, -3, 0, 3, 7, 10) pixels horizontally (a) and vertically (b) from C-frame ................................................................. 104
7.5 NC score drops when A-frame shifts from C-frame horizontally (a) and vertically(b). ................................................................. 105
7.6 Illustration of our proposed approach. ....................................... 106
7.7 The $\rho$ function and its influence function $\psi$. (a)(b)Quadratic $\rho$ function and its influence function. (c)(d) Lorentzian function and its influence function (reprinted from [87]) ................................................................. 109
7.8 Illustration of the interpolation process. .................................... 111
7.9 Faces in difference images without motion compensation (a) and with motion compensation (b), and corresponding template images(c) for subjects in (a) and (b) ................................................................. 112
7.10 (a)Histogram of the relative motion in all sequences. (b)Average improvement in NC score after using motion compensation in terms of the relative motion. ................................................................. 113
List of Figures
List of Tables

4.1 Training and test set configurations for tests on manually registered data 63
4.2 Average error of Cross Session tests on manually registered data (in %) 67
4.3 Average error of the Cross Illumination tests on manually registered data (in %) 67
4.4 Average error of the Combined tests on manually registered data (in %) 68
4.5 Data set configurations for tests on manually registered data 69
4.6 No. of true claims/ false claims/ element tests 70
4.7 Average EER of the Cross Session tests on manually registered data (in %) 72
4.8 Average EER of the Cross Illumination tests on manually registered data (in %) 72
4.9 Average EER of Combined tests on manually registered data (in %) 76
6.1 Average error of identification tests on the automatically localised faces (in %) 99
6.2 Average error of identification tests on the Manually/Automatically localised faces (in %) 99
6.3 EER for verification tests on Manually/Automatically localised faces (in %) 100
6.4 EER for verification tests on automatically localised faces (in %) 100
7.1 Error Rates of Face Identification Experiments on Moving Faces (in %) 114
7.2 EERs of Verification Experiments on Moving Faces (in %) 115
List of Tables
Chapter 1

Introduction

1.1 Motivation

1.1.1 Biometrics

Nowadays the identity fraud has become a serious problem for the society. According to the last estimate from the Home Office Identity Fraud Steering Committee, the identity fraud costs UK economy 1.7 billion pounds over the last three years [50]. The cumulative loss in the United States is about 48 billion dollars in 2003 and 56.6 billion dollars in 2005 [51].

One of the main reasons for this is that traditional identity authentication is not safe. The traditional identity authentication is usually based on something you have, such as ID card, key, etc., or something you know, like password, Personal Identify Number (PIN). However, they can be easily lost, forgotten, or obtained by thefts.

Biometrics has emerged as a much more reliable solution than the traditional identity authentication approaches. Making use of the physiological or behavioral attributes of an individual, such as fingerprint, face, hand geometry, palm print, voice, iris, signature, the identity authentication based on biometrics can overcome the problem of the traditional approaches. In the past decade the biometrics authentication technique has matured and become widely accepted in various application areas, such as Access control/Attendance, Civil ID, customer ID, Criminal ID, Device/System Access, and
Chapter 1. Introduction

Surveillance. From this year, each newly issued Greece passport will have a biometric chip on it. The UK government has also announced that a biometric card will be issued with every new passport. According to the Biometric Market and Industry Report provided by the International Biometric Group (IBG), the global biometric revenues reached 1.5 billion US dollars in 2005 and are expected to experience rapid growth through 2010.

1.1.2 Face Recognition

Why Face?

According to IBG, face biometric system accounts for 19% of the whole biometrics market in 2006, which is the second largest share, surpassed only by fingerprint. There are several advantages of face over other biometric traits:

1. Non-intrusive: Compared to other biometrics, the least cooperation is needed for face biometrics. Face image capture works in a non-contact way and fast.

2. Best acceptance by clients: face is the biometrics which is exposed to other people or in the surveillance camera in everyday life, therefore it is less considered as private information. However, other biometrics like fingerprints, iris and etc. are often taken to be too private to be exposed.

3. Readable: People have been well trained to recognize a person through his/her face in everyday life. Therefore, human inspection can easily be taken as a remedy if a biometric system fails. In contrast, people would need more effort, or even need to be trained, to recognise other biometrics traits.

4. Easy to attain: People's faces are available in their ID card, photo album, surveillance recordings and etc. In contrast, the other biometric traits are difficult to find in everyday life.
1.1. Motivation

Figure 1.1: Diagram of face identification.

**Face Recognition: Tasks**

Depending on the application scenario, the face recognition can be formulated as one of the two tasks: *Face Identification* and *Face Verification*:

1. **Face Identification**: Given an image or a video sequence containing a face of a person, identify who this person is from a given face database.

   Face Identification is a process involving with *One-to-Many* matching as shown in Figure 1.1. Biometric features are extracted from a probe face image of a subject. These features are then compared to all the templates stored in a database. The subject is identified as the identity associated with the template which is the most similar to the probe face.

2. **Face Verification**: Given an image or a video sequence containing a face of a person, verify whether this person is the one he/she claims to be.

   In contrast with face identification, face verification is a *One-to-One* matching process (see Figure 1.2). The subject presents a face image and a credential of his/her identity. The system then performs a comparison between the feature extracted from the face image and the stored template features of the claimed identity. If the similarity is above a threshold then the claim is accepted, otherwise
rejected.

Current Capability of Face Recognition

The theories and methodologies of image processing, computer vision and pattern recognition provide the theoretical background for face recognition system, and in return, the research in face recognition system leads to the development of theories and techniques in image processing, computer vision and pattern recognition in return. Significant attention has been paid on the research related to face recognition in the past decade.

The effort of researchers in the last decade has led to significant advances in face recognition. Consequently the performance of face recognition systems in controlled environments has reached a satisfactory level which offers many commercial opportunities for face recognition; however, there are still many challenges posed by uncontrolled environments for practical applications. Those challenges are the problems caused by variations in the illumination condition, face pose, expression, and etc.

1.1.3 The Illumination Problem

The effect of variation in the illumination conditions, which causes dramatic changes in the face appearance, is one of those challenging problems [133] a practical face
1.1. Motivation

Figure 1.3: Examples of face images of the same subject while under different illuminations in Yale-B database.

recognition system needs to face. To be more specific, the varying direction and energy distribution of the ambient illumination, together with the 3D structure of the human face, can lead to major differences in the shading and shadows on the face. Such variations in the face appearance can be much larger than the variation caused by personal identity [75]. From Figure 1.3 and Figure 1.4 it can be seen how different the faces of the same person can be when the illumination changes. Even a human being will be unable to successfully recognise faces that are poorly illuminated. The variations of both global face appearance and local facial features also cause problems for automatic face detection/localisation, which is the prerequisite for the subsequent face recognition stage. Therefore the situation is even worse for a fully automatic face recognition system. Moreover, in a practical application environment, the illumination variation is always coupled with other problems such as pose variation and expression variation, which increase the complexity of the automatic face recognition problem.

A number of illumination invariant face recognition approaches have been proposed in the past years. Despite the good performance claimed in the papers, the approaches in
each group have their own disadvantages. The approaches based on physical modelling of face image formation usually rely on the assumption of Lambertian reflectance model, which is unrealistic for the human face. Those based on statistical modelling require a large number of training samples under different illumination conditions to ensure good performance. Advanced photometric normalisation approaches are usually sensitive to the choice of parameters. The approaches based on 3D face information require 3D reconstruction, which involves high cost in both device and computational time. The approaches employing images in the invisible spectra offer visible light invariant face representations; unfortunately, the environmental illumination (especially in the outdoor environment) contains energy in these invisible spectra. Therefore the invisible spectrum face image is not invariant to changes in environmental illumination.

The objective of the work in this thesis is to develop an automatic illumination invariant face recognition system which can overcome the limitations of conventional approaches.
1.2 Contributions

The contributions of this thesis can be summarised as follows:

The thesis presents an extensive literature review of the existing techniques for illumination invariant face recognition. This review includes not only the traditional passive approaches which have received most attention, but also the active approaches which involve active sensing device to capture illumination invariant modalities of the human face.

The thesis presents an investigation into the active Near-Infrared differential imaging technique for illumination invariant face recognition. Assuming a still scene, a linear response of the camera to the scene radiation and no saturation, the imaging modelling based on a radiometry analysis and a sensor modelling proves theoretically that this technique yields face images which are independent of the variation in the environmental illumination. The advantages of this technique over several representative illumination invariant approaches are examined and confirmed by extensive face recognition experiments across different time, across different illumination conditions and across both time and illumination.

The thesis presents a multistage approach to automatic face localisation in Near-Infrared image of the human subjects. The proposed multistage approach is the combination of a bright pupil detector and a FloatBoost face detector. Exploiting the circular shape of the bright pupil boundary, an novel technique based on edge following and chain code representation is proposed to detect the pupil boundary. The candidates of eye centers are then validated using Support Vector Machines classifiers based on the local eye appearance and global face appearance. In the situation when this pupil detector fails, a FloatBoost face detector is applied as an alternative solution. The proposed multistage approach achieves accurate face localisation, which leads to the excellent performance for a fully automatic face recognition system in a practical environment with varying illumination.

The thesis presents a motion compensation approach to address the motion problem encountered in active differential imaging. The imaging system performs a continuous
capture of one frame with active illuminant on (combined frame) followed by one frame with the illuminant off (ambient frame). If the subject is moving, then its location on different frames is different. The artifact is therefore introduced to the difference image of the combined and ambient frame. In the proposed approach the motion between the nearest two combined frames is estimated. Interpolation can then be performed to obtain a virtual combined frame “captured” at the time instance when the ambient frame was captured. The motion artifact is therefore removed from the difference image of the virtual combined frame and the ambient frame. It is shown by experimental results that the proposed approach significantly reduces the drop in the face recognition performance due to the motion problem.

1.3 Overview of Thesis

The organisation of the thesis is as follows:

In Chapter 2 the structure of a general face recognition system is first described. The basic methodology, including the subspace approaches, classifiers, and performance measures used throughout this thesis for face identification and verification recognition are then presented.

In Chapter 3 an extensive survey on approaches to illumination invariant face recognition is provided. The conventional approaches are classified into passive approaches, in which the images that have already altered by illumination variation are corrected, and active approaches, which involve the active sensing technique to obtain illumination invariant image modalities of the human face.

Chapter 4 presents the investigation of the Near-Infrared active differential imaging technique for illumination invariant face recognition. The image formation modelling for active differential imaging based on the radiometry analysis and sensor modelling is presented. The experimental imaging system and a face database captured by this system under varying ambient illumination conditions, at two different times, are then described. The Near-Infrared active differential imaging technique is compared to several advanced illumination normalisation approaches, including a photometric normal-
isation approach based on anisotropic smoothing [41], and two approaches based on illumination insensitive features: phase[92][44] and Fisherphase. The face identification and verification experiments are carried out under three test protocols, namely the Cross Session test, the Cross illumination test, and the Combined test. A superior performance is achieved by the Near-Infrared active differential imaging technique for all the tests.

In Chapter 5 the approaches for automatic face localisation are briefly reviewed. Most approaches can be classified into passive approaches, which include top-down and bottom-up techniques, and active approaches, which involve active illumination to facilitate face localisation.

Chapter 6 presents a multistage approach to face localisation for Near-Infrared face images. A novel approach for bright pupil detection based on edge following and chain-code representation is first proposed. A FloatBoost approach to face localisation is then investigated. The proposed multistage method combines the above two approaches and yields the best localisation performance on the captured face database. The face recognition experiments using the faces localised automatically by the proposed multistage approach are conducted. The excellent performance of the face recognition system in the automatic scenario is facilitated by the accurate face localisation provided by the proposed multistage approach.

In Chapter 7 the motion problem in face recognition based on the active differential imaging is addressed. The motion of the subject causes artifacts in the captured difference face image. A case study shows that the face similarity to the template decreased dramatically due to the motion. An approach based on motion compensation is then proposed to deal with the motion problem. It is shown that motion artifacts are removed using the proposed approach. Face recognition experiments on a moving face database show that the proposed approach yields significant improvements in the performance of the face recognition system when motion is present.

The thesis is concluded in Chapter 8 where the directions of future work are also suggested.
Chapter 1. Introduction
Chapter 2

Face Recognition

This chapter briefly presents the stages of a face recognition system, the algorithm for face recognition, and the evaluation of face recognition systems.

2.1 Face Recognition System

In general, a face recognition system consists of the five stages shown in Figure 2.1.

- Sensing

  One sensor or multiple sensors are used in this stage to capture face information for the claimant. Various sensors can be applied to obtain different modalities of faces, such as visible spectral images, infrared images, 3D representations etc. This thesis focuses on 2D face images.

- Detection/Localisation

  In a practical situation, the captured image contains one or more faces in an unpredictable background. The faces need to be precisely located in the image for recognition.

- Preprocessing

  The preprocessing stage usually involves photometric normalisation and geometrical normalisation.
Chapter 2. Face Recognition

Figure 2.1: Diagram of face recognition system.

- *Photometric normalisation* aims at removing the brightness difference caused by uneven illumination, or different responses from different sensors. A literature review on various photometric normalisation techniques can be found in Chapter 3. Histogram equalisation is one of the commonly adopted techniques to obtain face region with even distribution in greyscale histogram within a predefined dynamic greyscale range.

- *Geometrical normalisation* means the process of registering faces to a predefined position in the resulting image. Faces can be of different size, position, and pose in the image, and therefore a geometric normalisation is necessary to remove this variation. Usually this registration is based on the eye coordinates. A geometric transform, including scale, translation and rotation, based on eye coordinates in the original image can be applied to bring the query and the target image into registration. The registered image will then be cropped to a rectangle region with size $W \times H$ and the line linking the eye positions in this cropped image at distance $h$ from the top of the box as shown in Figure 2.2. The inter-ocular distance is set to equal to $D$. Instead
of a rectangular region, a polygonal or elliptical region is sometimes used in the literature as the bounding face region. Please note that the transformation we discuss here is 2D transformation. Face is inherently a 3D object and therefore the correction introduced by geometric normalisation based on 2D transform is inadequate for faces with large 3D pose variation. Nevertheless, it is enough for faces with near-frontal pose.

- Feature extraction

The features extracted for face recognition include local features, holistic features, or their combinations. The early attempts of face recognition were based on geometrical features such as physical facial feature positions, shapes or distances between those facial features. In [18] Brunelli and Poggio showed the advantage of the statistical holistic feature over the geometric features, such as nose width and length, mouth position and chin shape. Holistic feature based approaches make use of the whole face or global face representations in feature spaces such as Gabor wavelet, Local Binary Pattern(LBP), Fourier Transform, etc. Those representations are often projected into subspaces to achieve dimension reduction or better discriminative power for classification.
• Classification

The Classification based on the extracted features provides the decision for the face recognition system. The classification can be performed using a single classifier, such as Nearest Neighbor classifier, Bayesian classifier or Support Vector Machine(SVM), or a cascaded classifier, such as AdaBoost and FloatBoost, or by a consensus decision of multiple classifiers.

2.2 Subspace Approaches

2.2.1 Principle Component Analysis

Principle Component Analysis (PCA) is a dimension reduction technique also known as Karhunen-Loeve Transform. It involves an orthogonal transform of the image signal and has been introduced to face recognition by Sirovich and Kirby [103], and Kirby and Sirovich [57]. PCA gained its popularity after the work of Turk and Pentland in [113], in which the term “eigenface” was introduced to represent the PCA bases of face images.

Given a set of $n$ vectors $X = \{x_i\}_{i=1}^n$, each of which has $m$ dimensions, the PCA basis is formed by the eigenvectors of the $m \times m$ covariance matrix $\Sigma$ as defined by

$$
\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T
$$

(2.1)

where $\bar{x}$ is the mean of the observations in $X$

$$
\bar{x} = \sum_{i=1}^{n} x_i
$$

(2.2)

By solving the eigenvalue problem

$$
\Sigma \Phi - \Phi \Lambda = 0
$$

(2.3)

we obtain $\Phi = [\phi_1, ..., \phi_m]^T$, which is a matrix composed of orthogonal eigenvectors $\{\phi_i\}_{i=1}^m$, and a diagonal matrix $\Lambda$ with eigenvalues $\{\lambda_i\}_{i=1}^m$ on the main diagonal and arranged as $\lambda_1 > ... > \lambda_m$. 
2.2. Subspace Approaches

Each eigenvalue $\lambda_i$ reflects the variance of $x$ in the direction represented by corresponding eigenvector $\phi_i$. Without any information loss, a vector $x$ can be decomposed into linear combinations of the eigenvectors $\{\phi_i\}_{i=1}^{m}$ as described by Equation 2.4.

$$x = \bar{x} + \sum_{i=1}^{M} a_i \phi_i$$ (2.4)

where $a_i$ is the projection coefficient of $x - \bar{x}$ in the direction $\phi_i$

$$a_i = \phi_i^T (x - \bar{x})$$ (2.5)

However, high dimensional data usually contains redundant information. This is especially true for images of faces, which contain smooth texture, and share similarity in appearance. The redundancy is reflected by the sharp drop in the magnitude of the ordered eigenvalues and the fact that most eigenvalues are close to zero. Therefore little loss will be incurred if we represent $x - \bar{x}$ using just projection to the top $k$ eigenvectors $\{\phi_i\}_{i=1}^{k}$. The $k$ dimensional PCA projection coefficient vector $a$ can then be obtained by

$$a = \Phi_k^T (x - \bar{x})$$ (2.6)

where $\Phi_k = [\phi_1, ..., \phi_k]$. There are several different ways to choose $k$. A common approach is based on the ratio of the retained to total energy as defined by:

$$e = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$ (2.7)

Since $k << m$, the dimensionality is greatly reduced so that both the computational complexity and the number of training samples needed to provide a good estimate for the face model is reduced. For example, the original dimension for a 55*50 face region is 2750. For face images in the XM2VTS database, if we choose $e = 0.95$, then $k = 294$.

2.2.2 Linear Discriminant Analysis

PCA concentrates on a subspace with much lower dimension while retaining most of the variance of the original data. However this is not necessarily helpful to distinguishing different classes. Linear Discriminant Analysis (LDA) aims at finding a linear subspace which maximizes the discriminability of classes after subspace projection.
For a set of data vectors with $C$ classes $\{X_c\}_{c=1}^C$, the within-class scatter matrix $S_W$ is defined as

$$S_W = \sum_{c=1}^C \sum_{x \in X_c} (x - \bar{x}_c)(x - \bar{x}_c)^T$$

(2.8)

and between-class scatter matrix $S_B$ as

$$S_B = \sum_{c=1}^C N_c(x_c - \bar{x})(x_c - \bar{x})^T$$

(2.9)

where $\bar{x}_c$ is the mean of class $c$, $N_c$ is the number of samples in this class, and $\bar{x}$ is the global mean of the whole data set. One of the criteria for choosing the optimal projection matrix $W_{lda}$ is the Fisher Discriminant Criterion introduced to face recognition by Belhumeur et al. \[8\]:

$$W_{lda} = \arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|}$$

(2.10)

Since in high dimensional spaces $S_w$ is almost always singular, the denominator of formula 2.10 is zero. Therefore a dimension reduction must be applied beforehand to project the original data set into a subspace in which the $S_w$ is non-singular.

It can be shown that $W_{lda}$ is the solution of the eigenvalue problem

$$S_B W_{lda} - S_W W_{lda} \Lambda = 0$$

(2.11)

It can also be shown that the dimensionality after projection with $W_{lda}$ is at most $C - 1$. Therefore the dimension reduction is achieved together with a better separability of the classes.

With the Fisher Discriminant Criterion, the LDA bases are referred to as “Fisherfaces”. Bellumeur et al. \[8\] showed the “Fisherface” approach significantly outperformed the “Eigenface” approach in a face recognition experiment with large variations in illumination condition.

An interesting case of LDA is so-called Client Specific LDA (CSLDA)\[68\], in which one specific LDA is built for each client. In this case the problem is a two-class problem with one client class and one non-client class. Therefore after CSLDA projection the coefficient vector contains only one scalar.
2.3. Classifiers

2.2.3 Others

There exist other subspaces methods, such as Independent Component Analysis (ICA) [3], Kernel PCA (KPCA) [124], Kernel LDA (KLDA) [125], etc. In ICA a non-orthogonal transformation is selected such that the projected data is statistically independent. KPCA and KLDA are kernel versions of PCA and LDA where a nonlinear mapping is applied to the original data before the PCA or LDA projection.

2.3 Classifiers

2.3.1 Nearest Neighbor

When a class has too few samples to obtain a good estimate of the class distribution, a non-parametric classification method such as Nearest Neighbor (NN) method is preferable.

According to the NN rule, given a training set with samples from \( C \) classes and an input sample \( x \), if \( x_c \) is the nearest sample to \( x \) among the whole training set, then \( x \) belongs to the same class as \( x_c \). It can be shown that the error rate for a classifier based on the NN rule is less than twice the Bayes error rate [24].

An important issue for the NN classifier is the distance measure. A variety of distance measures have been adopted in the literature, including Euclidean Distance, Mahalanobis Distance, Normalised Correlation (NC), etc. The normalised correlation has been shown to be particularly effective in face recognition. Given two vectors \( x_1 \) and \( x_2 \), it is defined as

\[
NC(x_1, x_2) = \frac{x_1 \cdot x_2}{\|x_1\| \cdot \|x_2\|}
\]

(2.12)

where \( \| . \| \) is the vector magnitude \( L^2 \)-norm. The range of NC distance is \([-1,1]\).

2.3.2 Support Vector Machine

Support Vector Machine (SVM) is a popular classification method for two-class problems. SVM finds the optimal linear discriminant plane for a linearly-separable classes,
Chapter 2. Face Recognition

but relies on embedding data (by nonlinear mapping) into a high dimensional space, in which the two classes can be separated by a hyperplane [27]. The aim of SVM training is to find the separating hyperplane exhibiting the largest margin for both classes. In this way good generalization can be achieved. The support vectors are those training samples that define the optimal separating hyperplane.

A multi-class problem can be solved by the combination of multiple SVMs, each designed for a two-class problem [11]. For instance, the \( k \)th SVM is constructed to distinguish samples from class \( k \) and those from the other classes. This is called one-versus-the-rest approach. An alternate approach is called one-versus-one approach, in which a SVM is trained for every possible pair of classes, and the test sample is classified into the class with the highest number of "votes".

2.4 Performance Evaluation of Face Recognition

There are three methodologies for the performance evaluation of face recognition system: technology evaluation, scenario evaluation and operational evaluation [85]. In technology evaluation, a face recognition system is trained and tested on a pre-captured face database. Scenario evaluation measures the overall system performance for a prototype scenario that models the application domain. An operational evaluation is similar to the scenario evaluation but more specific about the system to test and the application environment. This thesis focuses on the technology evaluation.

The face recognition system is trained and tested on different sections of the data set. For the face identification experiment, the data set can be divided into a training set and a test set. In addition to the training set and test set, face verification experiments sometimes need an evaluation set (also called validation set). An example is the experimental protocol of the XM2VTS database which involves the evaluation set for setting the decision thresholds and designing potential fusion stage.
2.4.1 Verification Test and Performance Evaluation

In verification the task is to confirm whether the claimant is the client he/she claims to be. The decision on whether or not accept a claim is made based on whether the similarity score $S$ is above a threshold $T$. There can be two different types of errors:

- **False Acceptance**

  This occurs when the claimant makes false claim (claims to be someone else) but the system accepts this claim. The frequency at which this error takes place is called False Acceptance Rate (FAR), or False Match Rate (FMR).

  $$FAR = \frac{\text{Number of Accepted False Claims}}{\text{Number of Total False Claims}}$$  \hfill (2.13)

- **False Rejection**

  This occurs when the claimant makes true claim (claims to be himself/herself) but the system rejects this claim. The frequency at which this error takes place is called False Rejection Rate (FRR), or False Non-Match Rate (FNMR).

  $$FRR = \frac{\text{Number of Rejected True Claims}}{\text{Number of Total True Claims}}$$  \hfill (2.14)

Half Total Error Rate (HTER) is defined as the average of FRR and FAR. Assuming that true claim and false claim are equally likely, the HTER reflects the system performance. Both FAR and FRR are functions of the threshold $T$. For a given value of $T$, there is a pair of $FAR(T)$ and $FRR(T)$. They can be plotted against each other as a curve in 2D (see Figure 2.3)

$$ROC(T) = (FRR(T), FAR(T))$$  \hfill (2.15)

and this curve, known as Receiver Operating Characteristic (ROC), expresses the behavior of FAR and FRR. Equal Error Rate (ERR) is defined as the error rate when FRR is equal to FAR. The threshold is chosen as the ERR point on the ROC for the evaluation set. It is then applied to the test set to obtain FAR and FRR and consequently HTER. Different error causes different level of risk for a specific application. For instance, forensic applications require a low FRR, high security applications need a low FAR. In civilian applications a system with low EER is preferable.
2.4.2 Identification Test and Performance Evaluation

In a face identification problem, the whole database is searched and the identification engine returns the possible matches for the input query face image. There can be three modes of operation[14]: Threshold-based mode, in which 1:1 verification is repeated against all the identities in the database and those matches with scores exceeding the threshold are returned. Rank-based mode, in which all the similarities between the input and the identities in the database are sorted and the identities with top-K similarities are returned. Hybrid mode, in which both thresholding and ranking are considered.

Different measures have been proposed for identification systems with different modes. For the threshold-based identification systems since it is close to face verification, the performance evaluation tools, such as ROC, FAR and FRR, can be applied. For the rank-based mode, a Cumulative Match Curve(CMC) is usually adopted. For a given rank $k$, $CMC(k)$ is defined as the frequency with which the true identity occurs among the top-$k$ matches.

This thesis concentrates on the rank-based mode. In our experiments the database contains all the identities for the test sample. Since the databases captured for our experiment are not of large scale, the rank-1 error rate is considered in this thesis. The rank-1 error rate is the probability that the input face is the most similar to a gallery
face of another identity.

2.5 Conclusion

In this chapter the stages of a general face recognition system, have been presented. Aiming at an illumination invariant face recognition system, work has been carried out on some of those stages, and will be presented in the following chapters of this thesis. An active Near-IR differential imaging technique is investigated for ambient illumination removal at the sensing stage. An automatic localisation technique has been proposed to address the face localisation problem for Near-IR face images. For the scenarios when people are moving their faces during the capture, a preprocessing approach based on motion compensation has been developed to remove the motion effects.

The face recognition algorithms which has been described in this chapter will be applied in this thesis. Principle component analysis is used to reduce the high dimensionality of face representation. Linear discriminant analysis is used to provide discriminative subspace for face recognition. Support Vector Machines and Nearest Neighbor classifiers with normalisation correlation are applied to perform classification for probe face images.

The performance of face verification is measured by Equal Error Rate and Receiver Operation Curve. Rank-1 identification error rate is used to measure the performance of face identification.
Chapter 3

Literature Review on Illumination-Invariant Face Recognition

3.1 Introduction

In this chapter a literature review on illumination-invariant face recognition is presented. Existing approaches addressing the illumination variation problem fall into two main categories. We call the approaches in the first category “passive” approaches, since they attempt to overcome this problem by studying the visible spectrum images which have been altered by illumination variations. The other category contains “active” approaches, in which the illumination variation problem is overcome by employing active imaging techniques to obtain face images captured under consistent illumination condition or images of illumination invariant modalities.

3.2 Passive Approaches

Passive approaches can be further divided into two groups. Approaches in the first group attempt to model the behavior of the face appearance as a function of illumina-
In the other group, the goal is to remove the influence of illumination changes from face images or extract face features that are invariant to illumination.

3.2.1 Illumination Variation Modelling

The modelling of face images under varying illumination can be based on a statistical model or physical model. For statistical modelling, no assumption concerning the surface property is needed. Statistical analysis techniques, such as PCA (Eigenface) and LDA (Fisherface), are applied to the training set which contains faces under different illuminations to achieve a subspace which covers the variation of possible illumination.

In physical modelling, the model of the process of image formation is based on the assumption of object surface reflectance property. The recognition based on all the methods of physical illumination modelling involves the same procedure: For a query image, find the most similar image among all the images described by the model for each candidate identity, then use the distance between the query image and this image as the similarity between the query image and this identity. The query image will be classified to the identity with the highest similarity. When a linear subspace is used as the model, the most similar image is simply the one giving the maximum projection of the query image into the subspace.

Linear Subspaces

Hallinan [42] models the whole set of a face under arbitrary illumination as a linear subspace composed of the faces under different point sources. The principle component analysis is applied to the faces under light directions which are evenly distributed samples of the hemisphere with a 15° step. The experiments showed that five eigenfaces were sufficient to represent the face images under a wide range of lighting condition. This PCA model does not involve any assumption of surface properties such as Lambertian reflectance.

Shashua proposed Photometric Alignment approach to find the algebraic connection between all images of an object taken under varying illumination conditions [95]. An order k Linear Reflectance Model for any surface point $p$ is defined as the scalar product
3.2. Passive Approaches

\[ x \cdot a, \text{ where } x \text{ is a vector in the } k\text{-dimensional Euclidean space of invariant surface properties (such as surface normal, surface albedo, and so forth), and } a \text{ is an arbitrary vector. The Lambertian model of reflection is an order 3 linear reflectance model, where the grey value } I(p) \text{ for point } p \text{ can be represented by } I(p) = n_p \cdot s \text{ with } n_p \text{ being the surface normal multiplied with albedo, and } s \text{ the lighting source vector. The image intensity } I(p) \text{ of an object with an order } k \text{ reflection model can be represented by a linear combination of a set of } k \text{ images of the object.}

\[ I(p) = \alpha_1 I_1(p) + \alpha_2 I_2(p) + \cdots + \alpha_k I_k(p) \quad (3.1) \]

For Lambertian surface under distant point sources and in the absence of shadows, all the images lie in a 3D linear subspace of the high dimensional image space, which means that they can be represented by a set of 3 images, each from a linearly independent source. Given three images of this surface under three known and linearly independent light sources, the surface normal and the albedo can be recovered. This process is known as Photometric Stereo. Shashua claimed the attached shadows, which is caused by points where the angle between surface normal and the direction of light source is obtuse \((n_p \cdot s < 0, \text{ therefore } I(p) = 0)\), do not have a significant adverse effect on the photometric alignment scheme. However, the cast shadows, which is caused by occlusion, cannot be modelled using the above framework.

Belhumeur et al. [8] presented the so-called 3D linear subspace method for illumination invariant face recognition, which is a variant of the photometric alignment method proposed by Shashua in [95]. In this linear subspace method, three or more images of the same face taken under different lighting are used to construct a 3D basis for the linear subspace. The recognition proceeds by comparing the distance between the test image and each linear subspace of the faces belonging to each identity. The Fisher Linear Discriminant (also called FisherFace) method is also proposed in [8] in order to maximise the ratio of the between-class scatter and the within-class scatter of the face image set to achieve better recognition performance. A comparison of several recognition methods, namely Correlation, Eigenfaces, 3D linear subspace and Fisherfaces, was carried out on Harvard face database for the face identification experiments. In the Harvard face database, all faces of the same person are in the same pose, but each face
is illuminated by a different lighting source. The space of light source directions which can be parameterized by spherical angles, was sampled in 15° increments. 330 images of 5 people (66 of each) were used for experiments. They were divided into 5 subsets, each subset containing images for which both the longitudinal and latitudinal angles of light source directions were within a certain degree of the camera axis. The lighting directions in subset 1 are close to frontal, then starting with subset 2 the directions increasingly deviate from the camera axis. The two experiments performed on this database were the extrapolation and interpolation of illumination. In the extrapolation experiment each method was trained on subset 1 and tested on subsets 1, 2, 3. In the interpolation experiment, each method was trained on subset 1 and 5, then tested on subset 2, 3 and 4. It was noticed that the first 3 PCA components are sensitive to illumination changes. The fisherface method outperformed the other methods for all the experiments.

The 3D linear subspace approach requires the assumption that there does not exist any shadow. In the presence of shadows the constructed subspace is erroneous. A segmented linear subspace model is proposed by Batur and Hayes in [6] to generalize the 3D linear subspace model so that it is robust to shadows. Each image in the training set is segmented into regions that have similar surface normals based on the $k$-Mean clustering, then for each region a linear subspace is estimated. Each estimation only relies on the specific region, therefore it is not influenced by the regions in shadow. The recognition experiment is performed on Yale B face database, and with less computing effort, the method leads to the same result as the illumination cone method discussed in the next section.

Illumination Cone

Belhumeur and Kriegman [9] proved that all images of a convex object with Lambertian surface from the same viewpoint but illuminated by an arbitrary number of distant point sources form a convex Illumination Cone, and the dimension of this illumination cone is the same as the number of distinct surface normals. This illumination cone can be constructed from as few as three images of the surface, each under illumination from an unknown point source. The illumination cone is a convex combinations of extreme
3.2. Passive Approaches

rays, which can be given by $x_{i,j} = \max(B_{s_{ij}}, 0)$, where $s_{i,j} = b_i \times b_j$, and $b_i, b_j$ are two different rows of a matrix $B$ where each row is the product of albedo with surface normal vector.

Kriegman and Belhumeur showed in [61] that for any finite set of point sources illuminating an object viewed under either orthographic or perspective projection, there is an equivalence class of object shapes having the same set of shadows. In the perspective projection case, two surfaces share the same shadow if they differ by a so-called *Generalized Perspective Bas-Relief (GPBR)* transformation, and in the orthographic case, this transformation is the *Generalized Bas-Relief (GBR)* transformation.

These observations are exploited by Georghiades et al. [37] for face recognition under variable lighting. In their experiment, 660 images from 10 people (66 of each) taken from the Harvard database are used. Recognition methods are trained on subset 1 and tested on subset 2-5. The performance with the illumination cone methods is superior to that achieved by Eigenfaces, Eigenfaces without the first three components, correlation method and 3D linear subspace method. The advantages are significant especially on subsets with extreme illumination. The authors further tested the algorithm on Yale B face database [38], which contains images of 10 individuals under 64 different lighting conditions and 9 different poses, and extended it to the pose-invariant face recognition. The extrapolation in the illumination experiment similar to [37] and [8] is carried out with all the methods considered in [37] and the superiority of their method over the other methods is confirmed once again.

**Spherical Harmonics**

The Spherical Harmonics method is proposed by Basri and Jacobs [4], and contemporarily by Ramamoorthi and Hanrahan [89]. Assuming arbitrary light sources (point sources or diffuse sources) distant from an object of Lambertian reflectance property, Basri and Jacobs [5] show that ignoring cast shadow the intensity of object surface can be approximated by a 9-dimensional linear subspace based on a *Spherical Harmonic* representation. The reflected radiance at location $p$ on a convex surface is given by:

$$ r(p) = \rho \int_0^{2\pi} \int_0^\pi k(\theta) L(\theta, \phi) \sin \theta d\theta d\phi $$
where \( \rho \) is the albedo for point \( p \), and \((\theta, \phi)\) are the coordinates in the local unit spherical coordinates system. \( k(\theta) = \max(\cos(\theta), 0) \) is the Lambertian kernel, and \( L(\theta, \phi) \) is the radiance from the light sources. The spherical harmonics \( Y_{nm} \), with \( n = 0, 1, 2, \ldots \) and \(-n < m < n\), are a set of functions that form an orthonormal basis for any \( L^2 \)-integrable function defined on a sphere, i.e.

\[
Y_{nm} = \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_{nm}(\cos \theta) e^{im\phi} \tag{3.3}
\]

where \( P_{nl} \) are the associated Legendre functions. \( Y_{nm} \) is called the \( n \)'th order harmonic, and can be mapped to a function expressed in terms of the corresponding local Cartesian coordinates \( Y_{nm}(x, y, z) \). It can be shown that

\[
r(p) = \rho \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \left( \sqrt{\frac{4\pi}{2n+1}} k_{nm} l_{nm} \right) Y_{nm} \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} l_{nm} b_{nm}(p) \tag{3.4}
\]

where \( k_{n} \) and \( l_{nm} \) are the coefficients of the harmonic expansion of \( k(\theta) \) and \( L(\theta, \phi) \). \( b_{nm}(p) = \rho \sqrt{\frac{4\pi}{2n+1}} Y_{nm} \) are harmonic images. Eq.3.4 shows the image radiance in any point \( p \) can be approximated as a linear combination of harmonic images of up to order \( N \). It can be shown that with \( N = 3 \) (in total 9 harmonic images), the approximation retains more than 99% of the energy.

Basri and Jacobs carried out a face identification experiment on the NEC face database which contains models of 42 faces, each including the 3D shape of the face. The albedos were estimated and the harmonic model for each individual was built. The distance between the query image and the subspace spanned by the harmonic basis images for each individual is used as a similarity measure for recognition. According to the CMC curve reported in [5], the rank-1 error rate is about 14%.

Lee et al. [63] showed that there exists a configuration of nine point source directions such that a subspace resulting from nine images of each individual under these nine lighting sources is effective at recognition under a wide range of illumination conditions. They reported very good results on the Yale B face database. The advantage of this method is that there is no need to obtain a 3D model of surface as in the spherical harmonics approach [5], or to collect a large number of training images as in the statistical modelling approaches.
3.2. Passive Approaches

Zhang and Samaras [129] proposed two methods for face recognition under arbitrary unknown lighting by using the spherical harmonics representation, which requires only one training image per subject and no 3D shape information. In the first method [128] the statistical model of harmonic basis images are built based on a collection of 2D basis images. For a given training face image, the basis images for this face can be estimated based on Maximum A Posterior estimation. In the second method the 3D morphable model and the harmonic representation are combined to perform face recognition with both illumination and pose variation.

Linear Lambertian Objects

Recently, Zhou et al. [134] analyzed the images of the face class with both Lambertian reflectance model and linear subspace approach. The human face is claimed to be an example of a so-called Linear Lambertian Object, which is not only an object with Lambertian surface, but also a linear combination of basis objects with lambertian surfaces. The albedo and surface normal vectors of each basis object for the face class form a matrix called class-specific albedo/shape matrix $W$, which can be recovered through a Generalized Photometric Stereo process from the bootstrap set. Face recognition is performed on the coefficients of the linear combination. The identification experiments were conducted on frontal faces of 68 subjects in the CMU-PIE face database. The gallery set consists of faces under one of 12 illuminations, and the probe set under a different illumination. The model is trained using Vetter's 3D face database [13]. Excellent performance was reported. The work was further extended to the case with multiple light sources.

3.2.2 Illumination Invariant Features

Adini et al. [1] presented an empirical study that evaluates the sensitivity of several illumination insensitive image representations to changes in illumination. These representations include edge map, image intensity derivatives, and image convolved with a 2D Gabor-like filter. All of the above representations were also followed by a log function to generate additional representations. With varying parameters, the total number
of representations considered was 107. However, the recognition experiment on a face database with lighting variation indicated that none of these 107 representations is sufficient by itself to overcome the image variation due to the change of illumination direction. Even the best representation among them can only achieve a 20% error rate.

Features Derived from Image Derivatives

Line Edge Map [36] is proposed for face recognition by Gao and Leung. The edge pixel is grouped into line segments, and a revised Hausdorff Distance is designed to measure the similarity between two line segments.

Kryszczuk and Drygajlo [62] applied a segmentation method based on the local variance of image gradient. The regions with severe shadow or specular reflections are discarded when performing face recognition.

Chen et al. [21] showed that for any image, there are no discriminative functions that are invariant to illumination, even for objects with Lambertian surface. However, they showed that the probability distribution of the image gradient is a function of the surface geometry and reflectance, which are the intrinsic properties of the face. The direction of image gradient is revealed to be insensitive to illumination change. The recognition performance using gradient direction is close to the illumination cone approach.

Relative Image Gradient feature is applied by Wei and Lai [116] and Yang et al. [122] for robust face recognition under lighting variation. The relative image gradient $\tilde{G}(x, y)$ is defined as

$$\tilde{G}(x, y) = \frac{\nabla I(x, y)}{\max_{(u,v) \in W(x,y)} |\nabla I(u, v)| + c}$$

(3.5)

where $I(x, y)$ is the image intensity, $\nabla$ is the gradient operator, $W(x, y)$ is a local window centered at $(x, y)$, and $c$ is a constant value to avoid dividing by zero. In their experiment on the frontal faces in the CMU-PIE database, three training images, each with left, right and ambient lighting respectively, for each subject are used. A 1.47% error rate is achieved when testing on the remaining frontal images with varying lighting.
3.2. Passive Approaches

Zhao and Chellappa [132] presented a method based on *Symmetric Shape from Shading* for illumination insensitive face recognition. Traditional shape from shading approach does not work for face due to the complex shape and varying albedo of the face surface. They utilize the symmetry of every face and the shape similarity among all faces in their approach. A prototype image with normalized illumination can be obtained from a single training image under unknown illumination. Their experiments showed that using the prototype image significantly improved the face recognition based on PCA and LDA.

Sim and Kanade [101] developed a statistical shape from shading model to recover face shape from a single image and to synthesize the same face under new illumination. The surface radiance $i(x)$ for location $x$ is modelled as

$$i(x) = n(x)^T \times s + e \quad (3.6)$$

where $n(x)$ is the surface normal with albedo, $s$ is the light source vector, $e$ is an error term which models shadows and specular reflections. A bootstrap set of faces with labelled varying illuminations is needed to train the statistical model for $n(x)$ and $e$. The illumination for an input image can be estimated using kernel regression based on the bootstrap set, then $n(x)$ can be obtained by Maximum A Posterior estimation and the input face under a new illumination can be synthesized.

**Quotient Image**

Shashua and Riklin-Raviv [96] treat face as an *Ideal Class of Object*, by which they mean the objects that have the same shape but differ in the surface albedo. Therefore, under lambertian assumption, the image class of faces is given by $\rho_i(p)n(p)^T s_j$, where $\rho_i(p)$ and $n(p)$ are the albedo and surface normal for the point $p$ of face $i$, and $s_j$ is an arbitrary illumination direction. The *Quotient Image* $Q_y$ of object $y$ against object $a$ is defined by

$$Q_y(u, v) = \frac{\rho_y(u, v)}{\rho_a(u, v)} \quad (3.7)$$

where $u$ and $v$ range over the image. The image $Q_y$ depends only on the relative surface texture information, and is independent of illumination. A bootstrap set containing $N$
faces under three unknown independent illumination directions is employed. A probe image $Y(u, v)$ under an unknown illumination is transformed into its quotient image against the mean of images in the bootstrap set:

$$Q_y(u, v) = \frac{\rho_y(u, v)}{\rho_A(u, v)}$$  \hspace{1cm} (3.8)

where $\rho_A(u, v)$ is the average of the albedos for all images in the bootstrap set. It can be shown that $Q_y$ can be calculated as

$$Q_y(u, v) = \frac{Y(u, v)}{\sum_j A_j(u, v)x_j}$$  \hspace{1cm} (3.9)

where $A_j(u, v)$ is the average of images under illumination $j$ in the bootstrap set, and $x_j$ can be determined by all the images in bootstrap set and $Y(u, v)$. Then the recognition is performed based on the quotient image. The results on Vetter's face database of 1800 face images demonstrate the significant advantage of the Quotient image method over the correlation method in the original face image space, and the performance is similar to eigenfaces method.

Based on the assumption that faces are an Ideal Class of objects as in the Quotient Image approach, Shan et al. [94] proposed Quotient Illumination Relighting. When the illumination in the probe image and the target illumination condition are both known and exist in the bootstrap set, the rendering can be performed by a transformation learnt from the bootstrap set. This approach had a better performance in experiments on Yale B and Harvard face databases than the histogram equalisation or gamma intensity correction (see Section 3.2.3) or their combinations.

Chen and Chen [20] proposed a Generic Intrinsic Illumination Subspace approach. Given the Ideal Class assumption, all objects of the same ideal class share the same generic intrinsic illumination subspace. Considering attached shadows, the appearance image of object $i$ in this class under a combination of $k$ illumination sources $\{l_i\}_{i=1}^k$ is represented by

$$I_i(x, y) = \rho_i(x, y) \sum_{j=1}^{k} \max(n(x, y)l_j, 0)$$  \hspace{1cm} (3.10)

where $\rho_i(x, y)$ is the albedo, and $n(x, y)$ is the surface normal vector of all objects in
3.2. Passive Approaches

The illumination image is defined as

\[ L(x, y) = \sum_{j=1}^{k} \max(n(x, y)l_j, 0) \]  

(3.11)

The illumination images of a specific Ideal Class form a subspace called Generic Intrinsic Illumination Subspace, which can be obtained from a bootstrap set. For a given image the illumination image can be estimated by

\[ L = Bl \]  

(3.12)

where \( l = \arg\min||Bl - L^*|| \). Here \( B \) is the basis matrix of the intrinsic illumination subspace, and \( L^* \) is an initial estimation of illumination image based on smoothed input image. Finally \( \rho(x, y) \) can be obtained by \( \rho(x, y) = I(x, y)/L(x, y) \). The method was evaluated in CMU-PIE and Yale B face database and showed significant better results than Quotient Image method. It is also shown that enforcing non-negative light constraint will further improve the results.

Transformation domain features

Recently methods based on the frequency domain representation have received attention. Savvides et al. [92] showed that there is correlation between the PCA in the image domain and the PCA in the frequency domain, therefore PCA in frequency domain does not bring any advantage over PCA in image domain. However, the phase spectrum of a face image carries illumination insensitive information. They performed PCA in the phase domain and achieved impressive results on the CMU-PIE database [100]. This so-called *Eigenphase* approach improved the performance dramatically compared to Eigenface, Fisherface and 3D linear subspace approach. Meanwhile, they further showed that even with partial face images the performance of the Eigenphase approach remains excellent and the advantages over other approaches are even more significant. Heo et al. [44] showed that applying Support Vector Machines directly on phase can lead to a even better performance than the Eigenphase approach mentioned above. Llano et al. [69] examined the sensitivity of several frequency domain representations of face image to illumination change. Those representations are the magnitude, phase,
real part and imaginary part of the Fourier spectrum of original face image, and those of gradient image. The gradient image is defined as an image where each pixel has a complex value with the horizontal gradient of the original image as the real part, and the vertical gradient as imaginary part. The sensitivity is measured by the overlap between the intra-person distribution and inter-person distribution of the correlation score based on each representation. The experiments are performed on the normal illumination set and the darkened set of the XM2VTS face database. The results show that the real part of the Fourier spectrum of the gradient image is less sensitive to illumination change than other representations.

In [120] a quaternion correlation method in a wavelet domain is proposed and good performance is achieved on the CMU-PIE database with only one training sample per subject. Each image is decomposed to LL, HL, LH and HH frequency band by applying the discrete wavelet decomposition. The subband images are then encoded into a 2-D quaternion array as

\[ f(x) = f_{LL}(x) + f_{LH}(x)i + f_{HL}(x)j + f_{HH}(x)k \]  

(3.13)

Quaternion Fourier Transform is performed to transfer the quaternion image to quaternion frequency domain, then quaternion correlation filter is applied.

Qing et al. [86] showed that the Gabor phase is tolerant to illumination change and has more discriminative information than phase in Fourier spectrum. The recognition is performed using the probabilistic model of intra- and inter- person distance based on the Gabor phases with different scales and orientations. Good results are reported for experiments on CMU-PIE and Yale B face databases.

Local Binary Pattern

Local Binary Pattern (LBP) is a local feature which characterizes the relationship between a pixel and its neighbors. For a given pixel position, LBP is defined as an ordered set of binary comparisons of pixel intensity between the center pixel and its surrounding pixels. The binary code obtained in each comparison is concatenated to form a binary string and can be converted to a decimal representation. LBP is
3.2. Passive Approaches

unaffected by any monotonic grayscale transformation in that the pixel intensity order is not changed after such a transformation. Furthermore, for a region with a number of pixels, a histogram for the LBP patterns associated with each pixels within this region tends to be a good feature for face recognition. LBP has been used in [65][66][45] as an illumination insensitive feature.

**Illumination Insensitive Eigenspaces**

Bischof et al. [10] proposed *Illumination Insensitive Eigenspaces* approach by incorporating a set of gradient based filters into the eigenspace recognition method. In the eigenspace approach, an image $I(p)$ can be represented by

$$I(p) = \sum_{j=1}^{m} a_j e_j(p)$$ (3.14)

where $a_j$ is the coefficient for the $j$th eigenimage $e_j(p)$. Due to the linearity of the equation above, the filtered version of image $I(p)$ can be presented by

$$f \ast I(p) = \sum_{j=1}^{m} a_j (f \ast e_j(p))$$ (3.15)

with $k$ filters and $n$ points, the coefficients $\{a_j\}_{j=1}^{m}$ can be obtained by solving the linear system of the above equations. A bank of gradient-based filters is chosen to obtain illumination insensitivity. The approach shows robustness to illumination change in experiments in general object recognition.

3.2.3 Photometric Normalization

Various photometric normalization techniques have been introduced to pre-process face images. *Histogram Equalisation* [40] is the most commonly used approach. By performing histogram equalisation, the histogram of the pixel intensities in the resulting image is flat. It is interesting that even for images with controlled illumination (such as face images in the XM2VTS database), applying histogram equalisation still offers performance gain in face recognition [97].
Shan et al. [94] proposed Gamma Intensity Correction for illumination normalisation. The corrected image \( G(x, y) \) can be obtained by performing an intensity mapping

\[
G(x, y) = cI(x, y)^{\frac{1}{\gamma}}
\]  

(3.16)

where \( c \) is a gray stretch parameter, and \( \gamma \) is the Gamma coefficient.

In Homomorphic filtering approach [40] the logarithm of the equation of the reflectance model is taken to separate the reflectance and luminance. The reflectance model often adopted is described by

\[
I(x, y) = R(x, y) * L(x, y)
\]  

(3.17)

where \( I(x, y) \) is the intensity of the image, \( R(x, y) \) is the reflectance function, which is the intrinsic property of the face, and \( L(x, y) \) is the luminance function. Based on the assumption that the illumination varies slowly across different locations of the image and the local reflectance changes quickly across different locations, a high-pass filtering can be performed on the logarithm of the image \( I(x, y) \) to reduce the luminance part, which is the low frequency component of the image, and amplify the reflectance part, which corresponds to the high frequency component.

Du and Ward [26] performed illumination normalization in the wavelet domain. Histogram equalisation is applied to low-low subband image of the wavelet decomposition, and simple amplification is performed for each element in the other 3 subband images to accentuate high frequency components. Inverse wavelet transform is then employed to reconstruct the image based on the modified 4 subband images. Face recognition on YaleB database based on the reconstructed image achieved better results than histogram equalised images.

In Retinex approaches the luminance is estimated by the smoothed image. The image can then be divided by the luminance to obtain the reflectance. A single Gaussian function is applied to smooth the image in the single scale retinex approach [55], and the sum of several Gaussian functions with different scales is applied in the multi-scale retinex approach [56]. Logarithm transform is employed to compress the dynamic range in [55] and [56].

Wang et al. [115] defined Self-Quotient Image, which is essentially a multi-scale retinex approach, however instead of using isotropic smoothing as in [56], anisotropic smoothing
functions with different scales are applied. Each anisotropic smoothing function is a Gaussian weighted by a thresholding function. The convolution window is divided into two subregions, one of them containing pixels with intensities above local average, and the other contains the rest of the pixels. For those pixels in the subregion with more pixels than the other, the thresholding function has a value one, and for pixels in the other subregion the thresholding function has a value zero. Finally a nonlinear transformation combining Arctangent function and Sigmoid function is adopted for a dynamic range compression.

Gross and Brajovic [41] proposed to solve luminance \( L \) by minimizing an anisotropic function:

\[
J(L) = \int_{\Omega} \rho(x, y)(L - I)^2 dx dy + \lambda \int_{\Omega} (L_x^2 + L_y^2) dx dy
\]

(3.18)

where \( \rho(x, y) \) is space varying permeability weight which controls the anisotropic nature of the smoothing. \( \Omega \) refers to the whole image region, \( L_x \) and \( L_y \) are the spacial derivatives of \( L \), and \( I \) is the intensity image.

The isotropic version of function \( J(L) \) can be obtained by discarding \( \rho(x, y) \).

\[
J(L) = \int_{\Omega} (L - I)^2 dx dy + \lambda \int_{\Omega} (L_x^2 + L_y^2) dx dy
\]

(3.19)

Multigrid method [17] can be used to solve \( L \) numerically.

Xie and Lam [121] proposed an illumination normalization method which is called *Local Normalization*. They separate the face region as a sequence of triangular facets, the area of which is small enough to be considered as planar patch. The main idea of Local Normalization approach is to normalize the intensity values within each facet to be with zero mean and unit variance.

Short et al. [98] compared five photometric normalization methods, namely illumination insensitive eigenspaces, multiscale Retinex method, homomorphic filtering, a method using isotropic smoothing to estimate luminance, and one using anisotropic smoothing [41]. Each method is tested with or without histogram equalisation performed in advance. Interestingly it was found that histogram equalisation helped in every case. Comparison of the above photometric normalisation algorithms are made based on their performances as a pre-processing technique for face verification on three
contrasting face databases: the Yale B database [38], the BANCA database [2] and the XM2VTS database [72]. Yale B contains a large number of different illuminations, XM2VTS is captured under constant illumination, and BANCA is captured in realistic environment. It is shown that the anisotropic smoothing method yields the most consistent performance across the three test databases.

### 3.2.4 3D Morphable Model

Blanz and Vetter [13] proposed face recognition based on fitting a 3D morphable model. A database containing 3D face scans of 100 males and 100 females is used to build the model. The 3D morphable model describes the shape and texture of face separately based on the PCA analysis of the shape and texture obtained from the database of 3D scans. To fit a face image under unknown pose and illumination to the model, an optimisation process is needed to optimize shape coefficients, texture coefficients along with 22 rendering parameters to minimise the difference of the input image and the rendered image based on those coefficients. The rendering parameters include pose angles, 3D translation, ambient light intensities, directed light intensities and angles, and other parameters about the camera and color channels. The illumination model of Phong is adopted in the rendering process to describe the diffuse and specular reflection of the surface. After fitting both the gallery images and the probe images to the model, the recognition can be performed based on the model coefficients for shape and texture. Good recognition performance across pose and illumination is achieved in experiments on CMU-PIE and FERET face database.

### 3.3 Active Approaches

In active approaches additional devices (optical filters, active illumination sources or specific sensors) usually need to be involved to actively obtain different modalities of face images that are insensitive to or independent of illumination change. Those modalities include 3D face information [15] and face images in those spectra other than visible spectra, such as thermal infrared image [60] and near-infrared hyper-spectral
image [81]. Nevertheless, 3D imaging devices are relatively expensive and 3D reconstruction is computationally expensive. The problem with using multi-spectral images is that although invisible spectra images can be invariant to changes to visible illumination, there can be variation in those invisible spectra themselves in real application environments. For example, the infrared component varies greatly between indoor and outdoor environments.

3.3.1 3D information

3D information is one of the intrinsic properties of a face, which is invariant to illumination change. The surface normal information is also used in some passive approaches described in the previous section, however, they are recovered from the intensity images captured by the camera. In this section we focus on the 3D information acquired by active sensing devices.

3D information can be represented in different ways. The most commonly used representations are range image, profile, surface curvature, Extended Gaussian Image(EGI), Point Signature, and etc. Surveys on 3D face recognition approaches can be found in [15][16] and [93]. The 3D modality can be fused with 2D modality, i.e. texture, to achieve better performance [16][19]. Nevertheless, it should be noticed that the 2D face images which are combined with 3D face info as reported in [16][19] are captured in controlled environment. It is still not clear how much the fusion will help in the case of uncontrolled environment due to the impact of uncontrolled illumination on the 2D face intensity images.

Kittler et al. [58] reviewed the full spectrum of 3D face processing, from sensing to recognition. The review covers the currently available 3D face sensing technologies, various 3D face representation models and the different ways to use 3D model for face recognition. In addition to the discussion on separate 2D and 3D based recognition and the fusion of different modalities, the approach involving 3D assisted 2D recognition is also addressed.
3.3.2 Infrared

Electromagnetic Spectrum

The electromagnetic spectrum is shown in Figure 3.1. Visible light spectrum ranges from 0.4μm-0.7μm. The spectral bands below the visible spectrum such as X-rays and ultraviolet radiation are harmful to human body and thus not useful for face recognition. The infrared spectrum ranges from 0.7μm -1mm. It can be divided into 5 bands, namely: Near-Infrared(Near-IR) (0.7-0.9μm), the Short-Wave Infrared (SWIR) (0.9-2.4μm), the Mid-Wave Infrared(MWIR) (3.0-5.0μm), the Long-Wave Infrared(LWIR) (8.0-14.0μm), and Far-Infrared(FIR) (14.0μm-1mm). Near-IR and SWIR belong to reflected infrared (0.7-2.4μm), while MWIR and LWIR belong to thermal infrared (2.4μm-14μm). Similar to the visible spectrum, the reflected infrared contains the information about the reflected energy from the object surface, which is related to the illumination power and the surface reflectance property. Thermal Infrared directly relates to the thermal radiation from object, which depends on the temperature of the object and emissivity of the material [60]. Optical filters which block the energy from a specific range of spectrum and infrared sensors with response to a certain range of spectrum are available now in the market to obtain the image with/without any of the
3.3. Active Approaches

subbands. Therefore the modality independent of visible lighting can be obtained.

**Thermal Infrared for Face Recognition**

The human face and body emit thermal radiation in both MWIR and LWIR bands, however with much higher emission in the former band than in the latter. The thermal patterns of faces are derived from the patterns of superficial blood vessels under the skin, therefore, the uniqueness of the vessel pattern leads to the uniqueness of the thermal face image. It is interesting to note that even identical twins who look similar to each other will have different thermal patterns. As it is independent of visible lighting, thermal imaging can be applied to face recognition under poor illumination conditions. A survey on visual and infrared face recognition is presented in [60].

Wilder at al. [117] showed that with minor illumination changes and for subjects without eyeglasses, applying thermal image for face recognition does not lead to significant difference compared to visible images. However, for scenarios with huge illumination changes and facial expressions, superior performance was achieved based on radiometrically calibrated thermal face images than that based on visible image[107][104]. The experiments in [22] shows the face recognition based on thermal images degrades more significantly than visible images when there is a substantial passage of time between the gallery images and probe images. This result was proved to be reproducible by [106], however it is shown that with a more sophisticated recognition algorithm the difference of recognition performance across time based on thermal face and visible face can be small.

Despite the independence from visible light, the thermal imagery has its own disadvantages. The temperature of the environment, physical conditions and psychological conditions will affect the heat pattern of the face[7]. Meanwhile, the thermal infrared is opaque to eyeglasses. All the above motivate the fusion of thermal infrared image with visible images for face recognition. Various fusion schemes have been proposed [60][22][7][105] and lead to better performance than recognition based on either modality alone. The thermal face recognition experiments are usually conducted on a face database from the University of Notre Dame [22] or the Equinox face database [31].
The former contains the visible spectrum images and LWIR images of 240 subjects without glasses, but with different lighting and facial expressions. The latter was collected by Equinox Corporation and the visible images and LWIR images of a total of 115 subjects are available.

While most of the experiments are carried out on indoor face data, Socolinsky and Selinger [105] performed thermal face recognition in an operational scenario, where both indoor and outdoor face data of 385 subjects are captured. When the system is trained on indoor sessions and tested on outdoor sessions, the performance degrades no matter using thermal imagery or visible imagery, but the thermal imagery substantially outperformed visible imagery. With the fusion of both modalities, the outdoor performance can be close to indoor face recognition.

Active Near-IR Illumination

The Near-IR band falls into the reflective portion of the infrared spectrum, between the visible light band and the thermal infrared band. Thus it has advantages over both visible light and thermal infrared. Firstly, since it can be reflected by objects, it can serve as an active illumination source, in contrast to thermal infrared. Secondly, it is invisible, making active Near-IR illumination unobtrusive. Thirdly, unlike thermal infrared, Near-IR can easily penetrate glasses.

Pan et al. [81] performed face recognition in hyperspectral images. A CCD camera with a liquid crystal tunable filter was used to collect images with 31 bands over near-infrared range. It was shown the hyperspectral signatures of the skin from different persons are significantly different, while those belonging to the same person are stable. During the camera calibration the reflectance of the skin can be decided based on the images of a white spectralon and a black spectralon. The reflectances at five facial regions, namely forehead, left cheek, right cheek, hair and lips are used for recognition. For each region the reflectance values in 31 bands are normalised and combined into a spectral reflectance vector. The Mahalanobis distance between two spectral reflectance vectors from two different person's faces is computed, and the weighted sum of the five Mahalanobis distances of the five regions, represents the distance between these two
3.3. Active Approaches

persons. The reflectance of skin proved to be more important than the reflectance of hair and lips. The hyperspectral face image databases contains multiple sessions of images from 200 subjects, with variations in pose, expressions and time. Above 91% rank-one correct identification rate is obtained in the recognition experiments on frontal images.

Most recently, Li et al. [65][66] proposed a face recognition system based on active Near-IR lighting. Near-IR Lighting-Emitting Diodes(LEDs) with wavelength 850nm are mounted on the camera to provide the Near-IR lighting. An optical filter is applied to cut off the visible spectrum to obtain face images in the Near-IR spectrum. The Near-IR face image captured by this device is subject to a monotonic transform in the gray tone, then LBP feature is extracted to compensate for this monotonic transform to obtain an illumination invariant face representation. Zhao and Grigat [130] performed face recognition in Near-IR images based on Discrete Cosine Transform(DCT) feature and SVM classifier.

Hizem et al. [47] described two different types of specific sensors for face recognition. The first one is a differential CMOS imaging system. In their system the image subtraction is performed by an on-chip computation circuit in the CMOS sensor.

Active Near-IR illumination has also been widely used in various aspects of face processing other than face recognition. In [126] infrared structured patterns are projected onto the human face to reconstruct the 3D face shape. The pattern with dense texture improved the computational efficiency and accuracy in correspondence searching for the stereo reconstruction. The face texture is captured by several colour cameras with an infrared blocking optical filter. Therefore the face shape and texture can be captured simultaneously.

Dowdall et. al performed face detection on Near-IR face images [25]. Most of the efforts in face detection have been made for visible images which are susceptible to lighting changes and variability of human facial appearance. Human skin has a high reflectance to the spectrum below 1.4\(\mu\)m and low reflectance above 1.4\(\mu\)m. Face region can then be detected based on the difference of skin responses to the upper band(0.7-1.4\(\mu\)m) and the lower band(1.4-2.4\(\mu\)m) of reflective infrared illumination. This unique property of
skin reflectance has been used for automatic detection and counting of vehicle passenger [83], and it is very useful for disguise detection [82] and liveness test.

Morimoto and Flickner [74] apply two sets of Near-IR light to obtain bright pupils. Similar eye detectors using active illumination are used in [53] for 3D face pose estimation and tracking. In [64][131][138] facial feature detection in Near-IR images is studied.

Although infrared image is invariant to visible illumination, it is not independent of the environmental illumination. This is because environmental illumination contains energy in a wide range of spectrum, including infrared. The variation of the infrared component in the environmental illumination will impose variation on the captured image.

One solution to maximize the ratio between the active source and the environmental source is to apply synchronized flashing imaging by Hizem et al. [47]. In the first place, a powerful active illumination source is desirable. Illuminants such as LEDs can provide very powerful flash, but this flash must only last for very short time to avoid the internal thermal effects which might destroy the LEDs. The idea in Hizem et al. [47] is to synchronise the sensor exposure time with the powerful flash. The sensor is exposed to the environmental illumination only for the same short exposure time as the flash. Since the power of the flash is usually much stronger than the environmental illumination, the contribution of the environmental illumination to the capture image will be minimised.

Nevertheless, the illumination variation problem can only be alleviated but not completely solved by the above mentioned approach. For indoor environment, the infrared energy in environmental illumination is low and will not cause much problem, while in outdoor environment, the infrared energy can be very strong.

Safety Issues

Despite all the advantages of active Near-IR illumination, the safety issue should be considered. A long time exposure under strong Near-IR illumination is hazardous to
3.4 Conclusion

A review on the existing illumination invariant techniques in literature has been presented in this chapter. Although good performance is usually reported for most techniques, each technique has its own drawbacks.

The illumination modelling methods require training samples from controlled illumination. The physical modelling of the image formation generally requires the assumption that the surface of the object is Lambertian, which is violated for real human faces.

The statistic modelling methods require training samples from as many different illumination conditions as possible to ensure a better performance.

The performance of many photometric normalisation methods severely depends on the choice of parameters. For example, the anisotropic smoothing method by Gross and Brajovic [41] yields the best results for face verification experiment across Yale B, XM2VTS and BANC A face database, however, it is shown by Short et al. [99] that the face verification results are sensitive to the choice of smoothing parameter of this method.

3D acquisition is computationally expensive and the cost of the 3D devices is usually high. The cooperation from the client is required. Meanwhile, it is worthwhile investigating the performance of 3D acquisition in an uncontrolled environment. A number of 3D acquisition techniques make use of intensity/infrared image pairs for 3D reconstruction. The variation in the environmental illumination might cause problem for the accuracy of 3D reconstruction in those systems.

The outdoor environment contains illumination across a wide range of spectrum, so the simple idea to apply filter to obtain images in those spectrum other than visible band will not lead to an image which is independent of environmental illumination.
However, a technique called *Active Differential Imaging* can be applied to solve the illumination variation problem completely, even for the outdoor application. In active differential imaging, an image is obtained by the pixel-to-pixel differentiation between two successive frames: one with an active illuminant on and one with the illuminant off. The difference image of these two frames contains only the scene under the active illuminant, and is completely independent of the environmental illumination. This technique and its application in face recognition will be presented in the following chapters.
Chapter 4

Near-IR Active Differential Imaging for Face Recognition

4.1 Introduction

Near-IR active differential imaging is applied in this thesis to obtain ambient illumination invariant face images for face recognition. Active differential imaging is an important variant of active sensing to minimize the illumination problem. Assuming still scene, linear response of the camera to the scene radiation and no saturation, it can be shown that the active differential imaging yields a face image independent of ambient illumination. Face recognition experiments carried out on the face images captured by our Near-IR active differential imaging system achieved very low error rates even in the scenario with dramatic ambient illumination changes. These results are significantly better than the results on the corresponding ambient faces images even if an advanced illumination normalisation technique is applied to photometrically normalise ambient face images or the recognition is based on illumination insensitive features extracted from ambient face images.

The Near-IR active differential imaging is introduced in Section 4.2, the modelling of the imaging process is presented in Section 4.3. Face recognition applications using Near-IR active differential imaging are reviewed in Section 4.4. In Section 4.5 a face capture
system based on Near-IR active differential imaging is presented and a face database captured by this system is described. The details and results of face identification and verification experiments are presented in Section 4.6, and conclusions are drawn in Section 4.7.

4.2 Near-IR Active Differential Imaging

Differential imaging is a technique to obtain the difference image of two frames, which is widely used to detect a moving object and remove the static background based on the assumption that the intensity change of a pixel is only caused by the difference in the reflectance properties of the moving object and the background. Active differential imaging also involves frame differencing, however, the two frames are not under the same illumination. An active differential imaging system consists of an active illuminant and an imaging sensor. During the imaging process, two images are captured: the first one is taken when the active illuminant is on, the second one is taken when the illuminant is off. In the first image the scene is illuminated by the combination of both the ambient illumination and the active illumination, while in the second image the scene is illuminated only by ambient illumination. Assuming no motion takes place in the scene, no saturation, and a linear response of the sensor to the scene radiation, the difference of these two images then contains the scene viewed under the active illumination only, which is completely independent of the ambient illumination. The active illuminant is generally fixed on the camera, therefore all the difference images captured by the active differential imaging system are under the same illumination direction. Figure 4.1 demonstrates the idea of active differential imaging.

Specific sensors which can perform differential imaging have been proposed\[73\]\[110]\[78\]. The first dedicated CMOS imager is proposed in \[73\], where two analog memories for each pixel are used; one for the actively illuminated frame, and one for the ambient frame. The image differencing is performed at the image readout phase. In \[110\], a high speed frame-mode image sensor with a digital frame buffer and a digital subtractor is introduced to perform differential imaging. Ni and Yan \[78\] designed a CMOS active differential imaging device with single in-pixel analog memory. The image sensor can
Figure 4.1: Active differential Imaging: ambient illumination is removed by taking the
difference between an actively illuminated frame and an ambient frame.

work in frame-mode or line-mode. For the actively illuminated frame, the pixel array
is exposed in a frame mode. For the ambient frame, the line-mode is adopted. The
first line of pixel of the actively illuminated frame is loaded to a line buffer, then
reset and exposed to ambient illumination and loaded to another line buffer. After
the differencing operation involving these two line buffers is executed, the next line
is processed following the same steps as the previous line. All these operations are
performed on-chip, and the output of the chip is the difference image.

Theoretically the active illumination in active differential imaging can be any reflective
illumination. However the reflective infrared would be preferable due to its invisibility.
We use Near-Infrared LEDs as the active illumination source, because they are cheap
and easily available in the market, and every normal black-and-white CCD sensor has
response to Near-Infrared spectrum.

4.3  Image Formation Analysis for Active Differential Imaging

The process of image formation can be analyzed based on radiometry and sensor mod-
ing. As shown in Figure 4.2, an object is illuminated by one or more incident il-
mination sources. The reflection from the object surface reaches the camera sensor
through the lens. Photon-electron conversion process takes place in the sensor and
Chapter 4. Near-IR Active Differential Imaging for Face Recognition

Figure 4.2: Image formation.

after the electronic amplification and analogue-digital conversion, the digital image is obtained with quantized brightness value for each pixel.

4.3.1 Radiometry Analysis

The irradiance received by an object surface point $P$ due to the incident radiance $L_i$ can be defined as [34]

$$Ir = L_i(P, \theta_i, \phi_i) \cos \theta_i \delta \omega_i$$

(4.1)

where $Ir$ describes the incident power per unit area not foreshortened, $L_i$ stands for the incident power per unit solid angle, per unit area perpendicular to the incident direction indicated by $(\theta_i, \phi_i)$. $\theta_i$ is the angle between incident direction and the surface normal $\vec{n}$ at $P$. $\delta \omega_i$ is the solid angle subtended by the illumination source with respect to $P$.

The local reflection can be modelled by Bidirectional Reflectance Distribution Function (BRDF), which is defined as the ratio of the radiance in the outgoing direction to
4.3. Image Formation Analysis for Active Differential Imaging

the incident irradiance. Based on the definition, the outgoing radiance $L_0$ from $P$ in
the direction indicated by $(\theta_o, \phi_o)$ can be represented by

$$L_0(P, \theta_o, \phi_o) = \rho_{bd}(P, \theta_i, \phi_i, \theta_o, \phi_o)I_r = \rho_{bd}(P, \theta_i, \phi_i, \theta_o, \phi_o)L_i(P, \theta_i, \phi_i) \cos \theta_i \delta \omega_i$$  \hspace{1cm} (4.2)

where the $\rho_{bd}(P, \theta_o, \phi_o, \theta_o, \phi_o)$ denotes the BRDF. Considering multiple illumination
sources, and denoting the outgoing radiance $L_0^m$ due to the $m$th illumination source $S^m$, the incident light can be expressed by:

$$L_0(P, \theta_o, \phi_o) = \sum_m L_0^m(P, \theta_o, \phi_o)$$  \hspace{1cm} (4.3)

Based on the thin lenses model, the irradiance $E$ arrived at the image point $P'$ on the
image plane can be shown to be [34]:

$$E(P') = \left[ \frac{\pi}{4} \left( \frac{d}{z'} \right)^2 \cos^4 \alpha \right] L_0(P, \theta_o, \phi_o)$$ \hspace{1cm} (4.4)

which is proportional to the reflected radiance $L_0$ from $P$. For all points in the image
plane, both the effective diameter of lens $d$ and the distance $z'$ between the lens and
image plane are constant. Meanwhile, considering the small size of the actual sensor,
the angle $\alpha$ between the light ray $PP'$ and the optical axis is close to zero, and therefore
$\cos^4 \alpha$ can also be taken as a constant. We can then express $E(P')$ as:

$$E(P') = kL_0(P, \theta_o, \phi_o) = k \sum_m L_0^m(P, \theta_o, \phi_o)$$  \hspace{1cm} (4.5)

For Lambertian surface (also called ideal diffusion surface), the BRDF is independent
of the incident and outgoing directions. The directional hemispheric reflectance, which
is defined to measure the fraction of the incident irradiance from a given direction that
is reflected by the surface whatever the direction of reflection, is also independent of
directions and called diffuse reflectance or albedo.

$$\rho_{dh}(\theta_i, \phi_i) = \int L_o(P, \theta_o, \phi_o) \cos \theta_o d\omega_o = \int \rho_{bd}(P, \theta_i, \phi_i, \theta_o, \phi_o) \cos \theta_i d\omega_i$$  \hspace{1cm} (4.6)

where $\Omega$ is the exit hemisphere. It can be shown that the albedo $\rho$ and the BRDF $\rho_{bd}$
of a Lambertian surface point satisfy:

$$\rho_{bd} = \rho / \pi$$  \hspace{1cm} (4.7)
4.3.2 Sensor Model

In this section, we consider the sensor model for black-and-white CCD camera. Assuming the illumination is not strong enough to cause the overflow of the electrons stored in the cell (i.e., image is not saturated at this pixel), the number of electrons recorded at the cell located at row \( r \) and column \( c \) of a CCD array can be modelled as [34]

\[
I(r, c) = T \int_{\lambda} \int_{P' \in S(r, c)} E(P', \lambda) R(P') q(\lambda) dP' d\lambda
\]

(4.8)

Here, \( T \) is the integration time for electron collection, and the collection is for all points \( P' \) within area \( S(r, c) \). \( \lambda \) is the wavelength of the incident light, \( R \) is the spatial response, and \( q \) is the quantum efficiency of the device, which stands for the number of electrons generated by per unit of incident light energy. The photon-electron conversion process is influenced by the uncertainty introduced by the quantum phenomenon. The brightness of pixel \( (r, c) \) in the image can be modelled as [34]

\[
B(r, c) = \gamma (N_I(r, c) + N_{DC}(r, c) + N_B(r, c) + R_d(r, c)) + Q(r, c)
\]

(4.9)

\( \gamma \) is the combined gain of the amplifier and camera circuitry. \( N_I(r, c) \) is a random integer variable obeying a Poisson distribution with mean \( \beta(r, c) I(r, c) \), where \( \beta(r, c) \) reflects the variation of the spatial response and quantum efficiency. The random integer variables \( N_{DC}(r, c) \), \( N_B(r, c) \), \( R_d(r, c) \) and \( Q(r, c) \) reflect the influences of dark current, additional electrons introduced by the CCD electronics, read-out noise due to the amplifier and the quantization noise. Ignoring the noise terms and the spatial response variation within \( S(r, c) \), the brightness can be expressed as

\[
B(r, c) = \gamma \beta(r, c) I(r, c) = T \gamma \beta(r, c) S \int_{\lambda} E(r, c, \lambda) q(\lambda) d\lambda
\]

(4.10)

\[
= T \gamma \beta(r, c) S k \int_{\lambda} \sum_m [E^m_0(P, \theta_o, \phi_o, \lambda)] q(\lambda) d\lambda
\]

\[
= \sum_m [T \gamma \beta(r, c) S k \int_{\lambda} E^m_0(P, \theta_o, \phi_o, \lambda)] q(\lambda) d\lambda
\]

\[
= \sum_m B_m(r, c)
\]

where \( B_m(r, c) \) is the brightness due to the \( m \)th illumination source \( S^m \).

4.3.3 Near-IR Difference Image

In the proposed imaging approach, at time \( t \) we capture one image with only ambient illumination, denoted by \( B_{amb}(t) \), then turn on the active Near-IR illuminant, and
capture another image $B_{combined}(t + \Delta t)$ at time $t + \Delta t$. Based on Equation 4.10 and the assumption that the image under the combined illumination of ambient and active Near-IR light sources is unsaturated, we have:

$$B_{combined}(t + \Delta t) = B_{amb}(t + \Delta t) + B_{led}(t + \Delta t)$$

(4.11)

The difference image of the above two images is

$$B_{diff}(t, \Delta t) = B_{combined}(t + \Delta t) - B_{amb}(t) = B_{amb}(t + \Delta t) + B_{led}(t + \Delta t) - B_{amb}(t)$$

(4.12)

Assuming there is no change on the image brightness caused by any motion or ambient illumination variation during the time interval of $\Delta t$, which means $B_{amb}(t) = B_{amb}(t + \Delta t)$, the difference image $B_{diff}$ satisfies:

$$B_{diff} = B_{led}(t + \Delta t)$$

(4.13)

Therefore the difference image is the face illuminated only by the active illumination source, and thus independent of the ambient illumination. Note that as Equation 4.10 holds for an arbitrary object surface, Equation 4.13 does not require the Lambertian surface assumption for faces, which is the essential assumption for most illumination-invariant face recognition approaches based on physical modelling of illumination [8, 9, 38, 5, 88, 127, 96].

## 4.4 Active Differential Imaging for Face Recognition

When applying active differential imaging to face recognition, with the illuminant fixed on the camera, all the difference face images in both gallery set and probe set are under a consistent frontal illumination from the active illumination source. The problem caused by variation in the environmental illumination for face recognition is completely solved. We applied this idea for illumination-invariant face recognition and achieved much better results on the difference face image compared to the ambient faces[137][139] when ambient illumination varies.

Contemporary to our work, Ni et al. [77], Hizem et al. [47] performed face recognition experiment using their specific sensor which performs active differential imaging. In
Chapter 4. Near-IR Active Differential Imaging for Face Recognition

[77], a face database of 25 subjects is captured indoor with two cameras: an active differential imaging camera and a normal CCD camera. Faces are exposed to four different illuminations conditions: Normal Light (with ambient daylight and artificial light), No Light (with little daylight and no artificial light), Frontal Light (with ambient daylight and frontal artificial light), and Side Light (with daylight and artificial light from right side). 10 images for each person are captured under each illumination condition, and five of them are used for training and the rest for testing. The face preprocessing includes face region segmentation based on intensity binarization, geometrical normalisation based on face region size, photometrical normalisation based on histogram equalisation, and Gaussian smoothing. Face verification experiment is carried out on the processed face images across every two different illuminations. The correlation score with Euclidean distance in image intensity domain is applied as a distance measure. Even with the simple preprocessing and face recognition algorithm, a prominent advantage of active differential imaging is demonstrated by the experimental results. In [47] a face database captured with a synchronized flash infrared camera (see Section 3.3.2) is added to the one in [77]. All faces are manually registered, histogram equalised and filtered with a Gaussian mask. For faces captured by normal CCD camera, an additional preprocessing stage is applied to remove the illumination variation by averaging the left part and the right part of a face. This preprocessing did help a lot for faces captured by normal camera, however, the best result is still achieved on the images captured by active differential imaging camera, and it is much better than the results on faces captured by normal camera or synchronized flash infrared camera.

A thorough investigation of applying active differential imaging to overcome illumination variation in face recognition is presented in this thesis. Our work is different to the work by Ni et al. in [77] and Hizem et al. in [47] in several aspects.

- The face data is captured in a different way. Our system outputs both images under ambient illumination and those under combined illumination. Assuming no motion taking place, the ambient face image and the corresponding difference image are perfectly registered and in exactly the same pose. The only difference between an ambient image and the corresponding difference image is the difference in illumination condition, which is the only factor which causes the difference
between the recognition performance on difference face image and that on ambient face images. And therefore, the face recognition experiments on the ambient face images and difference images captured by our system provides a strict justification for the advantage of active differential imaging in dealing with the illumination problem. In contrast, the active differential imaging sensor in [77][47] outputs only difference image. The ambient face images and difference face images are captured by two different cameras. If there are more or less pose variation between the times instances when the ambient images and the difference image are acquired, this pose variation will contribute to the difference between face recognition results on difference images and ambient images, which might unrealistically magnify the apparent advantage of active differential imaging.

- The active differential imaging approach is compared with several state-of-the-art illumination invariant face recognition approaches. These approaches include Fisherface method, a photometric normalisation technique in which the luminance function is estimated by anisotropic smoothing, and illumination insensitive approaches based on phase and Fisherphase. Those methods have been shown to have good performance in addressing the illumination problem. In contrast, in [47] comparison was made between the active differential imaging approach and a simple illumination correction method adopted based on face symmetry. The superiority of active differential imaging approach over those advanced illumination-invariant approaches, as shown in our work, demonstrates more convincingly the advantage of the active differential imaging over the conventional approaches in overcoming the illumination problem.

- We applied a more advanced face recognition algorithm. It has been shown that the Fisherface with normalised correlation approach is usually superior to the simple correlation method based on Euclidean distance adopted in [77][47].

- We also have the data captured at two different times so that we can measure the performance across time lapse.

- Both verification and identification experiments are carried out.
We developed a specific automatic eye localisation approach for face localisation and carried out face recognition in automatic scenario based on this localisation approach, as will be presented in Chapter 6.

- We addressed the problem caused by object motion when applying active differential imaging, which will be presented in Chapter 7.

### 4.5 Capture System and Face Database Acquisition

The hardware of our face capture system consists of an LED light source, a camera, the frame grabber card and a computer. The LED lamp in our system has a set of 42 LEDs with adjustable output and a photocell. The LED beam angle is 30 degrees. The peak output wavelength is 850nm, lying at the invisible and reflective light portion of the Electromagnetic spectrum.

The sensor we use is the HITACHI KP-M3RP CCD camera, which is a monochrome device with Near-IR sensitivity (see Figure 4.3). The peak sensitivity lies at 620nm.
4.5. Capture System and Face Database Acquisition

and the sensitive spectrum ranges from less than 400nm to 1000nm. To ensure the difference of the images taken when the LED lamp is on and off represents the face illuminated just by the LED lamp, the Automatic Gain Control (AGC) in the camera is disabled and the aperture is tuned to ensure that the camera will have a linear response to the scene radiance, and that there is no saturation in the captured images. If the images are unsaturated, the face image under the combined LED and ambient illumination is exactly the addition of the face images obtained separately under just LED illumination and under just the ambient illumination, respectively.

The LED lamp is attached close to the camera so that the reflective component of the Near-IR light from the eyes will be projected straight into the camera. This allows us to obtain face images with prominent bright pupils. A ring with 4 fluorescent lamps is used to provide varying illumination conditions. This is not a part of the envisaged practical system but just for the purpose of capturing a dataset on which the comparison with the LED illuminated face can be made. Figure 4.4 shows a picture of the face capture system for our experiments.
A database of face images of 40 subjects has been captured indoors using the system described in the previous Section. This database contains two subsets: ambient faces (face images under ambient illumination) and LED faces (face images under LED illumination). Two capture sessions were held over a period of several weeks. For each session, 4 different illumination configurations are used with light sources directed individually from the left, bottom, right and top. 6 recordings were acquired for each illumination configuration. For each recording a face image under ambient illumination only and one under combined ambient and LED illumination are captured. Therefore, we have $40 \times 2 \times 4 \times 6 = 1920$ ambient faces and the same number of ambient plus LED faces. 1920 LED faces are then obtained as difference images. The size of each image is 382*288 pixels. Examples of images are shown in Figures 4.5, 4.6, and 4.7.

During each recording, the subject was asked to sit as still as possible, without talking, in front of the lighting ring, facing frontally the camera. Just one of those four lamps was switched on, and the LED lamp was turned on and off. We captured face images before and after the LED lamp was turned on to obtain the set. Assuming little or no movement of the subject between the takes, we can then compute the difference image of the above two images to obtain the face under just LED illumination, which is independent of the ambient illumination. The face is in exactly the same pose in both ambient face image and LED face image within the same recording, which means the illumination conditions are the only distinctive factor in the image formation. Hence the difference in the face recognition results on the two sets of images respectively are solely due to the difference in the illumination conditions rather than any other factors such as variations in pose and expression. Meanwhile, although each ambient face is in the same pose with the corresponding LED face, there are small pose variations among all ambient faces under the same or different illumination conditions. This is also true for LED faces. This makes the database more practical compared to CMU-PIE database, in which all face images of the same subject within the same subset (the *Illum* subset or the *Lights* subset) under different illumination directions are in exactly the same pose.
Figure 4.5: Ambient faces (the left column), combined illumination faces (the middle column) and LED illuminated faces (the right column) under 4 different illumination configurations. The changes in ambient illumination caused significant differences in the appearance of the whole face. All important facial features look very different in different illumination conditions. Ambient faces and LED faces are relatively dark because the aperture of the lens is adjusted to avoid the saturation of the combined illuminated faces.
Figure 4.6: Ambient face images of two subjects for the two different sessions.
Figure 4.7: LED face images of two subjects for the two different sessions.
Figure 4.8: Resulting images after the histogram equalization is performed for manually registered ambient faces ((a)(b)) and for corresponding LED faces ((c)(d)). It is obvious that data from LED faces exhibits much less variation as compared to the data from ambient faces. Bright pupils are prominent in LED faces.

4.6 Face Recognition Experiments on Manually Registered Data

4.6.1 Face Identification Experiment

The first set of face identification experiments is performed on manually registered ambient faces and LED faces in order to investigate the advantages of the proposed imaging approach under the situation of uneven illumination. As mentioned in Section 4.5, the corresponding ambient face and LED face are in the same pose. Each pair of these two images are geometrically registered and cropped according to the same manually marked eye locations.

For identification experiments, the whole dataset was divided into different subsets to serve as training sets and test sets in different test protocols[139]. The rules of naming
Table 4.1: Training and test set configurations for tests on manually registered data

<table>
<thead>
<tr>
<th>Test Protocol</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Session</td>
<td>$MS_{j_0}$</td>
<td>$MS_{j \neq j_0}$</td>
</tr>
<tr>
<td>Cross Illumination</td>
<td>$MC_{i_0}$</td>
<td>$MC_{i \neq i_0}$</td>
</tr>
<tr>
<td>Combined test</td>
<td>$MC_{i_0}S_{j_0}$</td>
<td>$MC_{i \neq i_0}S_{j \neq j_0}$</td>
</tr>
</tbody>
</table>

A subset are listed as below:

- $S_i$ for data in Session $i$, $i = 1, 2$;
- $C_i$ for data in Illumination Condition $i$, $i = 1..A$;
- $M$ for manually registered data, $A$ for automatically registered data.

The three tests on manually marked data are Cross Session test, Cross Illumination test and Combined test (cross both session and illumination). Table 4.1 shows the training set and test set configuration for these three tests. In a Cross Session test the training set and the test set are from different sessions. In a Cross Illumination test, the training set contains data with one illumination and the test set contains data acquired in the other illumination conditions. The error rate under a specific test protocol is the average error among all tests under this protocol. For example, the test error under the first protocol is the average of the errors of 2 subtests. In one of these two subtests, data from Session 1 is used for training and Session 2 data for testing, while in the other, Session 2 data is used for training and Session 1 data for testing.

The face identification tests are carried out based on different face representations and classifiers. These representations include two in the intensity domain and two in the phase domain:

1 Fisherface [8],
2 Fisherface after anisotropic smoothing [41],
3 Phase [92][44], and
4 Fisherphase.

All the fisher subspaces involved in the representations No.1, No.2, and No.4 are determined using the XM2VTS face database. PCA retaining 95% energy of the data is applied for dimensionality reduction before the linear discriminant analysis. For representations No.1, histogram equalization is adopted to normalise face data photometrically. For representation No.2, a more advanced photometrical normalisation approach, in which an anisotropic smoothing is used to estimate the luminance function, is applied to preprocess the full images. The two classifiers we adopted are a SVM classifier with linear kernel and a Nearest Neighbor classifier with normalised correlation as the similarity measure. For comparison, Affinity™ face authentication SDK from Omnipercetion [80] is also applied to both ambient faces and LED faces.

The identification results are reported in Table 4.2, Table 4.3 and Table 4.4. Several observations can be found from these results:

1. Regardless of the choice of face representations and classifiers, identification on LED faces achieved much lower error rates than on ambient faces for all the three test protocols.

2. The best results on LED faces are achieved in the intensity domain. Those results are close to perfect. The anisotropic smoothing offers a small or no improvement in the results of three test protocols on LED faces, which confirms experimentally that illumination variation in ambient illumination has no influence on LED faces. In sharp contrast, the illumination difference caused severe problem for ambient faces. This can be found from Table 4.3 and Table 4.4. The identification performance on ambient faces is improved greatly by applying photometric normalisation, but it is still far worse than that on LED faces.

3. Compared to the results achieved using the representations in the original intensity domain, those based on the phase domain information are disappointing, although very impressive performance of phase-based approaches and their superiority over Fisherface, Eigenface and 3D linear subspace approaches is reported in [92][44] on CMU-PIE database. According to our experiment results, the phase
Figure 4.9: Images after the photometric normalisation based on anisotropic smoothing. 
(a) ambient face images with lighting from four directions. (b) corresponding LED faces.
Chapter 4. Near-IR Active Differential Imaging for Face Recognition

Figure 4.10: Face regions after cropping and histogram equalisation for images normalised photometrically by the method based on anisotropic smoothing. (a) two sessions of ambient face images of a subject with lighting from four directions, (b) corresponding LED faces.

representations are very sensitive to appearance change which can be shown from the poor performance in the Cross Session test and the Combined test. In the Cross Illumination test, the phase-based recognition does not deliver the expected advantages over ambient faces. A possible reason for the poor performance is that the phase-based method relies heavily on the perfect registration. In our database, although faces are manually registered, there is variation in pose and eye localisation among faces, which contrasts with the exactly same pose and eye position in faces in CMU-PIE database. This explains the big difference between the identification performance on the two databases.

4. The Affinity™ face authentication SDK gave the best performance among all identification approaches on both LED faces and ambient faces. Zero error rates are reached in the Cross Session test and Cross Illumination test. Even in the Combined test, which is the most difficult test, an error rate as low as 0.07% is achieved.

4.6.2 Face Verification Experiments

Face verification experiments are carried out on manually registered data to investigate the advantage brought by active differential imaging. We use the same four repre-
### 4.6. Face Recognition Experiments on Manually Registered Data

#### Table 4.2: Average error of Cross Session tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Classifier</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>SVM</td>
<td>1.61</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>0.83</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface</td>
<td>SVM</td>
<td>0.32</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Anisotropic Smoothing</td>
<td>N.N.</td>
<td>0.37</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>SVM</td>
<td>9.53</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>10.42</td>
<td>5.47</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>SVM</td>
<td>19.27</td>
<td>10.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>25.32</td>
<td>9.84</td>
</tr>
<tr>
<td>5</td>
<td>Affinity SDK</td>
<td></td>
<td>0.21</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Table 4.3: Average error of the Cross Illumination tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Classifier</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>SVM</td>
<td>42.57</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>34.96</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface</td>
<td>SVM</td>
<td>24.62</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Anisotropic Smoothing</td>
<td>N.N.</td>
<td>16.84</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>SVM</td>
<td>54.71</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>55.35</td>
<td>1.51</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>SVM</td>
<td>59.41</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>56.75</td>
<td>4.01</td>
</tr>
<tr>
<td>5</td>
<td>Affinity SDK</td>
<td></td>
<td>7.15</td>
<td>0</td>
</tr>
</tbody>
</table>
sentations as in Section 4.6.1 and the normalised correlation as a distance measure. Since the database contains only 40 subjects, we conducted the verification experiment based on the Leave-One-Out protocol. Similar to the identification experiments in the previous section, we investigated the scenario across different time, across different illumination and across both time and illumination. Three verification test configurations, namely the Cross Session test, the Gross Illumination test and the Combined test, are adopted. Each of these tests consists of a number of element tests. Each element test for the Cross Session test corresponds to a specific subject and a specific session, each element test for the Gross Illumination test corresponds to a specific subject and a specific illumination, and each element test for the Combined test corresponds to a specific subject, a specific illumination and a specific session.

For each element test, each database (ambient faces or LED faces) is divided into 4 independent data sets, namely Enrolment set, Training set, Evaluation set and Test set. Following the naming rules in Section 4.6.1, and using \( I_i \), \( i \in \{1, \ldots, 6\} \) for the \( i \)th image of a certain session under a certain illumination, \( ID_i \), \( i \in \{1, \ldots, 40\} \) for the \( i \)th subject, the protocols are described in Table 4.5.

**Enrollment set** contains only images of current subject \( ID_{i0} \), and the template of this subject is obtained from this set.

### Table 4.4: Average error of the Combined tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Classifier</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>SVM</td>
<td>52.95</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>43.30</td>
<td>0.19</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface</td>
<td>SVM</td>
<td>34.39</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Anisotropic Smoothing</td>
<td>N.N.</td>
<td>25.82</td>
<td>0.28</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>SVM</td>
<td>67.55</td>
<td>13.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>65.21</td>
<td>10.78</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>SVM</td>
<td>72.17</td>
<td>23.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N.N.</td>
<td>69.1</td>
<td>19.62</td>
</tr>
<tr>
<td>5</td>
<td>Affinity SDK</td>
<td></td>
<td>11.38</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table 4.5: Data set configurations for tests on manually registered data

<table>
<thead>
<tr>
<th></th>
<th>Cross Session test</th>
<th>Cross Illumination test</th>
<th>Combined test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enrolment Set</strong></td>
<td>$MID_{i_o} S_{j_o}$</td>
<td>$MID_{i_o} C_{p_o}$</td>
<td>$MID_{i_o} C_{p_o} S_{j_o}$</td>
</tr>
<tr>
<td><strong>Training Set</strong></td>
<td>$MID_{i \neq i_0} S_{j_o}$</td>
<td>$MID_{i \neq i_0} C_{p_o}$</td>
<td>$MID_{i \neq i_0} C_{p_o} S_{j_o}$</td>
</tr>
<tr>
<td><strong>Evaluation Set</strong></td>
<td>$MID_{i \neq i_0} S_{j \neq j_0} I_{i=1,..,4}$</td>
<td>$MID_{i \neq i_0} C_{p \neq p_0} I_{i=1,..,4}$</td>
<td>$MID_{i \neq i_0} C_{p \neq p_0} S_{j \neq j_0} I_{i=1,..,4}$</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td>$MID_{i \neq i_0} S_{j \neq j_0} I_{i=5,6} \cup MID_{i_o} S_{j \neq j_0}$</td>
<td>$MID_{i \neq i_0} C_{p \neq p_0} I_{i=5,6} \cup MID_{i_o} C_{p \neq p_0}$</td>
<td>$MID_{i \neq i_0} C_{p \neq p_0} S_{j \neq j_0} I_{i=5,6} \cup MID_{i_o} C_{p \neq p_0} S_{j \neq j_0}$</td>
</tr>
</tbody>
</table>

**Training set** consists of some images of subjects $\{ID_{i \neq i_0}\}$ to build the template for each of those subjects.

**Evaluation set** is the collection of some images of subjects $\{ID_{i \neq i_0}\}$ to make true and false claims to the templates in the training set. The threshold which gives the best separation for true claim scores and false claim scores in the evaluation stage will be applied to the testing stage to make the decision to accept or reject a claim of identity $ID_{i_0}$.

**Test set** is a combination of an imposter test set, which contains some images of subjects $\{ID_{i \neq i_0}\}$ to make false claims to identity $ID_{i_0}$, and a client test set, which contains some images of identity $ID_{i_0}$ to make true claims to subject $ID_{i_0}$.

The numbers of true claims and false claims in the evaluation stage and testing stage are provided in Table 4.6.

Recall we have two session data of 40 subjects, 4 illumination conditions for each session, 6 images for each illumination condition. Taking the Cross Session test as an example: for the subject $ID_{i_0}$, session $S_j$, the training set contains all the images of the other subjects in session $S_j$, the template for each of these 39 subjects is obtained from the $4 \times 6 = 24$ images belonging to each of them. In the other session 4 out of 6 images of these subjects under all 4 illuminations for each subject constitute the evaluation set, and the other 2 images is used as test set to make false claim to the identity $ID_{i_0}$. During the evaluation stage, all images in the evaluation set are used to claim each of
Chapter 4. Near-IR Active Differential Imaging for Face Recognition

Table 4.6: No. of true claims/ false claims/ element tests

<table>
<thead>
<tr>
<th></th>
<th>Cross Session test</th>
<th>Cross Illum. test</th>
<th>Combined test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval</td>
<td>No of true claim</td>
<td>4<em>4</em>39=624</td>
<td>4<em>3</em>2*39=936</td>
</tr>
<tr>
<td></td>
<td>No of false claim</td>
<td>4<em>4</em>39*38=23712</td>
<td>4<em>3</em>2<em>39</em>38=35568</td>
</tr>
<tr>
<td>Testing</td>
<td>No of true claim</td>
<td>4*6=24</td>
<td>2<em>3</em>6=36</td>
</tr>
<tr>
<td></td>
<td>No of false claim</td>
<td>4<em>2</em>39=312</td>
<td>2<em>2</em>3*39=468</td>
</tr>
<tr>
<td></td>
<td>No of Element tests</td>
<td>40*2=80</td>
<td>40*4=160</td>
</tr>
</tbody>
</table>

the 39 subjects identities in the training set, which means we have 4*4*39=624 true claims and 4*4*39*38=5928 false claims during evaluation. The template of subject ID_{i0} is obtained from the 24 images in session S_j, and the 24 images of himself/herself in the other session are used to make true claims to this template. So we have 24 true claims and 2*4*39=312 false claims for subject ID_{i0} in session S_j in the testing stage. This element test is repeated for all 40 subjects and 2 sessions. Thus in total there are 24*40*2=1920 true claims and 312*40*2=24960 false claims. The EERs in all 40*2=80 element tests are averaged and reported.

The distributions of the true claims scores and the false claim scores in the testing stages of the three face verification tests based on Fisherface are shown in Figure 4.11. It can be observed that for any test on ambient faces, the histogram of true claim scores has a large overlap with the histogram of false claim scores. In sharp contrast, this overlap for LED faces is always very small, which means the true/false claims based on LED faces are much easier to distinguish. Therefore much lower error rate of the verification system can be achieved.

The ROC curves of the three face verification tests are presented in Figures 4.12, 4.13 and 4.14, and the EERs of those tests are provided in Tables 4.7, 4.8, 4.9. These results reflect the same observations we have found in the identification experiments:

1. Independent of the choice of face representation, all the experiments on LED faces produced much better results than on ambient faces.
Figure 4.11: The distributions of the true claim and the false claim scores for ambient faces and LED faces in the Cross Session test (a)(b), the Cross Illumination test(c)(d) and the Combined test(e)(f) based on Fisherface representation.
Table 4.7: Average EER of the Cross Session tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>16.02</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface with Anisotropic Smoothing</td>
<td>4.14</td>
<td>1.08</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>10.62</td>
<td>8.14</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>12.96</td>
<td>8.74</td>
</tr>
</tbody>
</table>

Table 4.8: Average EER of the Cross Illumination tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>31.39</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface with Anisotropic Smoothing</td>
<td>17.44</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>25.26</td>
<td>4.25</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>25.58</td>
<td>3.96</td>
</tr>
</tbody>
</table>

2. For the Cross Illumination tests on ambient faces, anisotropic smoothing based photometric normalisation offered a significant improvement over the histogram equalisation. Lower EERs are achieved based on Phase and Fisherphase than Fisherface. However, adopting any of these illumination insensitive representations of ambient faces does not lead to ERRs which is comparable to the ERRs on LED faces.

3. The best results for all tests are obtained on LED faces based on Fisherface, which suggests that this intensity domain representation is the best choice when there is no illumination variation.

4. In the Combined test, which is the most difficult test, the Fisherface on the LED faces can offer the lowest ERR rate of 1.13%.
Figure 4.12: ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Cross Session test.
Figure 4.13: ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Cross Illumination test.
Figure 4.14: ROC curves for face verification based on different representations of ambient faces (a) and LED faces (b) in the Combined test.
Table 4.9: Average EER of Combined tests on manually registered data (in %)

<table>
<thead>
<tr>
<th>No.</th>
<th>Representation</th>
<th>Ambient Faces</th>
<th>LED Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fisherface</td>
<td>32.28</td>
<td>1.13</td>
</tr>
<tr>
<td>2</td>
<td>Fisherface with Anisotropic Smoothing</td>
<td>18.53</td>
<td>2.06</td>
</tr>
<tr>
<td>3</td>
<td>Phase</td>
<td>28.50</td>
<td>10.87</td>
</tr>
<tr>
<td>4</td>
<td>Fisherphase</td>
<td>29.34</td>
<td>10.92</td>
</tr>
</tbody>
</table>

4.7 Conclusion

The active differential imaging technique for illumination invariant face recognition is investigated in this chapter. Through the radiometry analysis and sensor modelling, it can be shown that an ambient illumination independent image of a still object, Lambertian or not, can be obtained by the active differential imaging. A face database, which contains face images under ambient illuminations from different directions and corresponding face images under only active illumination, has been captured based on active differential imaging. The face identification experiments and verification experiments on manually registered faces in this database have proved the significant advantage of active differential imaging for scenarios with illumination changes. It has been shown that applying an advanced photometric normalisation method or illumination insensitive feature can improve the performance of face recognition system under varying illumination but the improvement is limited. It is evident that employing active differential imaging to obtain ambient illumination invariant face images is preferable to trying to remove illumination variation from ambient face images captured under varying illumination.
Chapter 5

Literature Review on Face Localisation

5.1 Introduction

Automatic face localisation is the first step of an automatic face recognition system and its performance is crucial for the performance of the whole system. Face localisation is closely related to face detection, which has been addressed in many papers and carefully reviewed in [48][123]. The role of face detection is to detect the existence of faces in a given image and give the coarse positions of the faces detected. However, in face localisation the existence of a face is assumed, and the aim is to precisely localise the positions of facial features for the purpose of recognition. Face detection and localisation are rather challenging due to both the intrinsic variations in face appearance associated with identity, age, gender and facial expression, and extrinsic variations caused by changes in view point, illumination, and imaging system [39].

Most of the “passive” approaches to face localisation from a color/graylevel face image can be classified into top-down approaches or bottom-up approaches. The difference between these two groups is based on whether local facial feature information is used before or after the holistic face info. The exceptions from the above two groups are the approaches based on Active Shape Model/Active Appearance Model [112], in which
the face detection and facial feature localisation are achieved simultaneously by an optimisation process. Moreover, some active approaches have also been proposed, where face properties such as the skin response to invisible spectrum, retro-reflectivity of eyes are exploited to facilitate face detection/localisation. A review of top-down, bottom-up and active approaches is presented in Sections 5.2.1, 5.2.2 and 5.3.

The eyes are the most important facial features among all facial features that could be localised. Geometric normalisation of faces is usually based on eye centers. The eye center can be defined as the middle location between the locations of the left corner and the right corner of each eye. This definition is invariant to the eye-gaze direction. However, in a cooperative scenario where clients are looking into the camera, the pupil locations can serve as the basis for the geometric normalisation instead of the physical eye center locations. The problem of eye localisation commands special interest not only because eye locations are crucial for face recognition, but also because they are important for human-machine interface applications. The eye localisation is a necessary step in any face localisation approach. We presents a review of eye localisation techniques in Section 5.2.3.

A face localisation performance measure is introduced in Section 5.4 and the chapter is concluded in Section 5.5.

5.2 Passive Approaches

5.2.1 Top-down Approaches

Most approaches to face localisation can be considered as top-down approaches which can be divided into two stages: the face detection stage and the stage of facial feature localisation. A holistic face model or skin color model is usually used for the face detection stage. For the numerous face detection methods please refer to the survey provided in [48][123]. Viola and Jones [114] proposed a rapid face detector based on Haar-like features and cascaded AdaBoost classifiers. They introduced so-called “Integral Image” which allows fast computation of Haar-like features. Li et al. [67] presented a
5.2. Passive Approaches

FloatBoost based face detector, in which the floating search of feature selection is incorporated into AdaBoost. Cristinacce et al. [23] applied the face detector developed by Viola and Jones [114] to find the approximate scale and location of the face. AdaBoost classifiers are then applied to locate facial features in the corresponding regions of the detected face, and facial feature candidates are combined using pairwise probability constraints. Finally, the facial feature locations are refined by an edge/corner Active Appearance Model. Jesorsky et al. [52] proposed a coarse face detection based on the Hausdorff distance to an edge-based face model, then an eye model is used to refine the initially detected position. A Multi-Layer Perceptron trained with pupil-centred images is further applied for refinement.

5.2.2 Bottom-up Approaches

Another group of techniques for face localisation are bottom-up feature-based approaches in which facial feature candidates are searched directly in the whole image before being verified by global face models. Harnouz et al. [43] use Gabor filters based detectors to localise features, including eye corners, eye centres, nostrils and mouth corners. Face hypotheses are generated in the affine space using triplets of facial feature candidates. An appearance model is applied to pre-select face hypotheses and finally to pick out the best hypothesis. However, feature candidate detection is generally computationally expensive if there is significant variation in the scale or rotation of the face. In Wu and Zhou [119], the eye-analogue segments are firstly detected in the face image based on the fact that eye regions are relatively dark. Then a pair of eye-analogue segments are chosen as eye candidates based on the anthropomorphic characteristic of the human eye. Based on each candidate eye pair the face is cropped and verified by measuring its similarity to a few predefined face templates. As mentioned in [43] there are several major advantages of the feature-based approach over the top-down approach. First, local features are considered to be more robust to illumination change than the global appearance. Second, there is no need to discretize the search space as in the scanning windows approach, which might limit the precision of localisation. Other advantages include the simplicity of face and illumination distortion modelling.
5.2.3 Eye Detection/Localisation

Approaches for eye detection/localisation can be broadly divided into three categories [136]: appearance based approaches, template based approaches and feature based approaches. The appearance based approaches detect the eyes based on the intensity appearance of the eyes. Template(model) based approaches rely on abstract eye models, such as deformable models, which can be deformed to fit the local eye image using a cost function [91]. Nixon [79] applied the Hough transform to detect boundary of iris as a circle and the boundary of sclera as an ellipse. In the eye feature based approaches the eye corners are usually detected to locate the eyes. For example, Sirohey and Rosenfeld described two eye detection methods based on eye corners in [102]. The first one is a linear filtering approach based on the Gabor filters to detect the edges of the eye away from the eye corners. The other approach is based on a nonlinear “Wedge” filter employing the color information around eye corners.

The majority of eye detection approaches are based on the eye appearance. Similar to “Eigenfaces”, “Eigeneyes” are used to model the eye appearance by Pentland et al. [84]. Feng and Yuen [33] proposed an eye region detection method based on multi-cues. The first cue is the average low intensity of an image windows encapsulating the human eye, the second cue is based on the edge lines joining the eye centers, and the third cue is the response to a local variance filter. Within each eye window, the authors localise the eye center based on a Variance Projection Function (VPF), which describes the intensity variance along each column and each row of the eye window. Zhou and Geng [135] presented an eye detection method based on their Generalized Projection Function (GPF), which is a linear combination of the VPF and the Integral Projection Function (IPF). The candidate eye windows are provided by the approach proposed in [119] before the projection functions are computed. In [71] Ma et al determine the midline between the left eye and the right eye based on the VPF after the face is coarsely located. The candidates of the left eye and the right eye are then found by applying an eye detector based on a cascaded AdaBoost classifiers which outputs the probability for each hypothesis rather than a binary decision. The candidates are then sampled to form eye pairs. Both local eye images and eye-pair images are
considered in the final prune stage, and the final best eye locations are decided based on the probabilities obtained from the AdaBoost classifiers on the eye-pair images. Everingham and Zisserman [32] investigated regression and classification approaches to eye localization in the face regions found by the Viola and Jones’s face detector. In the regression approach a functional mapping between the input face image and the eye locations are learnt from training image. The two classification approaches considered include a simple Bayesian model of eye/non-eye appearance and a discriminative eye detector based on AdaBoost. The Bayesian model is shown to outperform the other two approaches. In the literature, other classifiers, such as Neural Network and SVM, have also been adopted to distinguish eye/non-eye appearance.

5.3 Active Approaches

In addition to the passive approaches described above, active approaches [25][29][28][74] based on active illumination have been proposed. Dowdall et al. [25] use two bands of reflective infrared illumination for face detection. Skin region is detected based on its different responses to each of the two bands. Their system performs eye detection and eye brow detection based on the integral projection in multi-band images, and if this detection approach fails, the system switches to another detection approach based on a dynamic thresholding model and template matching. Ebisawa and Satoh [29] applied two sets of Near-IR light sources and an image differencing method to detect pupils. The first set of light sources is set coaxial with the camera, therefore pupils appear bright when this light set switched on. The other set of light sources is set to be non-coaxial with the camera. With this set, pupils appear dark while the brightness of the the rest of the scene is similar to that achieved with the first set of light sources. In the difference image of the two images, each illuminated by the respective sets of light sources, the pupils are prominent and they are the only object left after an appropriate thresholding. In [28] a method for eliminating the reflection of light sources in eyeglasses based on morphological operations is presented to improve the performance of the device in [29]. This active Near-IR device is adopted by Morimoto and Flickner [74] for a multiple face detector and by Ji and Yang [53][136] for pupil
Chapter 5. Literature Review on Face Localisation

detection and tracking.

In [130] active Near-IR illumination is used by Zhao and Grigat to introduce bright pupils, and then the top-hat transformation is applied to obtain small regions with an abrupt change in brightness as candidate pupil regions. The top-hat transform is an operation in grayscale Mathematical Morphology defined as the difference of the original image and the image resulting from the application of an grayscale opening operation on the original image. The boundary of each candidate region is then extracted, and Hough transform is applied to detect circles. Those regions without circular shape are removed from the candidate list. In [131] they improved the eye localisation approach. A dynamic thresholding method is applied to the top-hat transformed image to obtain a binary image with the Euler number in a desired range. In this way the number of candidate pupil regions can be controlled. An appearance based model is then applied to verify the pupil candidates. The authors further extend their approach to the scenario with eyeglasses by employing a Generalised Symmetry Transform (GST) on the input face image. Eye candidates locations are supposed to have high value in the transformed image.

5.4 Localisation Performance Measures

It is necessary to have a quantifiable measure of performance of a face localisation algorithm. The performance is usually based on the displacements of localised eye positions from the ground truth eye positions. Jesorsky et al. [52] proposed an error measure which has been the most commonly adopted in the literature. This measure is defined by the maximum of the two displacements normalised by the ground truth inter-ocular distance:

$$d = \max(d_l, d_r) \frac{d_l}{\|C_l - C_r\|} \frac{d_r}{\|C_l - C_r\|} \quad (5.1)$$

where $d_l$ is the displacement between the detected position, $C_l^*$, and ground truth position $C_l$ of the left eye, $d_r$ is the displacement between the detected position, $C_r^*$, and ground truth position $C_r$ of the right eye. The normalisation factor in the denominator of this measure makes it particularly relevant to automatic face recognition, the
performance of which is very sensitive to the localiser performance. This is because geometrically normalisation of faces is usually based on their eye positions.

The performance of the face localiser on a face dataset can be characterized by a cumulative distribution curve $C(freq, d)$ of displacement error. This curve describes what percentage of faces is localised with an error less than any specific threshold $d$.

5.5 Conclusion

A review of automatic face localisation techniques is presented in this chapter. The top-down approaches rely on a face detector exploiting a holistic face model to locate the face region and then perform a localisation refinement in this region. The bottom-up approaches perform facial feature detection first and then verify those detected facial features based on local or holistic face appearance or human face characteristics. The facial features is more robust to occlusion and illumination than global face information, however the feature candidate search across the whole image is time consuming. Active approaches can be applied to enhance the global or local face features. In turn this will facilitate the automatic localisation.
Chapter 6

Automatic Face Localisation for Near-IR Faces

6.1 Introduction

Near-IR face images are similar to face images under visible lighting, because both Near-IR and visible light are in the reflective spectrum. Therefore most face localisation techniques developed for visible light images can be directly applied to Near-IR image face localisation. Near-IR active differential imaging removes the variation of face caused by changes in ambient illumination, and provides face images under consistent frontal active illumination, which facilitates the automatic face localisation. Moreover, when the active illuminant is placed close to camera, bright pupils will be introduced in the face image. These bright pupils are strong and stable features in Near-IR face. Some examples of face images captured using the Near-IR active differential imaging system are given in Figure 6.1.

In this chapter a novel automatic face localisation approach based on bright pupil detection is first proposed. The pupil boundary is detected using a novel edge following technique based on the chaincode. A FloatBoost face localiser based on holistic face appearance is then investigated. Finally a multistage approach combining the above two approaches is proposed, which yields better performance. These three approaches
are described respectively in Sections 6.2, 6.3 and 6.4. In Section 6.5 face localisation experiments are carried out on LED face images based on these three approaches. The face identification and verification tests based on the eye locations provided by the multistage approach are then performed. The results are presented in Section 6.6 and conclusions are drawn in Section 6.7.

6.2 Face Localisation Based on Bright Pupil Detection

It can be observed from Figure 6.1 that bright pupils are strong and stable features in Near-IR faces, which suggests it can be exploited for accurate face localisation. However the localisation of bright pupil is still challenging. Bright pupils are introduced by the illuminant fixed close to the camera in the same way as the red-eye effect in photography. However, the colour clue, which is always used for red-eye detection in digital photo processing, is not available in Near-IR face images. The challenges for pupil detection also come from the variation in intensity of the light reflected from the pupils across the population[76], the variation in the size of bright pupils, and possible occlusions by the upper eyelids.

We developed a bottom-up method of automatic face localisation[138] which exploits the shape property of bright pupil boundary, local eye appearance and global face
appearance. The circular shape of bright pupils, which is a feature invariant to scale and rotation, offers a firm handle for quick pupil candidate detection even when subject to variations in scale or rotation. Therefore, this method actually avoids the expensive computation which is generally required to detect feature candidates under changes in scale or rotation. We proposed a novel method to detect the circular boundary of bright pupil based on edge following. This rule-based edge following is robust to the change in the size and direction of the boundary. Accuracy in localisation is further improved by learning, based on both local facial features and global face appearance. The diagram of the approach is given in Figure 6.2.

6.2.1 Pupil Boundary Detection

Two significant features of a bright pupil are exploited here for pupil detection. First, the bright pupil has a circular boundary. As we know, the circle retains its shape under scaling, rotation, and translation. Therefore, a pupil detection approach based on its circular shape feature can work regardless of scaling, rotation or translation. Unlike the circular shape, the local appearance of a pupil is not invariant to the transforms considered (see Figure 6.4(g) and Figure 6.5(d)). Therefore, multi-scale, multi-directional templates/filters have to be applied to facilitate the detection based on local appearance, which requires expensive computation. Second, the brightness of the pixels inside the pupil is much higher than that of pixels surrounding it. This big difference in brightness is manifest in a stable and continuous edge around pupil regions. In contrast, the edges of eyelids or irises are often discontinuous and thus difficult to track. This motivates us to perform pupil detection on the edge image. Another reason for detecting the pupils in edge image is that edges are more robust to illumination variation than pixel intensities.

The Hough transform is often used to detect circles[79][130]. However, most pupils in our images are too small in size (eg. with 2-3 pixels in radius) for the Hough transform. What is more, due to the possible occlusion by the upper eyelid, pupils are often found with only lower part in the image (see Figure 6.4(j),6.4(k),6.4(l)). Therefore we choose to detect only the lower part of the pupil boundary based on a novel method employing
Figure 6.2: Diagram of the face localisation method based on pupil edge detection.
edge following. A Canny edge detector is applied for edge detection, then a rule-based edge following algorithm is performed to find the possible edge segments of the lower part of the pupil boundary. Based on the rules, the edge segments consistent with the lower half of a circle will be accepted as candidate segments for the pupil boundary. The direction chain code [108] from an edge point to its successor in its 8-neighborhood is defined in Figure 6.3. The rules governing the search of valid pupil candidate segments include:

1. The starting point of the candidate segment has a successor in the direction with chain code either 0 or 1, and the chain code from the predecessor of the ending point to the ending point is 3 or 4.

2. When following a candidate segment the chain code remains the same or increasing.

3. A candidate segment contains sub-segments with at least two different chain codes.

4. The length of any sub-segment in which chain code remains the same is within a certain interval.

5. The length of a candidate segment is within a certain range.

6. Following the edge segment anti-clockwise, the neighbor pixels on the left are brighter than those on the right.

The candidate segment search based on these rules is tolerant to the changes in scale. Examples of original images and edge images for faces, eye regions, and detected pupil edge candidates are shown in Figure 6.4 and Figure 6.5.

6.2.2 Candidates Sorting

For each detected candidate pupil boundary, the candidate pupil centre is determined as the center of the circle containing this pupil boundary. The candidate pupil centres are pruned based on the co-existence of pupil pairs. For every candidate of one eye
Figure 6.3: (a) direction chain code definition. (b)(c) valid segments with chain code string 00122334 and 11122334.

Figure 6.4: Face images (a)(d), edge images (b)(e), detected candidate pupil edge segments (c)(f), and zoomed images (g)-(l) of eye regions for the corresponding images (a)-(f).
6.2. Face Localisation Based on Bright Pupil Detection

Figure 6.5: Face image(a), edge image(b), detected candidate pupil edge segments(c) after scaling(X1.5), rotation(25°) and translation have been performed on the face image in Figure 6.4(a). (d-f) are zoomed images of the eye regions for the corresponding images(a-c). True pupil edges are still successfully detected.

there should exist a candidate for the other eye so that they can form an eye pair with an appropriate inter-ocular distance. Meanwhile, the locations of each candidate pair must ensure that the cropped face region based on this pair falls completely within the image region. The positions in the neighborhood of the surviving pupil candidates are considered to be the coordinates of the candidate eye centers. They are sorted according to the distance to the eye templates. Rather than using one common template for both eyes, two templates, one specifically for each eye, are learnt from the training faces. Treating the left eye and the right eye separately increases efficiency. Adopting this strategy, the candidates for each eye will come from two different but shorter candidate lists respectively instead of both from a longer list. In consequence the number of candidate pairs will be much smaller. Each template is a histogram equalised version of the mean image of the rectangular image patches extracted from the training images. The patches are positioned at the corresponding manually marked eye centres. The image patch centered at each candidate is first histogram equalised, and then the
Chapter 6. Automatic Face Localisation for Near-IR Faces

normalized correlation score between the corresponding template and this histogram equalised image patch is computed. The score distributions of true candidates and false candidates are obtained from the training set. For each eye, the candidates with scores above a threshold obtained from the above distributions are sorted by the score value to form a list. This sorting of candidates ensures that true eye positions will be more likely to be considered first in the following stage of the analysis.

6.2.3 Appearance Based Validation by SVM Classifiers

By minimizing the structural risk rather than minimizing the empirical error from the training data, SVM has been considered as a classifier with good generalisation ability (see Section 2.3.2). We employ SVMs based on the local appearance of eye region and the global appearance of face region to validate candidate eye positions. For the same reason described in the last section, we use two SVMs, one for each eye. Since the eyebrow, mouth and eye regions are sometimes similar in appearance, there can be false candidates which pass the validation of the eye SVM. Therefore, another SVM validation based on the global face appearance is used as a final step to validate a pair of eye candidates.

The positive samples used to train each eye SVM come from image patches centred at the corresponding ground truth eye centres. The size of the image patch is 15*25 pixels. The negative samples are chosen from image patches centred at false candidates with the highest normalized correlation score to the template. Examples of the training samples for the eye SVMs are given in Figure 6.6. For the training of the face SVM, the positive samples are normalized faces of size 22*20 pixels, and the negative samples come from misaligned face regions normalized according to the false candidate positions. Examples of the training samples for the face SVMs are given in Figure 6.7.

6.3 Face Localisation Based on FloatBoost

Li et al. [67] proposed a novel boosted classifier learning procedure named FloatBoost. In contrast to AdaBoost [35], which minimizes an upper error bound expressed as
Figure 6.6: Positive training samples (a)(c) and negative samples (b)(d) for left eye SVM and right eye SVM, respectively.

Figure 6.7: Positive training samples (a) and negative samples (b) for face SVM.
an exponential function of the margin on the training set, FloatBoost incorporates a feature selection mechanism into AdaBoost aiming at achieving minimum error rate with fewer weak classifiers. After each iteration of AdaBoost, a backtrack mechanism is applied to select the combination of weak classifiers with the best performance. In their paper, FloatBoost learning is applied to face detection. Simple Harr wavelet like features are adopted, which are similar to those used by Viola and Jones [114] except that displacements among subregions of Harr wavelet window are allowed.

We employed the face detector based on the FloatBoost classifier mentioned above to find a coarsely located face region. The eyes candidates are then detected using another FloatBoost classifier in the upper left and upper right face regions, respectively. In each region, a candidate position is selected as the final output if the image patch centred at it best matches the corresponding eye template.

6.4 Multistage Approach

The face localiser based on bright pupil detection proposed in Section 6.2 completely relies on the successful detection of bright pupils. There are several possible reasons which can lead to the failure of this localiser. Sometimes the pupils in a face image will be missing because the subject does not open his/her eyes widely enough. Additionally, there is a possibility that the SVM classifiers occasionally fail. In this situation we apply the FloatBoost face localisation method described in Section 6.3 as an alternative solution. The adopted FloatBoost face localiser is based on the global face appearance and is insensitive to the missing bright pupils. The diagram of the multistage approach combining the pupil based localiser and the FloatBoost based localiser is given in Figure 6.8.

6.5 Face Localisation Experiments

Face localisation experiments were conducted to evaluate the three approaches mentioned in Section 6.4: namely bright pupil based localiser, FloatBoost localiser and the
6.5. Face Localisation Experiments

Figure 6.8: Diagram of the multistage face localisation method.

multistage localiser. The classifiers implementing the above approaches were trained on data recorded in one session and tested on data acquired in the other session.

We adopted the localisation error measure proposed by Jesorsky et al. [52] as described in Section 5.4. The cumulative distributions of the displacement error using the above two methods and the multistage approach are shown in Figure 6.9. According to [43], the empirical threshold of localisation accuracy to ensure the success in the subsequent face recognition stage is 0.05. Therefore 0.05 is chosen here as the threshold to distinguish a successful localisation. The success rate achieved by the first method is 96.5%, which is 6% higher than using the method based on FloatBoost. A further improvement of 1% is achieved by the multistage approach. Note the success rate exceeds 99% when the threshold $d_{eye}=0.07$. Examples of the localisation results for images in Figure 6.1 are shown in Figure 6.10.

A significant contribution to the 3.5% localisation error incurred by the first method is due to the missing bright pupils for the subject in Figure 6.11. The pupils of the right eye are missing in nearly all 48 images of this subject. This contributes 2.5% error. There are several reasons for the failure to obtain bright pupil. First, his right eye was almost never opened widely enough, so the pupil region is occluded by the eyelid. Second, there are some imperfections in our capture system: the LED illuminant is placed on the top of the camera, and the camera has a very narrow field-of-view. Hence the subject has to sit far away from the camera to ensure the whole face region is captured. Consequently, the bright pupil region is relatively small. It can be expected that in a better designed system with LED rings surrounding the camera, a larger bright pupil region will be obtained, and it will be less likely to be missing. Except
missing bright pupil, another cause which can lead to error is a failure of the eye or face SVM classifier systems. In these situations, in the multistage approach, the FloatBoost based face localiser is applied as a remedy.

6.6 Face Recognition on Automatically Localised Data

6.6.1 Face Identification Experiment

It can be argued that the displacement measure adopted is unable to provide the information required to distinguish errors in translation, rotation and scale [90]. Moreover, this measure can be influenced by the uncertainty in marking up the ground truth eye positions which is caused by the ambiguity in the definition of precise eye centre position, especially for small faces. Therefore, it is meaningful to measure the performance of localisation in terms of recognition error.

We conducted two sets of face identification experiments, one on the automatically
6.6. Face Recognition on Automatically Localised Data

Figure 6.10: Localisation results.

Figure 6.11: Images with false detection using the first approach.
Chapter 6. Automatic Face Localisation for Near-IR Faces

localised faces (cross Auto/Auto faces), which are registered using the proposed multi-stage localisation approach, and the other set on cross Manu/Auto faces, by which we mean manually registered faces used for training and automatically localised faces for testing. The three test protocols are similar to those in Section 4.6.1. The representation we used is Fisherface, which has been proved to be the best choice for LED faces in Section 4.6. The fisher subspace is built in the same way as described in Section 4.6. The SVM and Nearest Neighbor classifiers are applied as in the other experiments. Several observations can be made from the results shown in Table 6.1 and Table 6.2:

1. Very low error rates are achieved for all the tests on automatically localised faces. Nearly all the error rates on the Cross Session and the Cross Illumination tests are below 0.8%, whether trained on manually registered data or automatically registered data. Actually these error rates are just slightly inferior to the best results obtained on the manually registered faces (shown in Tables 4.2, 4.3 and 4.4), which confirms once again that the proposed multistage approach provides accurate face localisation.

2. In the Combined test, which is the most difficult test, the Nearest Neighbor classifier gives significantly better performance than SVM.

3. A practical application scenario is best represented by the Combined test on across Auto/Auto faces, yielding an surprisingly low error rate of less than 0.8%. This error rate is better than the result obtained using manually registered training images.

6.6.2 Face Verification Experiment

The face verification experiments are conducted on across Manu/Auto faces and Auto/Auto faces based on the three protocols defined in Section 4.6.2. The Fisherface representation is used here with normalised correlation as a similarity measure. From the ERRs reported in Table 6.3 and Table 6.4, it can be observed that:

1. Compared to the results obtained on manually registered data shown in Tables 4.7, 4.8 and 4.9), the results on automatically localised faces are just a little worse.
Table 6.1: Average error of identification tests on the automatically localised faces (in %)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>LED Faces</th>
<th>Test Protocol</th>
<th>Training Set</th>
<th>Test Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.68</td>
<td>$AS_{jo}$</td>
<td>$AS_{j\neq jo}$</td>
<td>Cross Session</td>
<td></td>
</tr>
<tr>
<td>N.N.</td>
<td>0.52</td>
<td>$AC_{i0}$</td>
<td>$AC_{i\neq i0}$</td>
<td>Cross Illum.</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.52</td>
<td>$AC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>N.N.</td>
<td>0.58</td>
<td>$AC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Average error of identification tests on the Manually/Automatically localised faces (in %)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>LED Faces</th>
<th>Test Protocol</th>
<th>Training Set</th>
<th>Test Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.73</td>
<td>$MS_{jo}$</td>
<td>$AS_{j\neq jo}$</td>
<td>Cross Session</td>
<td></td>
</tr>
<tr>
<td>N.N.</td>
<td>0.68</td>
<td>$MC_{i0}$</td>
<td>$AC_{i\neq i0}$</td>
<td>Cross Illum.</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.61</td>
<td>$MC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>N.N.</td>
<td>0.64</td>
<td>$MC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2.55</td>
<td>$MC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>N.N.</td>
<td>1.28</td>
<td>$MC_{i0}S_{jo}$</td>
<td>$AC_{i\neq i0}S_{j\neq jo}$</td>
<td>Combined</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.3: EER for verification tests on Manually/Automatically localised faces (in %)

<table>
<thead>
<tr>
<th>Test Protocol</th>
<th>Cross Session test</th>
<th>Cross Illumination test</th>
<th>Combined test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR</td>
<td>1.98</td>
<td>1.08</td>
<td>2.25</td>
</tr>
</tbody>
</table>

Table 6.4: EER for verification tests on automatically localised faces (in %)

<table>
<thead>
<tr>
<th>Test Protocol</th>
<th>Cross Session test</th>
<th>Cross Illumination test</th>
<th>Combined test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR</td>
<td>0.96</td>
<td>0.77</td>
<td>1.37</td>
</tr>
</tbody>
</table>

2. In the most practical scenario, which is best represented by the *Combined* test on cross Auto/Auto faces, an EER, as low as 1.37%, is achieved.

3. Interestingly the results across Auto/Auto faces are better than across Manu/Auto faces, just as in the identification experiments. This is because our localisation approach provides consistent automatic face localisation result across the training and test images of each subject. Although some localisation results sometimes deviate from ground truth position, the consistency of the localisation errors across training and operational phase of the system guarantees good performance.

6.7 Conclusion

We addressed the automatic face localisation problem for faces captured by a Near-IR active differential imaging system. A localisation approach based on bright pupil detection has been proposed. It includes a novel edge following approach to detect the pupil boundary. A multistage approach which combines this pupil detector and a FloatBoost face detector is shown to yield the best localisation performance. The low error rates in face recognition on faces localised by the proposed multistage approach confirm the excellent performance of our automatic face recognition system in a practical environment with varying illumination.
Chapter 7

Motion Analysis for Face Recognition System based on Active Differential Imaging

7.1 Introduction

In the previous chapters the assumption has been made that there is no motion taking place during the capture of the Combined frame and Ambient frame. However, the active differential imaging system is detrimentally affected by any motion of the subject in the field of view. A moving subject will appear at two different locations in the combined illumination frame (C-frame) and ambient illumination frame (A-frame), respectively, which results in a difference image with “artifacts” as shown in Figure 7.2. In comparison, Figure 7.1 shows the difference image without motion artifacts.

To the best of our knowledge, this problem has not been addressed before. We propose a motion compensation technique developed to cope with this problem[140]. The face capture system captures a C-frame and an A-frame alternately. The motion between two C-frames is estimated, and a “virtual” C-frame is computed by performing motion interpolation. The difference image between this “virtual” C-frame, and the A-frame captured between the above two C-frames is used for recognition. The advantage
Chapter 7. Motion Analysis for Face Recognition System based on Active Differential Imaging

Figure 7.1: Face images under combined illumination (a), ambient illumination (b) and their difference image (c), without any motion.

Figure 7.2: Face images under combined illumination (a), ambient illumination (b) and their difference image (c), when the face is moving.
of applying the proposed approach is proved by the improvement in face recognition results on a database containing moving face sequences.

In Section 7.2, the motion problem is investigated and a case study is presented. The proposed motion compensation approach is presented in Section 7.3. A moving face database is captured as described in Section 7.4. The face recognition experiments carried out on face images with and without the proposed motion compensation approach are described in Section 7.5, with conclusions drawn in Section 7.5.4.

7.2 Motion Problem for Active Differential Imaging

Two problems can be introduced by a motion for a capture system based on active differential imaging. The first one is motion blur for each captured image. However, this is a general problem for all imaging systems. Fortunately face recognition based on the commonly adopted subspace approaches is relatively robust to the degradation in resolution and consequently it is not seriously affected by motion blur. Therefore motion blur is not a major problem to be addressed. The second one is posed by the motion artifacts in the difference image. Since the C-frame and A-frame are captured at different times, the difference image will be quite different to what it is supposed to be due to the position change of the object. The displacement between object positions in two frames is directly related to the time interval between these two successive captures and the moving speed. As shown in Figure 7.2, this problem is significant for a face capture system based on differential imaging. Artifacts are prominent especially around the edges, such as eyelids, nose and mouth. Therefore this thesis focuses on the second problem.

7.2.1 A Case Study

An example is given below to show how the face similarity degrades due to displacement of the faces. A C-frame and an A-frame are taken for a still face with 60 pixels in interocular distance as shown in Figure 7.3. We manually shifted A-frame by 1 to 10 pixels from its original position horizontally, or vertically. The faces in the resulting difference
Chapter 7. Motion Analysis for Face Recognition System based on Active Differential Imaging

Figure 7.3: Combined illumination frame (a) and ambient illumination frame (b)

Figure 7.4: original face template (a), and resulting difference image when A-frame was shifted (-10, -7, -3, 0, 3, 7, 10) pixels horizontally (a) and vertically (b) from C-frame images are registered with reference to the original position in the C-frame, cropped and normalised to 55*50 patches, and histogram equalised, as shown in Figure 7.4. Figure 7.5 shows that the similarities between the resulting difference images and the face template keep decreasing when the displacement between the C-frame and A-frame in any direction (left, right, up or down) increases. The similarity is measured by the Normalised Correlation (NC) score in the original image space, Principal Component Analysis (PCA) subspace, and Linear Discriminant Analysis (LDA) subspace. PCA and LDA subspaces are built from a Near-Infrared face database with 2844 images of 237 people. NC in an LDA subspace usually gives much better results for face recognition than NC in the image space and PCA subspace, however, NC in the LDA subspace is much more sensitive to the displacement than in the other two spaces, according to Figure 7.5.

As a result, the performance of a face recognition system based on active differential imaging will degrade when faces are moving. For a general-purpose camera, the time interval between two frames is 40ms (CCIR) or 33ms (EIA), which is so long that the motion effect maybe significant. A hardware solution is specific high speed sensors, such as the sensor developed by Ni et. al [78], which can provide a capture speed of
7.2. Motion Problem for Active Differential Imaging

Figure 7.5: NC score drops when A-frame shifts from C-frame horizontally(a) and vertically(b).
100 images per second. However, due to the high price of custom designed devices, a software solution is always desirable.

7.3 Frame Interpolation based on Motion

7.3.1 The Proposed Approach

The problem cannot be solved using only the difference image. First, simply applying subspace approaches does not work: as discussed above, none of the commonly used face subspace representations is both insensitive to this motion effect and discriminative enough for face recognition. Second, motion information cannot be recovered from the difference image to remove the motion effect. It is also difficult to align the faces in the successive C-frame and A-frame in that faces are in different illuminations in these two frames. We propose to use two nearest C-frames to obtain motion information and interpolate a virtual C-frame, the motion effect can then be removed from the difference image of the A-frame and the virtual C-frame. Assuming the face is moving
at the same speed in the same direction between \( t_i \), the time when the first \( C \)-frame \( C_i \) is captured, and \( t_{i+2} \), the time when the second \( C \)-frame \( C_{i+2} \) is captured. We can apply interpolation to obtain a virtual \( C \)-frame \( C'_{i+1} \) as “captured” at \( t_{i+1} \), which is at the same time when the frame \( A_{i+1} \) is captured. The faces in the frame \( C'_{i+1} \) and frame \( A_{i+1} \) are therefore at exactly the same location and the motion effect is removed. This approach is illustrated in Figure 7.6.

### 7.3.2 Motion Estimation Based on Robust Optical Flow

The first step of the proposed approach is to estimate the motion accurately. Techniques for motion estimation can be roughly divided into two major classes: differential techniques and feature-based techniques [111]. Those in the first class provide dense motion field, while those in the latter provide only sparse motion field. An overview for motion estimation methods can be found in [109]. Dense motion field, which is needed for further frame interpolation, can be approximated by the optical flow. By assuming the apparent brightness of a object remains constant when it is moving, the image brightness constancy equation

\[
\frac{dI(x(t), y(t), t)}{dt} = \frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t} = 0
\]  

(7.1)

provides the relationship between the spatial and temporal change of image brightness \( I \) and motion field \( \mathbf{v} = (u, v) \). However, from this equation only the component of motion field in the direction of spatial image gradient can be determined, which is known as the Aperture Problem [111]. This problem can be tackled by assuming the constancy of the motion field within a small image patch. Then the equation 7.1 for all positions in this patch provided a over-constrained system for determine the least squared solution for the motion vector[70]. An alternate way is to incorporate spatial smoothness constraint (also called spatial coherence constraint), by which the optical flow within a neighborhood is assumed to change gradually due to a single motion, and minimizing

\[
\int \left[ (\frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t})^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2) \right] \]

[49]. In some approaches the optical flow within region is modelled by affine or quadratic model, and the optical flow estimation is solved by searching the best model parameters
Robust estimation approaches are proposed to overcome the problem caused by violation of image brightness constancy or spatial smoothness, which always happens when there is multiple motion, specular reflection, shadows, depth discontinuity or illumination change. The robust estimation method by Black and Anandan [12] is adopted in this thesis. In [12] Black and Anandan apply a robust framework to standard formulations of optical flow problem to reduce their sensitivity to violations of their underlying assumptions.

Robust formulation for optical flow estimation

For the image intensity function \( I(x, y, t) \) defined on the region \( S \), the estimation of optical flow field \( U = \{(u_s, v_s) \mid s \in S\} \) can be treated as a minimization problem of a cost function \( E(u, v) \) of the residual errors from the data conservation constraint and the spatial coherence constraint:

\[
E(u, v) = \sum_{s \in S} \left[ \lambda_D \rho_D(I_x u_s + I_y v_s + I_t) + \lambda_S \left[ \sum_{n \in G_s} \rho_S(u_s - u_n) + \sum_{n \in G_s} \rho_S(v_s - v_n) \right] \right]
\]

where \( I_x, I_y \) and \( I_t \) are the partial derivatives of \( I(x, y, t) \), \( \lambda_D \) and \( \lambda_S \) are weights for the data conservation term and spatial coherence term, respectively. Generally, the \( \rho \) functions \( \rho_D \) and \( \rho_S \) are both of the quadratic form:

\[
\rho(x) = \frac{x^2}{2\sigma^2}
\]

then the solution to the problem of minimising \( E(u, v) \) involves a standard least-squares estimation. However, this estimation is optimal only if the measurement errors are normally distributed. The problem with the least square estimation is that the outliers in the set of measurements are assigned a high weight by the quadratic function. This can be seen easily from the influence function corresponding to the \( \rho \) error function, which is proportional to the derivative of the \( \rho \) function and characterizes the bias that the measurement imposes on the solution. In the case of the least-squares estimation, the influence of each measurement is proportional to \( x/\sigma \) and increases linearly and without bound. See Figure 7.7 (a)(b). The robust formulation of \( E(u, v) \) presented in
7.3. Frame Interpolation based on Motion

Figure 7.7: The $\rho$ function and its influence function $\psi$. (a)(b) Quadratic $\rho$ function and its influence function. (c)(d) Lorentzian function and its influence function (reprinted from [87])

[12] is as below:

$$E(u,v) = \sum_{s \in S} \left[ \lambda_D \rho_D(I_xu_s + I_yv_s + I_t, \sigma_D) + \lambda_S \left( \sum_{n \in G_s} \rho_S(u_n - u, \sigma_S) + \sum_{n \in G_s} \rho_S(v_n - v, \sigma_S) \right) \right]$$

(7.4)

where $\rho$ is a Lorentzian function:

$$\rho(x, \sigma) = \log(1 + \frac{1}{2} \left( \frac{x}{\sigma} \right)^2)$$

(7.5)

and $\sigma_D, \sigma_S$ are parameters that control the shapes of the Lorentzian functions. As shown in Figure 7.7(c)(d), the derivative function, $\psi$, of the Lorentzian function, $\psi(x) = \frac{2x}{\sigma^2 + x^2}$, is redescending and thus this $\rho$ function has the ability to reduce the effect of outliers. In our implementation, $\lambda_D$ is set to 10.0, and $\lambda_S$ is set to 1.0. The initial value of $\sigma_D$ is 20.0 and that of $\sigma_S$ is 1.0.

Coarse-to-fine strategy for large motion

To cope with large motions, a coarse-to-fine strategy is employed. A 4-level pyramid of spatial filtered and sub-sampled images is constructed. Both horizontal and vertical dimensions are sub-sampled to half size during the construction of the pyramid from
one level to its upper level. For each level of resolution 20 iterations are carried out, and during each iteration the optical flow at each site is updated. When the optimisation of the motion estimation at a level is finished the optical flow is projected to the next level in the pyramid and the first image is warped using this flow field so that:

\[
I_{\text{warped}}(x, y) = I(x - u(x, y), y - v(x, y))
\]

The image derivatives \((I_x, I_y)\) are then computed at this level using the warped image as opposed to the original image. After processing at the finest level, the value of \(\sigma_D\) and \(\sigma_S\) are lowered according to \(\sigma_{n+1} = 0.95\sigma_n\), and the entire coarse-to-fine process is repeated.

### 7.3.3 Frame Interpolation

If \(u\) and \(v\) represent the horizontal and vertical motion between frame \(C_i\) and \(C_{i+2}\), then the motion between \(C_i\) and \(C_{i+1}\) is defined by \(u \frac{t_{i+1}-t_i}{t_{i+2}-t_i}\) and \(v \frac{t_{i+1}-t_i}{t_{i+2}-t_i}\) based on linear interpolation. \(C'_{i+1}\) can be warped from \(C_i\) based on:

\[
C'_{i+1}(p) = \sum_{s=0}^{3} d^s C_i(p^s_0)
\] (7.6)

where \(C'_{i+1}(p)\) is the grey value of a pixel \(p\) with coordinates \((m, n)\) in frame \(C'_{i+1}\), \(\{p^s_0\}_{s=0,\ldots,3}\) are the 4 nearest neighbors of the original subpixel location \((m_0, n_0)\) of \(p\) in the frame \(C_i\) with \(m_0 = m - u(m, n) \times \frac{t_{i+1}-t_i}{t_{i+2}-t_i}\), \(n_0 = n - v(m, n) \times \frac{t_{i+1}-t_i}{t_{i+2}-t_i}\). \(\{d^s\}_{s=0,\ldots,3}\) are the weights related to the distances between \((m_0, n_0)\) and \(\{p^s_0\}_{s=0,\ldots,3}\). Figure 7.8 illustrates the interpolation process.

### 7.4 Capture System and Database Capture

An experimental face capture system based on active differential imaging was used to capture moving face data for our experiments. The image sensor has a high resolution of 755*400 pixels. A face database of 37 subjects was captured in an indoor environment near the window with sunlight coming in. Each subject sat 1 meter away from the camera, and was asked to keep moving his/her face. Two sessions of data were recorded.
7.5. Experiments

Figure 7.8: Illustration of the interpolation process.

for each subject, with 29 C-frames and 29 A-frames captured continuously for each session. Ambient illumination was mainly from left for one session and from right for the other. Another 6 difference images for each subject were captured when the subject sat still. These images will serve as gallery images for the identification and verification experiments.

7.5 Experiments

To show the advantage brought by the proposed motion compensation approach, face recognition experiments are conducted on two sets of difference face images for comparison. The first set contains the original difference face images between every pair of C-frame and A-frame, without considering the motion issue. For the second set, the proposed approach is applied to obtain the difference images without motion effect. Faces in both sets are geometrically normalised based on the manually marked eye positions in the corresponding C-frame, cropped to 55*50 image patches, and photometrically normalised using histogram equalisation. Examples for faces in both sets are shown in Figure 7.9. Each set contains 2072 faces (37 subjects\times2 sessions\times28 difference images).
7.5.1 Improvement in Face Similarity to Templates

A histogram of the motion between every two $C$-frames for the whole database is shown in Figure 7.10.(a). For every pair of $C$-frames, the motion value recorded in the histogram is the average length of all those motion vectors of the motion foreground pixels (pixels with motion vector magnitude above a threshold $0.8$). Since faces are normalised with reference to the eye positions, the same absolute displacement will have different influence for faces with different inter-ocular distances. Therefore relative motion, which is the motion value over the inter-ocular distance, is applied here to measure motion.
Figure 7.10: (a) Histogram of the relative motion in all sequences. (b) Average improvement in NC score after using motion compensation in terms of the relative motion.
Table 7.1: Error Rates of Face Identification Experiments on Moving Faces (in %)

<table>
<thead>
<tr>
<th></th>
<th>with motion compensation</th>
<th>without motion compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>full face</td>
<td>12.07</td>
<td>31.71</td>
</tr>
<tr>
<td>left face</td>
<td>14.23</td>
<td>32.09</td>
</tr>
<tr>
<td>right face</td>
<td>10.28</td>
<td>26.93</td>
</tr>
<tr>
<td>fusion of left/right faces</td>
<td>6.13</td>
<td>14.96</td>
</tr>
</tbody>
</table>

According to Figure 7.10.(b), NC score in the LDA subspace is improved by motion compensation more significantly than in the PCA subspace and the original image space. When motion is too small, applying motion interpolation brings a little improvement. But for moderate motion, the improvement is significant. When motion is too large the improvement decreases because large motion tends to be associated with a large pose change has a negative influence on motion estimation.

7.5.2 Face Identification Experiment

Face identification experiments are carried out using the Nearest Neighbor classifier based on full face, left/right half of the face, and the fusion of left/right halves of the face. The similarity is measured by NC score in the respective LDA subspaces for full face, left face, or right face. For the fusion case, the Sum rule [59] is applied to fuse the similarity score in the LDA subspace for left face and right face. As shown in Table 7.1, after motion interpolation the errors decrease to less than half of the errors achieved without performing motion compensation. Applying fusion technique gives the best identification result with an error rate 6.13%.

7.5.3 Face Verification Experiment

In the first round of verification experiments, the data of one session is used for evaluation, and that of the other session for testing. Then evaluation and test sessions are switched for the other round of testing. Every test image is used to make claim
Table 7.2: EERs of Verification Experiments on Moving Faces (in %)

<table>
<thead>
<tr>
<th></th>
<th>with motion compensation</th>
<th>without motion compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>full face</td>
<td>7.88</td>
<td>14.13</td>
</tr>
<tr>
<td>left face</td>
<td>8.69</td>
<td>15.63</td>
</tr>
<tr>
<td>right face</td>
<td>8.72</td>
<td>15.33</td>
</tr>
<tr>
<td>fusion of left/right</td>
<td>6.42</td>
<td>10.95</td>
</tr>
</tbody>
</table>

to all 37 identities in the gallery. So in total there are 1036 instances of true claims and 1036*36 instances of false claims during the evaluation and the same number of true claims and false claims in the testing stage. The average verification errors are reported in Table 7.2. Again, significant improvement is achieved by applying motion interpolation. Fusion has the best performance with the lowest error rate 6.42%.

7.5.4 Conclusion

Motion causes problems for face recognition systems based on active differential imaging. The similarities between probe images and the template decrease due to the artifact in the captured face image caused by motion. In this chapter we proposed an approach based on motion compensation to remove the motion effect in the difference face images. A robust error kernel defined by the Lorentzian function was deployed to minimise the effect of outliers on the estimation accuracy. A significant improvement has been achieved in the results of face identification and verification experiments on moving faces.
Chapter 7. Motion Analysis for Face Recognition System based on Active Differential Imaging
Chapter 8

Conclusion and Future Work

8.1 Conclusion

Biometrics based identity authentication has become widely accepted and has been receiving increasing attention. Face is one of the most popular biometric traits and it has several advantages over the other biometrics. The face recognition in controlled environment has been quite successful; however, it is still challenging to overcome the problem caused by uncontrolled imaging conditions. Illumination variation is one of the factors causing serious problem for face recognition.

An extensive literature survey of conventional approaches to illumination invariant face recognition has been present in this thesis. The approaches fall into two categories: passive approaches and active approaches. In the first category, the visible spectrum images which is sensitive to illumination variations were studied. It was noted that the illumination problem had been addressed by either modelling the behavior of face appearance change physically or statistically, or by removing illumination variation by photometric normalisation, or by extracting illumination insensitive features from the image. In the other category, active sensing techniques were involved to obtain illumination invariant modalities of face, such as 3D information and multi-spectral images. Despite the good performance claimed in the papers, the approaches in each group have their own disadvantages.
An active approach to address the illumination variation problem based on the active Near-Infrared differential imaging technique has been investigated in this thesis. Assuming a still scene, a linear response of sensor to the scene radiance and no saturation, a face image independent of ambient illumination can be obtained by this technique. This has been proved theoretically by means of the image formation modelling based on radiometry analysis and sensor modelling, and empirically by experiments on a face database captured under varying ambient illumination. In three scenarios: across different time, across different illumination, and across both different time and illumination, the performance of the face recognition system based on active Near-Infrared differential imaging has been compared with those achieved by several representative illumination invariant approaches. These include the advanced photometric normalisation approach based on anisotropic smoothing, and two based on illumination insensitive features: Phase and FisherPhase. The active differential imaging technique achieved excellent results in all scenarios and it has been shown to be superior to the other approaches.

The problem of automatic face localisation in Near-IR face images has been studied in this thesis. A multistage approach has been proposed. It is a combination of a novel pupil detection approach based on edge following and chaincode representations, and a FloatBoost based face localisation approach. Very good localisation performance has been achieved by the proposed approach, and this achievement led to the excellent performance of the fully automatic face recognition system developed in the thesis.

The motion problem in the active differential imaging system has been addressed in this thesis. The moving subject appears at two different locations in the Combined frame and Ambient frame, and therefore artifacts are introduced to the difference image. A motion compensation approach has been proposed to overcome this problem. The motion between two Combined frames is estimated to determine a virtual Combined frame by interpolation. This virtual frame is "captured" at the same time instance as the Ambient frame. In this way the difference image of the virtual Combined frame and the ambient frame is free of motion artifact. The performance of the face recognition system degraded by motion has been significantly improved by the proposed motion compensation approach.
8.2 Future Work

In the context of automatic face recognition based on active Near-IR differential image there are several areas which have not been investigated in this thesis.

First of all, the eye glasses problem has not been addressed in the thesis. The reflection from eye glasses due to the active illumination will change local appearance in eye region. Possible solutions might include adding eye/face images with eyeglasses to the training samples for building the eye/face detector, or detecting and removing the eye glasses from the images using the approaches available in [54] and [118].

Moreover, the active differential imaging system works under the assumption of no saturation. However, when the ambient illumination is too strong, the dynamic range of the active illumination will be limited. It is interesting to investigate how the system performance varies with the dynamic range of the active Near-Infrared illumination.

Finally, although the motion compensation approach successfully removes the motion artifact, the details in the difference face images are not always well preserved. It would be useful to develop an appropriate automatic approach to localise the face accurately in the degraded difference image.
Bibliography


