

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0020

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité “SIGNAL et IMAGES”

présentée et soutenue publiquement par

Rui MIN

12 Avril 2013

Reconnaissance de Visage Robuste aux Occultations

Directeur de thèse: **Jean-Luc DUGELAY**

Jury

Mme. Alice CAPLIER, Professeur, GIPSA-lab, Grenoble-INP	Président
M. Liming CHEN, Professeur, École Centrale de Lyon	Rapporteur
M. Abdenour HADID, Professeur Adjoint, University of Oulu	Rapporteur
M. Philippe ROBIN, Directeur Technique, Thales	Examinateur
M. Jean-Luc DUGELAY, Professeur, EURECOM	Directeur de thèse

TELECOM ParisTech
école de l’Institut Télécom - membre de ParisTech

THÈSE

The research presented in this thesis was supported under the French national project FR OSEO BIORAFALE.

Face Recognition Robust to Occlusions

Dedicated to my PhD study from 2010.3 to 2013.3

Abstract

Abstract

Face recognition is an important technology in computer vision, which often acts as an essential component in biometrics systems, HCI systems, access control systems, multimedia indexing applications, etc. In recent years, identification of subjects in non-controlled scenarios has received large amount of attentions from the biometrics research community. The deployment of real-time and robust face recognition systems can significantly reinforce the safety and security in public places or/and private residences. However, variations due to expressions/illuminations/poses/occlusions can significantly deteriorate the performance of face recognition systems in non-controlled environments. Partial occlusion, which significantly changes the appearance of part of a face, cannot only cause large performance deterioration of face recognition, but also can cause severe security issues.

In this thesis, we focus on the occlusion problem in automatic face recognition in non-controlled environments. Toward this goal, we propose a framework that consists of applying explicit occlusion analysis and processing to improve face recognition under different occlusion conditions. We demonstrate in this thesis that the proposed framework is more efficient than the methods based on non-explicit occlusion treatments from the literature. We identify two new types of facial occlusions, namely the sparse occlusion and dynamic occlusion. Solutions are presented to handle the identified occlusion problems in more advanced surveillance context. Recently, the emerging Kinect sensor has been successfully applied in many computer vision fields. We introduce this new sensor in the context of face recognition, particularly in presence of occlusions, and demonstrate its efficiency compared with traditional 2D cameras. Finally, we propose two approaches based on 2D and 3D to improve the baseline face recognition techniques. Improving the baseline methods can also have the positive impact on the recognition results when partial occlusion occurs.

In this thesis, occlusion-robust face recognitions are proposed for different scenarios, so that identification of subjects in non-controlled environments (i.e. video surveillance) becomes more robust in practical conditions.

Résumé

La reconnaissance faciale est une technologie importante en vision par ordinateur, avec un rôle central en biométrie, interface homme-machine, contrôle d'accès, indexation multimédia, etc. Au cours de ces dernières années, l'identification d'individus dans des scenarii non contrôlés, a attiré l'attention de la communauté des chercheurs en biométrie. Le déploiement de systèmes temps-réel et robuste de reconnaissance faciale peut augmenter de manière significative la sureté et sécurité des espaces publics ou privés. Cependant, les variabilités liées aux expressions/illuminations/poses/occultations peuvent dégrader sensiblement les performances d'un système de reconnaissance en environnements non contrôlés. L'occultation partielle, qui change complétement l'apparence d'une partie du visage, ne provoque pas uniquement une dégradation des performances en reconnaissance faciale, mai peut aussi avoir des conséquences en termes de sécurité.

Dans cette thèse, nous nous concentrons sur le problème des occultations en reconnaissance faciale en environnements non contrôlés. Pour cela, nous proposons une séquence qui consiste à analyser de manière explicite les occultations et à fiabiliser la reconnaissance faciale soumises à diverses occultations. Nous montrons dans cette thèse que l'approche proposée est plus efficace que les méthodes de l'état de l'art opérant sans traitement explicite dédié aux occultations. Nous identifions deux nouveaux types d'occultations, à savoir éparses et dynamiques. Des solutions sont introduites pour gérer ces problèmes d'occultation nouvellement identifiés dans un contexte de vidéo surveillance avancé. Récemment, le nouveau capteur Kinect a été utilisé avec succès dans de nombreuses applications en vision par ordinateur. Nous introduisons ce nouveau capteur dans le contexte de la reconnaissance faciale, en particulier en présence d'occultations, et démontrons son efficacité par rapport aux caméras traditionnelles. Finalement, nous proposons deux approches basées 2D et 3D permettant d'améliorer les techniques de base en reconnaissance de visages. L'amélioration des méthodes de base peut alors générer un impact positif sur les résultats de reconnaissance en présence d'occultations.

Dans cette thèse, des techniques de reconnaissance de visages robustes aux occultations sont proposés pour divers scenarii, de telle sorte que l'identification d'objets en environnements non contrôlés (i.e. vidéo surveillance) devient plus robuste en conditions réelles.

Acknowledgements

First of all, I would like to extend my deepest gratitude to my thesis advisor Prof. Jean-Luc Dugelay, who provides me this great opportunity to pursue a PhD at Telecom ParisTech in France. Prof. Dugelay is a great advisor with abundant experience in research and teaching. With his guidance and encouragement, I get to discover the exciting world of image processing research. It is indeed my great pleasure to work under the supervision of Prof. Dugelay, and I am really grateful for his continuous endorsement and persistent help on my thesis.

My sincere appreciation also goes to my thesis committee members for their enthusiastic supports. I would like to thank all the jury members for their precious time to read my manuscript. I appreciate the very helpful suggestions from Prof. Liming Chen and Prof. Abdenour Hadid to improve my manuscript.

I would like to thank Prof. Gerard Medioni and Dr. Jongmoo Choi for hosting me as a research intern in the computer vision lab at University of Southern California, where I spent a wonderful summer with unique research experience in Los Angeles.

My previous and current colleagues, including Antitza Dantcheva, Nesli Erdogmus, Carmelo Velardo, Hajer Fradi, Xuran Zhao, Neslihan Kose, have given me a lot of help from different aspects of my PhD study. I am very happy to work with them and share our experiences with each other. I must also thank the members of "Multimedia Chinese team" at EURECOM, including Yingbo Li, Dong Wang and Xueliang Liu, from whom I have learnt enormous knowledge in both research and life. I would like to thank all my friends at EURECOM for the colorful life I enjoyed during the past three years.

At last, I would give my heartfelt thanks to my family. My mother always encourages me in pursuing the truth and scientific essence. My father teaches me to think differently and always have a positive attitude. I write my thesis with the full proud of my parents. My love and appreciation goes to Isis Li, with whose accompany I am always inspired and blessed to face any challenge.

Contents

1 Introduction 1
1.1 Motivations . . . 1
1.2 Content of the Thesis . . . 5
1.3 Contributions . . . 7
1.4 Outline . . . 8

2 State-of-the-Art 11
2.1 Basic Tools and Concepts in Face Biometrics . . . 11
2.1.1 Standard Techniques . . . 12
2.1.2 Evaluation Metrics . . . 16
2.1.3 Challenges . . . 18
2.1.4 Face Database with Occlusions . . . 19
2.2 Review of Face Recognition under Occlusions . . . 20
2.3 State-of-the-Art Techniques . . . 22
2.3.1 Locality Emphasized Algorithms . . . 22
2.3.2 Sparse Representation based Classification (SRC) . . . 26
2.3.3 Occlusion Analysis + Face Recognition . . . 28
2.4 Conclusions . . . 32

3 Classical Occlusion Handling for Robust Face Recognition 35
3.1 Introduction . . . 35
3.2 Occlusion Analysis . . . 37
3.2.1 Occlusion Detection in Local Patches . . . 38
3.2.2 Occlusion Segmentation . . . 42
3.3 Face Recognition . . . 44
3.3.1 Improving LBP based Face Recognition . . . 45
3.3.2 Improving LGBP based Face Recognition . . . 49
3.4 Conclusions . . . 54

4 Advanced Occlusion Handling for Robust Face Recognition 55
4.1 Introduction . . . 55
4.2 Sparse Occlusion . . . 56
4.2.1 Overview . . . 57

4.2.2 Problem Statement . . . 58
4.2.3 Method . . . 59
4.2.4 Results . . . 62
4.3 Dynamic Occlusion . . . 66
4.3.1 Overview . . . 66
4.3.2 Challenges and Innovations . . . 67
4.3.3 Method . . . 68
4.3.4 Results . . . 73
4.4 Conclusions . . . 75

5 Exploiting New Sensor for Robust Face Recognition 77
5.1 Introduction . . . 77
5.2 Kinect Face Database . . . 79
5.2.1 Overview . . . 79
5.2.2 Review of 3D Face Database . . . 80
5.2.3 KinectFaceDB . . . 82
5.2.4 Experiments . . . 87
5.3 Depth Assisted 2D Face Recognition Under Partial Occlusions . . . 92
5.3.1 Overview . . . 92
5.3.2 RGB based Occlusion Analysis to Improve Face Recognition . . . 94
5.3.3 Depth based Occlusion Analysis to Improve Face Recognition . . . 96
5.3.4 Results . . . 98
5.4 Conclusions . . . 100

6 Improving Baseline Face Recognition Methods 103
6.1 Introduction . . . 103
6.2 Improving 2D Face Recognition via Combination of LBP and SRC . . . 104
6.2.1 Overview . . . 105
6.2.2 Background and Related Algorithms . . . 106
6.2.3 Method . . . 108
6.2.4 Results . . . 111
6.3 Improving 3D Face Recognition via Multiple Intermediate Registration . . . 114
6.3.1 Overview . . . 114
6.3.2 The System . . . 115
6.3.3 Results . . . 119
6.4 Conclusions . . . 122

7 Conclusions 123
7.1 Achievements . . . 123
7.2 Perspectives . . . 125
7.3 Conclusions . . . 126

Bibliography 127

List of Publications 140

Chapter 1

Introduction

1.1 Motivations

One fundamental problem in artificial intelligence is to understand and to mimic the remarkable ability of human visual system to process and recognize faces, which gained great attentions from different communities including neuroscience, computer science, statistics and psychology. Face recognition, as a inter-discipline research topic, has a long history of study [34, 39, 48, 95, 159] from the computer science community back to the early 80s. Along with the technological advancement, automatic face recognition has smoothly transferred from science fiction movies to daily life and becomes an essential technique for numerous real world applications. Enormous applications of face recognition can be found in intelligent systems, human-computer interaction, security management and access control system, video surveillance, ubiquitous computing, multimedia indexing, smart phones as well as search engines.

Face recognition is a particular entity of the more general biometric concept. According to ISO[1]/IEC[2] JTC1 SC 37, biometrics refers to the automated recognition of individuals based on their behavioural and physiological characteristics/traits. To attain a robust and reliable biometric system, good biometric characteristics shall satisfy the following desirable properties: universality, uniqueness, permanence, collectivity (measurability), and acceptability (user friendliness). Some biometric characteristics are known to have strengths on part of those properties but weaknesses on the others. A non-exhausting list of biometric traits is given here as following: the physiological traits: face, fingerprint, iris, retina, hand geometry, vein structure of hand, ear geometry, and Deoxyribonucleic acid (DNA); the behavioural traits: signature, voice, and keyboard strokes. Among all those biometric traits, face, fingerprint [81] and iris [49] are the most popular ones thanks to their robustness, reliability and accessibility.

[1] International Organization for Standardization
[2] International Electrotechnical Commission

As probably the most popular biometric trait, face possesses its intrinsic advantages for real world applications. Unlike a fingerprint scanner that requires touch access for data acquisition, and iris recognition that requires precision instruments and close-distance data capturing, recognition based on face requires less user cooperation and thus can be integrated in many advanced conditions (notably for video surveillance applications). With a single camera, recognition of people based on face can be accomplished either in controlled environments (such as access control system) or in uncontrolled environments (such as crowed scene in video surveillance).

Nevertheless, with emphasis on real world applications, face recognition suffers from a number of problems in the uncontrolled scenarios. Those problems are mainly due to different facial variations which can greatly change the facial appearance, including facial expression variations, illumination variations, pose changes as well as partial occlusions. In the last decade, enormous amount of works are proposed to overcome the expression/illumination/pose problems in face recognition, and significant progresses have been made. Except very drastic expression changes (e.g. yawing, screaming), most recent face recognition algorithms obtained good performance on face data with regular expression changes (e.g. smile, anger, cry). It is also well known that subspace based method can address the illumination problem with sufficient training based on the fact that Lambertian reflectance with arbitrary distant light sources lies close to a 9D linear subspace [20], and local texture descriptors such as local binary patterns (LBP) [17] is robust to monotonic gray level changes due to illumination. The pose variation can be easily handled in 3D, since matching of the corresponding parts can be done by rigid (ICP [24, 40]) and non-rigid (TPS [55]) registration. Similarly, multiple images can be approached to address the pose problem in 2D. In the literature, many algorithms reported their results which demonstrated their robustness to more than one of those variations (e.g. expression&illumination invariant recognition, or illumination&pose invariant recognition).

However, in comparison to the great number of works in the literature toward the expression/illumination/pose problems, the occlusion problem is the relatively overlooked one by the research community. Many of the important works, such as Eigenface [140], Fisherface [21] or LBP based face recognition [17], did not take into account the occlusion problem intentionally, and thus are not robust when partial occlusion occurs. Some previous works [62, 84, 89, 96, 107, 116, 120, 127, 138, 146, 150, 151, 156, 157, 160] do target on the occlusion problem and demonstrate certain robustness to the presence of partial occlusions. However, the results are still far inferior than the average performance of face recognition without occlusion. And many of those algorithms suffer from the generalization problem that are not suitable/compatible to the faces without occlusion and with/without other types of facial variations (e.g. expression/illumination/pose). Last but not the least, almost all of those algorithms are tested in well controlled lab conditions (i.e. high resolution face images wearing sunglasses or scarf under mild illumination conditions, or synthetic occlusion (normally by a black square) on the face databases without occlusion), and very like to fail in many more advanced occlusion conditions due to the large variety of facial occlusions in uncontrolled environments.

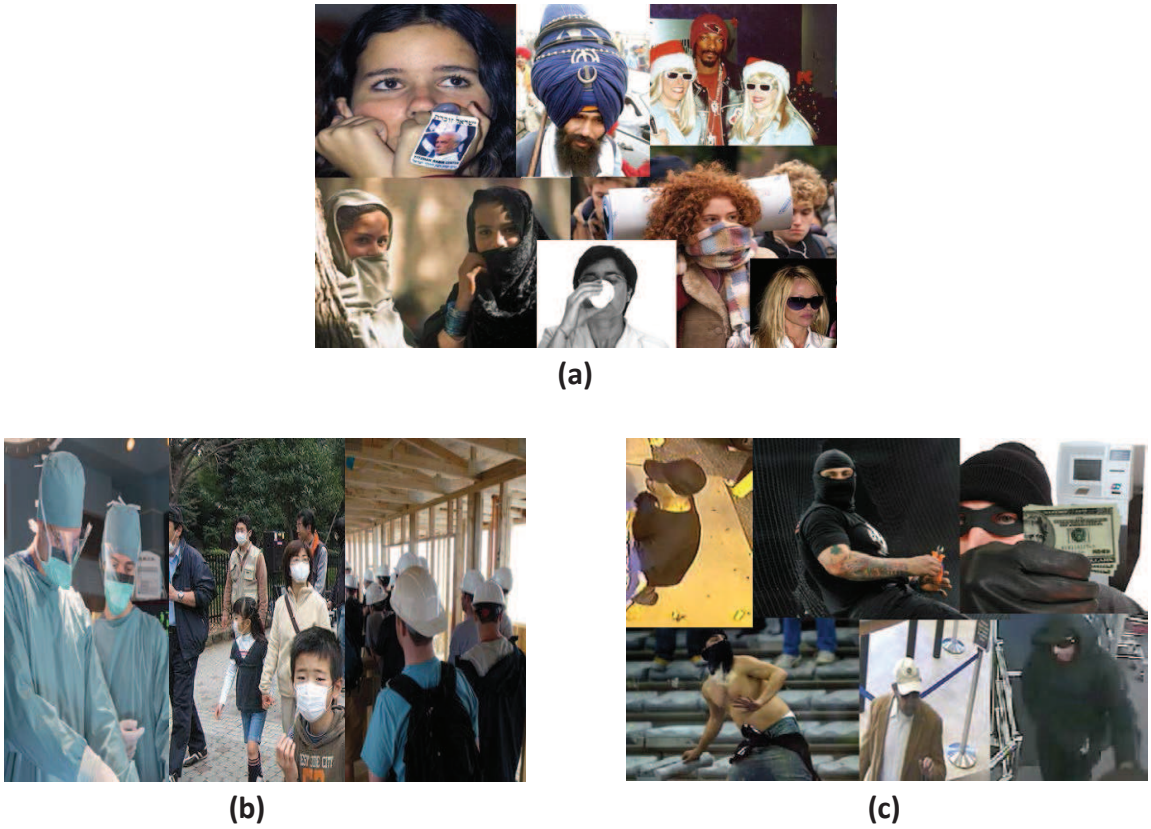


Figure 1.1: Examples of facial occlusion in different categories: (a) occlusions in daily life; (b) occlusion related to safety issues; (c) occlusion related to security issues.

Although limited attentions were drawn to the occlusion problem within the face recognition literature, the importance of this problem must be emphasized since the presence of occlusion is very common in uncontrolled scenarios and may be associated with several safety and severe security issues. From the user's perspective, facial occlusions may occur for several intentional or undeliberate reasons. First of all, facial accessories like sunglasses, scarf, facial make-up and hat/cap are quite common in daily life. Some other people also wear veils for religious convictions or cultural habits. Examples of occlusions observed in daily life are shown in Figure 1.1 (a). In the intelligence or surveillance applications, user friendliness is the most important property that must be considered, where no user cooperation can be expected. The system must be able to organize people no matter how big such noise appears. Secondly, occlusion is also appearing in safety scenarios. For example, surgical/procedure mask is required in the restricted areas of hospital, and it is often wearing by people in the eastern Asian (e.g. China, Japan) to prevent the exposures to air pollution, respiratory diseases or pollen allergy. Also in the construction areas, safety helmet is vital for human-beings in such areas [54]. Examples of safety related occlusions are displayed in Figure 1.1 (b). Last but not least, facial occlusions are often related to several severe security issues. Football hooligans and ATM criminals tend to wear scarf and/or sunglasses to prevent their faces from being recognized. Bank robbers and shop thieves usually wear a cap when entering in places where they commit illegal actions. The increasing amount of

recent police reports gave testimony to such issues. Some examples of the security related occlusions can be found in Figure 1.1 (c).

Handling all of above mentioned occlusions in face recognition is essential for both intelligence and security/safety/law enforcement purposes. As stated before, identifying people without any cooperation of removing occlusion due to facial accessories brings great convenience and comfortableness for users in numerous scenarios. On the other hand, identifying the presence of must occlusions in restricted places (e.g. hospital, construction area) and revealing the people's identity in such areas guarantee the safety in the environment. Similarly, detecting the presence of occlusion can identify suspicious people in certain areas (e.g. football stadium, ATM machine, shops, airport), and face recognition (in spite of the presence of occlusion) in such areas can help police to find criminals/fugitives. In sum, to achieve occlusion robust face recognition is very important and has many potential usages in the surveillance world.

The focus of this thesis is thus on the occlusion problem in face recognition. Based on the study of literature works (details of our study will be presented in Chapter 2), in order to make contributions to face recognition robust to occlusion, the following questions naturally arise:

1. What is the efficient way to eliminate the effect of occlusion in the face recognition process?
2. Can the prior information of occlusion (e.g. size, location, structure, intensity) be helpful to improve face recognition in presence of occlusion?
3. How to obtain such information of occlusion in an automatic manner?
4. How to effectively incorporate such information into the matching process of face recognition?
5. Other than the well-studied occlusions in the literature (such as sunglasses and scarf), are there other types of occlusion problems that must be addressed in practical scenarios? And what are they?
6. If such occlusions exist, how to efficiently address those problems?
7. How to leverage state-of-the-art computer vision and image processing techniques to handle the occlusion problem in face biometrics?
8. Can the emerging new sensor (such as Microsoft Kinect) be more helpful to improve face recognition under occlusion conditions than the traditional RGB cameras, and how?
9. Can 3D information be exploited to improve 2D face recognition in the sense of occlusion handling, or vice versa?

Toward above positioned questions, in this thesis, we present a complete study of face recognition robust to occlusions and our original contributions to the state-of-the-art. In the following sections, we first give a general description of this thesis in Section 1.2. Then we briefly summarize our contributions in Section 1.3. Outline of each chapter is given in Section 1.4.

1.2 Content of the Thesis

In this thesis, we give a complete study of the occlusion problem in automatic face recognition. In addition to the literature review, there are mainly four parts of works that are thoroughly investigated which are in correspondence to the content of Chapter 3-6. They can be summarized as bellow:

- Part 1: Handling the classical occlusion problems in face recognition.
- Part 2: Handling the more advanced occlusion problems in face recognition.
- Part 3: Exploiting new sensor (i.e. Kinect) to improve face recognition under occlusion conditions.
- Part 4: Improving the baseline face recognition algorithms.

Part 1 mainly studies the classical occlusion problems (dense and static facial occlusions, such as scarf and sunglasses) which are already extensively investigated in the literature. Our approach consists of first applying explicit occlusion analysis to exclude the occluded part in local feature based face representation (i.e. LBP [17] and LGBPHS [157]), then conducting face recognition from the non-occluded face part only. This approach takes both advantages of information selectivity from occlusion exclusion and the robustness from local feature based face representation. Toward this goal, we present a novel framework for improving the recognition of occluded faces due to facial accessories, in which new techniques to detect and segment facial occlusions are presented. Extensive experimental analysis is conducted, demonstrating significant performance enhancements in comparison to state-of-the-art methods under various occlusion conditions. Our work in this part advocates that the prior information of occlusion can be leveraged to improve even the most sophisticated face recognition algorithms.

In Part 2, we first figure out that works in the literature could not handle occlusions in some more advanced conditions. We identify two new types of facial occlusions. The newly identified facial occlusions - sparse occlusion and dynamic occlusion are in contrast to the classical ones with dense and static properties. Almost all works in the literature impose the dense and static assumption in their ways to address the occlusion problem and therefore they are likely to fail for the occlusions that are sparse or dynamic. By giving formal definitions, we find that facial painting, face dirt or face behind fence belong to the sparse occlusion category and cap detection for moving people in entrance surveillance is a typical example of

dynamic occlusion analysis. In this part, we present some first solutions dedicated to those problems in the context of face biometrics in video surveillance. To handle the sparse occlusion problem, we propose to automatically detect sparse occlusion on faces and then inpaint the occluded part. Occlusions are thus removed from the inpainted faces. Significant improvements are observed via extensive experiments for faces with various sparse occlusions. For the dynamic occlusion handling, we propose a system to detect the presence of cap for moving people in entrance surveillance. The system automatically detect, track and classify faces in complex surveillance scenarios and can be applied to a broad range of security and forensics applications.

Unlike the literature works based on traditional RGB camera (or laser scanner in the 3D case) for face recognition (as well as the occlusion problem), in Part 3, we leverage the capability of the emerging new sensor - Kinect for handling the occlusion problem in face recognition. For such a purpose, we build the first publicly available face database (KinectFaceDB) based on the Kinect sensor, which contains well-aligned 2D, 2.5D and 3D face data as well as multiple well organized facial variations including three types of occlusions. We also give the details for data acquisition from Kinect, benchmark evaluations, and comparison to another well recognized face database (i.e. FRGC) in terms of 3D facial data quality in biometrics. With the embodiment of the built KinectFaceDB and the RGB-D sensing capability of Kinect, we propose a new approach to address the partial occlusion problem for face recognition in this part. We exploit heterogeneous cues (both depth and RGB) from Kinect acquisition for occlusion analysis and face recognition, respectively. Similar to the framework we presented in previous parts, an explicit occlusion analysis module is included, but the novelty lies on the use of depth information. Where the occlusion analysis resulting from depth are then incorporated with face recognition based on RGB images. We demonstrate via experiments that in comparison to the traditional 2D sensors, Kinect do greatly improve face recognition in presence of partial occlusions.

Generally speaking, the framework of occlusion analysis plus local feature based face recognition is compatible with any local methods. In Part 4, we argue that improving standard face recognition algorithms can further improve the results for occluded faces under the proposed framework. In this part, we improve both 2D and 3D face recognition based on standard methods. In the 2D case, LBP based face representation is combined with Sparse Representation based Classification (SRC) with a multi-resolution architecture. The proposed method overcomes the "curse of dimensionality" problem via the divide-and-conquer and strengthens the discriminative power via its pyramidal architecture thanks to the information description from different levels. In the 3D case, we improve Iterative-Closest-Points (ICP) based face recognition by an indirect face registration architecture with multiple intermediate registration references. A fast and accurate online 3D face identification system based on Kinect/PrimeSensor is implemented, on which the proposed face registration scheme is justified. Competitive performances of both proposed 2D and 3D face recognition methods are observed on the testing sets.

1.3 Contributions

This thesis is dedicated to handle the occlusion problems in face recognition from different aspects. The main contribution of this thesis could be summarized as follows.

1. We focus on an important however somewhat overlooked problem in face recognition. The framework to combine explicit facial occlusion analysis and processing with local feature based face recognition is emphasized. The proposed framework is different from most of the works in the literature. We have justified its efficiency in various scenarios when facial occlusion occurs.

2. We give a complete review of works in the literature for face recognition under occlusion conditions. State-of-the-art works are carefully categorized and elaborated. Advantages and limitations of the literature works are carefully considered which can instruct the algorithm design for our own work and the successive works.

3. We present a series of processing methods including occlusion detection, occlusion segmentation, selective-LBP and selective-LGBPHS to address classical facial occlusion problems caused by facial accessories (such as sunglasses and scarf). We prove via experiments that occlusion exclusion is more efficient than occlusion weighting since information from the occluded part can be completely discarded in the recognition process. In comparison to the literature works, our approach achieves the state-of-the-art result and does not suffer from the generalization problem. (This work was published at EUSIPCO 2010 and FG 2011.)

4. We identify two new types of facial occlusions - the sparse occlusion and the dynamic occlusion which are never studied in the literature according to the best of our knowledge. Solutions to those more advanced occlusion scenarios are provided.

5. We propose an approach of first detecting the presence of occlusion in local pixels based on the framework of Robust-PCA, and then inpainting the occluded pixels given the information from the occlusion detection part for face recognition. This approach can significantly improve various face recognition algorithms in complex sparse occlusion scenarios. This work might also draw attentions from both face recognition researchers from the biometric community and inpainting researchers from the graphic community. (The proposed approach was published at ICIP 2012.)

6. We present a system to detect cap for moving people in the entrance surveillance scenario. Our system is able to find suspicious person for restricted areas (e.g. bank, football stadium) and provide occlusion information for face recognition. It may also be applied to many security management and surveillance systems. (The cap detection system was presented at ACM Multimedia 2012.)

7. We build the first publicly available face database - the KinectFaceDB based on the emerging Kinect sensor. The database consists of different data modalities (well-aligned and processed 2D, 2.5D, and 3D based face data) and multiple facial varia-

tions. We conduct benchmark evaluations on the proposed database using standard face recognition techniques, so as to provide the baseline results for face recognition studies using KinectFaceDB. We also report the performance comparisons between Kinect images and traditional high quality 3D scans (FRGC database) in the context of face biometrics. The published database received great interests from face recognition researchers (50+ requests in couples of weeks), and therefore serve as a bridge between face recognition research and the emerging Kinect technology. (Results for the proposed KinectFaceDB are in submitting to IEEE Transaction on System, Man and Cybernetics: Systems, special issue Biometric Systems and Applications.)

8. We propose an approach to conduct facial occlusion analysis based on the depth information from Kinect to improve face recognition based on the RGB image from Kinect. The proposed approach demonstrates its advantages over traditional occlusion analysis and face recognition solely based on 2D. This work also reveals the capability of the Kinect sensor in handling the occlusion problem in face recognition. (The proposed method has been submitted at ICIP 2013.)

9. We present a new algorithm which gives an improved combination of LBP based feature extraction and SRC based classification for 2D based face recognition. In comparison to the state-of-the-art work, we overcome the "curse of dimensionality" problem by the divide-and-conquer strategy and improve the discriminative power via a pyramidal architecture. (The improved combination of LBP and SRC was published at ICME 2011.)

10. We propose an intermediate face registration architecture via multiple references for 3D face recognition. Instead of registering a probe to all instances in the database, we propose to only register it with several intermediate references, which considerably reduces processing, while preserving the recognition rate.

11. We implement a real-time 3D face identification system using a depth camera. The system routinely achieves 100% identification rate when matching a (0.5-4 seconds) video sequence, and 97.9% for single frame recognition. These numbers refer to a real-world dataset of 20 people. The methodology extends directly to very large datasets. The process runs at 20fps on an off the shelf laptop. (This real-time 3D face identification system was published at ICPR 2012.)

The works presented in this thesis led to a number of publications on the international venues. The list of publications can be found at the end of the thesis.

1.4 Outline

The rest of this thesis is structured as follows:

Chapter 2 discusses the background and the state-of-the-art for face recognition under occlusion conditions. Starting by a more general look at face biometrics (Section 2.1), we elaborate the difficulties of occlusion problem in face recognition and summarized the works in the literature toward this problem (Section 2.2). Representative algorithms are then categorized and looked into more details (Section 2.3).

Chapter 3 presents our solutions to the classical occlusion problems in face recognition (mainly due to facial accessories such as sunglasses and scarf). For occlusion analysis (Section 3.2), we present our approach consisted of the occlusion detection algorithm (Section 3.2.1) and the occlusion segmentation algorithm (Section 3.2.2). For face recognition (Section 3.2), integration of the proposed occlusion analysis with LBP (Section 3.3.1) and LGBP (Section 3.3.2) based recognition are illustrated, respectively.

Chapter 4 reports two new types of facial occlusions, namely the sparse occlusion (Section 4.2) and the dynamic occlusion (Section 4.3), in advanced conditions. For handling the sparse occlusion problem, we propose explicit occlusion detection (Section 4.2.3.2) and then performing inpainting on the occluded pixels (Section 4.2.3.2). For dynamic occlusion, our system of detecting cap implements head detection/tracking (Section 4.3.3.1) and dynamic occlusion detection (Section 4.3.3.2).

Chapter 5 introduces the work based on a new sensor (i.e. Kinect) to improve face recognition in presence of occlusions. The proposed KinectFaceDB (Section 5.2) is explained with 3D face database review (Section 5.2.2) and acquisition details (Section 5.2.3.2). Then the benchmark evaluations (Section 5.2.4.1) and comparison with FRGC (Section 5.2.4.2) are presented. Our method of using depth based occlusion analysis (Section 5.3.3.2) to improve RGB based face recognition (Section 5.3.3.3) is given in Section 5.3.

Chapter 6 describes our approaches to improve the baseline face recognition algorithms in both the 2D (Section 6.2) case and the 3D case (Section 6.3), which could potentially upgrade the recognition rates in occluded conditions. Motivating the utilization of divide-and-conquer strategy (Section 6.2.3) and intermediate face registration via multiple references (Section 6.3.2) are presented in this chapter, respectively.

Conclusions are given in Chapter 7 which summarizes our major achievements as well as the potential future works.

Chapter 2

State-of-the-Art

Although state-of-the-art face recognition methods perform with high accuracy under controlled environments, variations such as extreme illuminations, pose changing and partial occlusion can significantly deteriorate the recognition results in uncontrolled environments (e.g. in video surveillance). Among the most common facial variations, partial occlusion is recognized as a very challenging problem in face recognition [56], and therefore increasingly received attentions from the biometric research community in recent years.

In this chapter, we will present the state-of-the-art of face recognition under occlusion conditions. To achieve this goal, we will firstly revisit some basic tools and concepts in face biometrics (for a complete review please refer to [95]) because the tools and concepts described in this section will be extensively reused in the evaluations of our works. Three most representative techniques – the Eigenface [140], the Fisherface [21] and the LBP based face recognition [17] are reviewed in Section 2.1. Within the same section, we will also present our choice of identification rates as the evaluation metric, as well as the standard face database (AR face database) in face occlusion research. Then in Section 2.2, we will give the literature review of face recognition in occluded conditions. The representative state-of-the-art techniques from different categories are reviewed in Section 2.3.

2.1 Basic Tools and Concepts in Face Biometrics

Face recognition [159], the least intrusive biometric technique from the acquisition point of view, has been applied to a wide range of commercial and law enforcement applications. In comparison to other popular biometric characteristics (such as fingerprint [81] and iris [49]), biometric recognition (for either identification or verification) using face requires less user cooperation and thus can be integrated in many advanced conditions (notably in video surveillance). In this section, we will give a brief review of general face recognition by introducing the following four parts: (1) standard techniques, (2) evaluation metrics, (3) challenges and (4) face database with occlusions. In the first part we will review the three most

representative techniques, namely the Eigenface [140], the Fisherface [21] and the LBP based face recognition [17]. In part two, the identification mode and verification mode with their corresponding evaluation metrics are described. We then discuss the most often facial variations including expression, illumination, pose and occlusion in the third part. Part four gives the overview of the AR face database.

2.1.1 Standard Techniques

In this section, we briefly review three most widely used (and the mostly cited as well) face recognition techniques, namely Eigenface [140], Fisherface [21] and LBP based face recognition [17]. All those techniques are considered as the benchmarks for more advanced algorithm development. The good understanding of these methods is a key to design and develop occlusion-robust face recognition algorithms.

2.1.1.1 Eigenface

The idea of projecting a high dimensional face representation into a lower dimensional feature space (namely a face subspace) relies on the fact that well-aligned face images possess a high degree of correlation from a statistical point of view, and thus can be reconstructed by a linear combination of fewer components. Eigenfaces [140] are normally considered as the "standardized face ingredients", where the principal component analysis (PCA) [86] is applied to face data which maximizes the total scatter across all images of all faces.

Given a set of N face images $\{x_1, x_2, ..., x_N\}$, where $x_k \in R^M$, the Eigenface method seeks a linear projection W' that translate the original M dimensional data x_k into a lower dimension D (where $D \ll M$ and $D < N$) as below:

$$y_k = W'^T x_k \quad k \in [1, N] \tag{2.1}$$

where $W' \in R^{M \times D}$, and $y_k \in R^D$ is the projected face using the first D eigenvectors in the sense of $l2$ error reconstruction.

In order to find the set of eigenvectors $W \in R^{M \times N}$ (the base for the projection matrix W') for dimensionality reduction, the covariance matrix $S \in R^{M \times M}$ of all training data (which reflects the total scatter of all faces) is calculated:

$$S = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T \tag{2.2}$$

where $\mu \in R^M$ is the mean of all faces. Here S is the scatter matrix of faces in the original feature space $\{x_k \in R^M, k \in [1, N]\}$,the purpose of PCA is to maximize the scatter matrix $\hat{S}$ of faces in the projected space $\{y_k \in R^D, k \in [1, N]\}$ by finding the optimal projection, which

can be written as:

$$\hat{W} = arg\max_{W} |W^T S W| \tag{2.3}$$

where $\hat{W} \in R^{M\times N}$ is the set of N eigenvectors $\{w_k \in R^M, k \in [1, N]\}$ ordered by their corresponding eigenvalues (in the descending order). By selecting the first D eigenvalues out of N, we can form the projection matrix $W' \in R^{M\times D}$ (as shown in Equation 2.1) and subsequently project a face into D-dimensional subspace. The projected representation $y_k \in R^D$ is then used for face recognition. The new representation is thus less redundant and more parsimonious for classification.

2.1.1.2 Fisherface

The Eigenface method conducts linear dimensionality reduction prior to face recognition in the sense of optimal reconstruction. However, in may practical scenarios (for example, extreme illumination and expression occurs), the difference between the same identity with different variations can be much larger than the difference between 2 different people (usually with same variation). Supposing there are sufficient training samples for all people and all different facial variations, in cooperating the class label information can be helpful to increase the discriminative power in the projecting subspace.

Fisherface [21] is then proposed to enforce the label information in the projection via Fisher Linear Discriminative (FLD) analysis [63]. Instead of computing the overall scatter of the input data, 2 different variances of the data are measured by considering the class/label information, namely the between-class scatter matrix S_B:

$$S_B = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \tag{2.4}$$

where C is the number of classes, μ_i is the mean face of class i, and μ is the mean face of all the faces in the training; and the within-class scatter matrix S_W

$$S_W = \sum_{i=1}^{C} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \tag{2.5}$$

where X_i is the set of faces in class i. Considering both the between-class scatter and within-class scatter, the objective which can maximize the discriminative power according to class label is therefore to find an optimal projection $\hat{W} \in R^{M\times C}$ by maximizing the between-class scatter meanwhile minimizing the within-class scatter, which can be written as follows:

$$\hat{W} = arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{2.6}$$

Similar to Eigenface, we seek the first D eigenvectors in the set of ordered eigenvectors $\hat{W}$ for face projection. Works in the literature have demonstrated the superiority of Fisherface to Eigenface. An visualization example of Eigenface and Fisherface is shown in Figure 2.1.

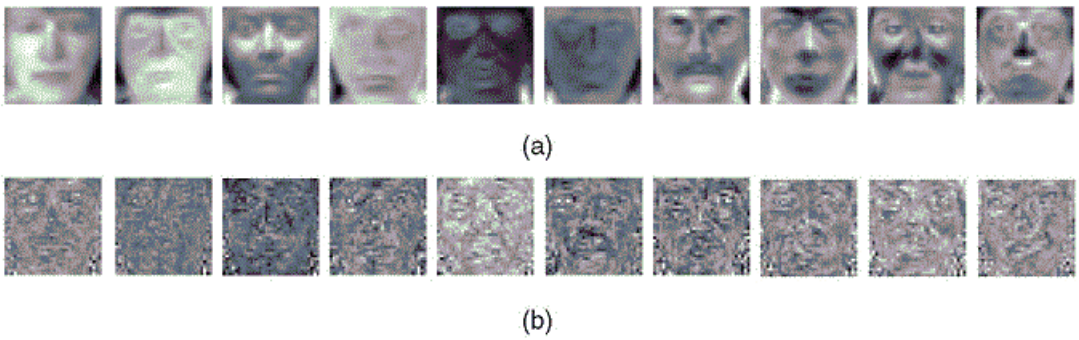


Figure 2.1: Visualization of (a) Eigenface and (b) Fisherface constructed from Yale face database [92] (Image excerpted from [71]).

2.1.1.3 Local Binary Patterns

Since firstly introduced by Ahonen et al. [17], the Local Binary Pattern (LBP) based representation has become the most popular technique for face recognition in recent years, due to its computational simplicity, discriminative power and robustness (notably for the invariance against monotonic gray level changes). LBP algorithm rapidly gained popularity among researchers and numerous extensions (e.g. [68, 76, 97]) have been proposed which prove LBP to be a powerful measure of image texture and in particular for face images (a complete survey can be found at [75]). The LBP method is used in many kinds of applications, including image retrieval, motion analysis, biomedical image analysis and distinguishably in face recognition.

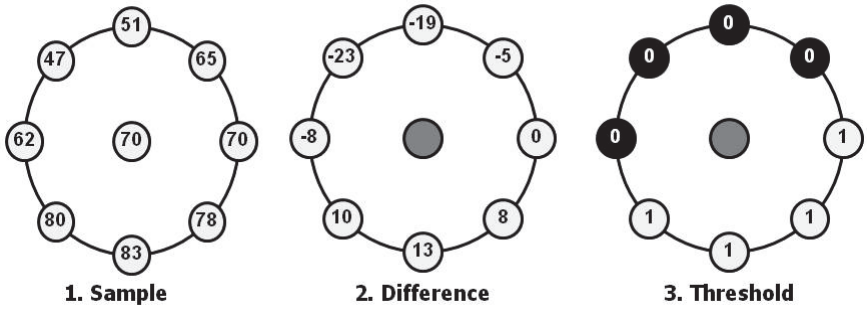


Figure 2.2: Visual illustration of LBP thresholding.(Image excerpted from [1])

The LBP operator originally forms labels for the image pixels by thresholding the 3×3 neighbourhood of each pixel with the center value and considering the result as a binary number (A visual illustration is shown in Figure 2.2). A histogram of these labels, is created as the

texture descriptor, by collecting the occurrences. The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{2.7}$$

where g_c corresponds to the gray value of the center pixel (x_c,y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \{ \begin{array}{ll} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{array} \tag{2.8}$$

Successively, the histograms that contain data about the distributions of different patterns such as edges, spots and plain regions are built (which are usually summarized from different local facial regions), and the classification is performed by computing their similarities. Among various proposed dissimilarity measures, the χ^2 distance is the mostly used one for LBP based face representation as below:

$$\chi^2(H_1.H_2) = \sum_{i}^{N} \frac{(H_{1,i} - H_{2,i})^2}{H_{1,i} + H_{2,i}} \tag{2.9}$$

where H_1 and H_2 are feature histograms and the special case is defined as $\frac{0}{0} = 0$.

2.1.1.4 Remarks

The three mostly used face recognition techniques we reviewed in this section have been proven to have good performance under controlled lab conditions. However, all of them suffer from real world variations especially for the occlusion problem.

Eigenface is a basic technique which can only deal with minor facial changes (e.g. small facial expressions). Fisherface can be more robust to larger variations, only if sufficient training samples of those variations are included in the training set. It reports good performance on many face databases because all those face databases have a large number of well organized facial variations (illuminations, expressions etc.) in the gallery. Notably, it cannot address the occlusion problem for the following two reasons: (1) occluded faces are not included in the gallery for practical systems; (2) occlusion appearance and location can be randomly distributed and unpredictable therefore cannot be exclusively considered in the training phase. In the literature, both Eigenface and Fisherface are considered as "holistic" method, because each coefficient in the projected subspace is correlated with all original pixels to some extent. Changes of the original pixel values can affect all components for the face representation. Therefore, Eigenface and Fisherface are known to be sensitive to facial occlusion.

LBP descriptor for face representation is well known for its robustness to monotonic gray scale changes caused by, for example, illumination variations. The use of histograms as features also makes the LBP approach robust to face misalignment and pose variations to some extent. In addition, because LBP based face representation is a local patch based method, changes of pixel values caused by partial occlusion can only affect parts of the entire face representation in the face recognition. Therefore, LBP based approach is more robust to occlusion than Eigenface and Fisherface. Nevertheless, without explicit handling of the corruptions, partial occlusion can still significantly deteriorate face recognition results based on LBP.

2.1.2 Evaluation Metrics

Since the standard biometric system can be implemented in two different modes [93] (i.e. the identification (one-to-many) mode and the verification (one-to-one) mode), face recognition algorithms are subsequently evaluated under two metrics: the identification metric and the verification metric. In this section, we briefly review the 2 face recognition modes and the corresponding evaluation metrics and justify our choice of evaluation method for the works presented in this thesis.

2.1.2.1 Identification

According to Jain et al. [82], in the identification mode, the system recognizes an individual by searching the templates of all the users in the database for a match. This can be done by the similarity measure between the probe face and all gallery faces, and find the matching of smallest distance, which takes a time complexity of $O(N)$ (assuming there is N faces in the gallery). The identification system is normally deployed in the non-cooperate systems (for example in video surveillance systems or intelligent interactive systems), so that more types of facial variations (e.g. extreme illumination conditions and occlusions) have to been considered in the system design.

The evaluation metric of a face recognition algorithm in the identification mode can be expressed in the following ways: **Rank-1 identification rate** and **Cumulative Matching Characteristics (CMC)** [95]. The rank-1 identification rate refers to the rate of correctly matched probe faces in the testing pool. It is the mostly used face biometric metric and usually named as "recognition rate". The other performance measure is the CMC score (see Figure 2.3, in the format of ROC (Receiver Operating Characteristic) curve), which shows the recognition rate as a function of the rank given to the probe identity by the face recognition system.

Another recently considered biometric problem is the open-set identification problem, where a probe face is firstly judged for whether or not it belongs to the people who are searching for, and then identify which ID it associates with in the gallery. In this case, a probe can either belong to one of the IDs in the gallery or it do not belong to any ID. This specific scenario is not considered in this thesis.

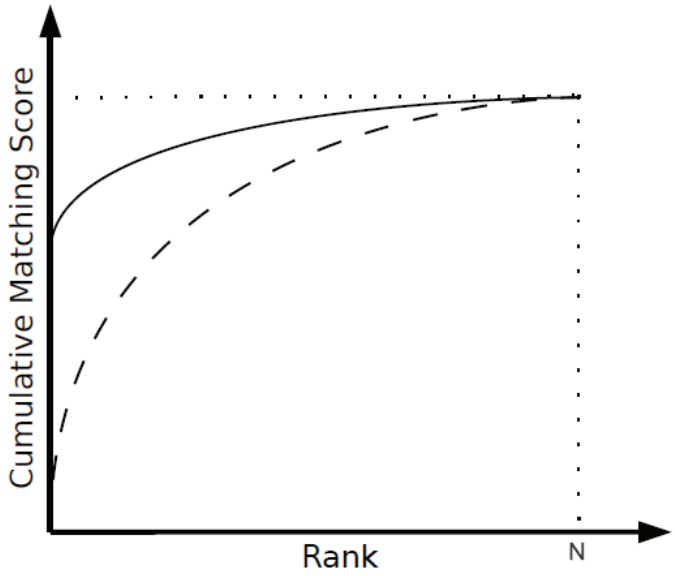


Figure 2.3: Illustration of the Cumulative Matching Characteristics (CMC) in the format of ROC curve. The solid line represents the system that performs better. N is the number of subjects in the gallery. (Image excerpted from [142])

2.1.2.2 Verification

According to Jain et al. [82], in the verification mode, the system validates a person's identity by comparing the captured biometric data with her/his own biometric template(s) stored in the system database. Which means given a probe face we compute the similarity score (distance) between itself and its claimed gallery face. If the distance is smaller than a predefined threshold, the probe is verified, otherwise it is rejected. This process takes only $O(1)$ time complexity. The verification system is normally deployed in the cooperative biometric systems (for example in biometric security management or access control system). Therefore less robustness (to extreme facial variations) is required. However, the reliability is highly desirable and several malicious issues (such as face spoofing [103, 104]) shall be carefully handled.

The evaluation metric of a face recognition algorithm in the verification mode can be summarized as follows: the **False Acceptance Rate (FAR)**, the **False Rejection Rate (FRR)** and the **Equal Error Rate (EER)**. FAR refers to the rate of wrongly accepted imposters (as clients) and FRR refers to the rate of wrongly rejected clients (as imposters), and EER is the rate where FAR=FRR. Those statistics are usually shown in the format of ROC (Receiver Operating Characteristic) curve. In general concepts, the error rates (FAR, FRR and EER) are the lower the better. However, since FAR is inversely proportional to FRR, the system is designed to full-fill the specification of either to be reliable or convenient. Figure 2.4 shows an example of clients/imposters distribution in the verification mode as well as the corresponding evaluation metric.

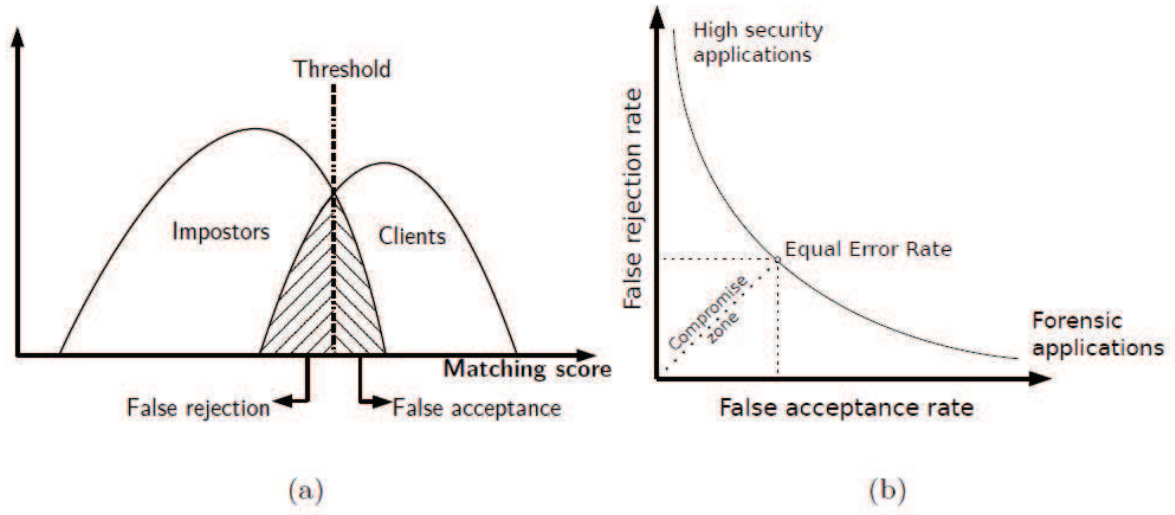


Figure 2.4: Illustration of typical examples of biometric system graphs, the two distributions (a) represent the client/impostor scores; by varying the threshold different values of FAR and FRR can be computed. A ROC curve (b) is used to summarize the operating points of a biometric system, for each different application different performances are required to the system. (Image excerpted from [142])

2.1.2.3 Remarks

In the test of general face recognition algorithms, either identification rates or verification rates can reflect the correctness of the proposed algorithm. Some works in the literature do report both for justification. However, almost all papers [62, 84, 89, 96, 107, 120, 122, 138, 146, 150, 151, 157, 160] related to the occlusion problem conduct experiments in the identification mode only. This is because occlusion problem does not occur in most cooperative verification systems (the purpose of both a client and an imposter is to be verified by the system, the later case refers to the spoofing attack). On the other hand, occlusion is an important issue in the identification systems, which can significantly deteriorate the system performance and often relates to several severe security issues. For above reasons, we report the identification rates (namely the recognition rates) for the works we presented in this thesis.

2.1.3 Challenges

The challenges to face recognition systems are often occurring in practical scenarios and can significantly deteriorate the recognition performance. Based on the works in the literature, here we summarize the main challenges of face recognition in the following categories:

- **Facial expression variations:** Changes caused by facial expressions (such as smile, anger, yawing, laughing etc.) can change the regular (the neutral expression) appearance and surface structure of a face and therefore affect the recognition result.
- **Illumination variations:** Strong lights from different directions can cause dramatic changes in facial appearance.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 [161].
- **Pose variations:** pose is a major problem in 2D based face recognition since the alignment of two 2D face images with different poses is a difficult problem. However it is a readily addressed problem in 3D, since the surface alignment can be easily achieved by rigid (ICP [24, 40]) and non-rigid (TPS [55]) registration.
- **Occlusion:** any external object (other than the original face surface) in between the face surface and the camera can cause the occlusion problem. Common occlusions are mainly due to facial accessories such as sunglasses, scarf and hat, as well as hand and phone on a face. However, some unusual occlusions can still occur by for example disguises, cosmetics, and face behind fence. We will discuss in details different types of occlusions in the later parts of this thesis.

In addition to those general problems, face recognition also suffers from some protocol problems such as single sample face recognition, spoofing attacks etc. The discussion to those challenges is out of the scope of this thesis.

While there has been an enormous amount of research on face recognition under pose/illumination/expression changes and image degradations, problems caused by occlusions are relatively overlooked. In this thesis, we focus on the occlusion problem and present a complete study of different aspects of the occlusion problem in face recognition.

2.1.4 Face Database with Occlusions

Face databases (for example, The Facial Recognition Technology (FERET) [125] and Face Recognition Grand Challenge (FRGC) [123] by National Institute of Standards and Technology (NIST)) challenges the standard techniques by the positioning of different variations (expressions, illuminations, poses etc.), large number of subjects, as well as new data modalities (2D, 3D, video etc.). On the other hand, development of robust and reliable face recognition systems greatly relies on the existing face databases.

Unlike the enormous amount of research on face recognition under pose/illumination/expression changes and image degradations, problems caused by occlusions are mostly overlooked. Very few face databases includes partial occlusion as a major variation (typical examples can be found in the AR face database [105] for 2D and UMB-DB [44] for 3D). From almost all the works in the literature, the AR face database is recognized as the benchmark database for the occlusion problem in 2D based face

recognition because it contains a large number of well organized real world occlusions (other alternatives are mostly artificially generated occlusions (e.g. black square) on the regular face databases such as FERET [125] or Yale face database [92]).

Variations in the AR face database can be summarized as follows:

- 01: Neutral expression
- 02: Smile
- 03: Anger
- 04: Scream
- 05: left light on
- 06: right light on
- 07: both side lights on
- 08: wearing sun glasses
- 09: wearing sun glasses & left light on
- 10: wearing sun glasses & right light on
- 11: wearing scarf
- 12: wearing scarf & left light on
- 13: wearing scarf & right light on
- 14: second session (same conditions as 01 to 13)

From above structure we can observe that almost half of the faces in the databases contains partial occlusion (by either a sunglasses or a scarf). The whole database contains more than 4000 face images of 126 subjects (70 men and 56 women) with different facial expressions, illumination conditions, and occlusions. The images were taken under controlled conditions but no restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed to participants. Each subject participated in two sessions, separated by two weeks (14 days) time. The original image resolution is 768x576 pixels. Some examples of face images from the AR face database are shown in Figure 2.5.

2.2 Review of Face Recognition under Occlusions

The classical methodology to address face recognition under occlusion is to find corruption-tolerant features or classifiers. Toward this goal, numerous previous works confirmed that locally emphasized algorithms are less sensitive to partial occlusions. Penev and Atick [122]



Figure 2.5: Example of occlusions in the AR face database.

proposed the local feature analysis (LFA) to extract local features by second order statistics. Martinez [107] proposed a probabilistic approach (AMM) which can compensate for partially occluded faces. Tan et al. [138] extended Martinez's work by using the self-organizing map (SOM) to learn the subspace instead of using the mixture of Gaussians. In [89], Jim et al. proposed a method named locally salient ICA (LS-ICA) which only employs locally salient information in constructing ICA basis. In [62], Fidler et al. presented a method which combines the reconstructive and discriminative models by constructing a basis containing the complete discriminative information. Park et al. [120] proposed to use a line feature based face attributed relational graph (ARG) model which encodes the whole geometric structure information and local features of a face. Zhang et al. [157] proposed a non-statistical face representation - Local Gabor Binary Pattern Histogram Sequence (LGBPHS) to exploit the multi-resolution and multi-orientation Gabor decomposition. In [84], Jia and Martinez proposed the use of partial support vector machines (PSVM) in scenarios where occlusions may occur in both the training and testing sets.

More recently, facial occlusion handling under the sparse representation based classification (SRC) [146] framework has demonstrated impressive performances in face recognition with occlusions. The idea of using SRC for occluded face recognition is first introduced by Wright et al. [146], where an occluded face is represented as a linear combination of the whole face gallery added by a vector of errors (occlusion) in the pixel-level and the classification is achieved by $l1$ minimization. Zhou et al. [160] extend [146] by including a Markov Random Fields (MRF) model to enforce spatial continuity for the additive error vector to address contiguous occlusions. In [150], Yang and Zhang applied compressible image Gabor features instead of original image pixels as the feature vector used in SRC to reduce computations in the presence of occlusions. Liao and Jain [96] incorporated the SIFT descriptor into the SRC framework to achieve alignment free identification. Yang et al. [151] proposed a Robust Sparse Coding (RSC) method which seeks the maximum likelihood estimation (MLE) solution of the sparse coding problem for non-Gaussian/Laplacian occlusions in an iterative manner. Even though the SRC based methods achieve significant identification results on occluded faces from standard face databases (i.e. AR face database [106]), the prerequisite of

those methods relies on the large number of training samples of each identity with sufficient variations. But in many practical face recognition scenarios, the training samples of each subject are often insufficient (the "curse of the dimensionality" [52] problem, in the extreme case only one template face per subject is available).

Lately, a few works have revealed that prior knowledge of occlusions can significantly improve the accuracy of local-feature/local-component based face recognition. Rama et al. [127] empirically showed that prior knowledge about occlusion (manually annotated) can improve Eigenface in local patches. In [116], Oh et al. have proposed an algorithm based on local non-negative matrix factorization (LNMF) [94], named Selective LNMF (S-LNMF) that automatically detects the presence of occlusion in local patches; face matching is then performed by selecting LNMF representation in the non-occluded patches. Zhang et al. [156] proposed to use Kullback-Leibler divergence (KLD) to estimate the probability distribution of occlusions in the feature space, so as to improve the standard LGBPHS based method [157] for partially occluded face.

In this thesis, we will demonstrate that explicit occlusion analysis and processing can significantly improve face recognition under occlusion conditions. Even if locally emphasized algorithms (e.g. LGBPHS [157]) can already show robustness to partial occlusions, our experiments reveal that the prior information of the occlusion (obtained by our occlusion analysis manners) can further improve it. Because occlusion analysis we presented is an independent module from the face recognition part, and we adopt face recognition methods without a learning step (e.g. LBP, LGBPHS) (unlike like Eigenface [140], Fisherface [21] or sparse representation [146]), our approaches are not restrained by the number of training samples. The summary of of literature works in occluded face recognition is illustrated in Table 2.1.

2.3 State-of-the-Art Techniques

According to our review in Section 2.2, works in the literature can be categorized into three classes. In this section, we will look into more details of the state-of-the-art techniques by showing the representative algorithms in each category. Three different ways in the locality emphasized algorithms are reviewed in Section 2.3.1. The standard SRC method for face recognition is reviewed in Section 2.3.2. Finally we show how the previous works exploiting explicit occlusion analysis can improve face recognition in Section 2.3.3.

2.3.1 Locality Emphasized Algorithms

In this section, we show three different ways to enforcing the locality in face recognition to cope with effects of the occlusion problem. In AMM [107], the holistic face representation (Eigenface) is divided into several local components and thus partial occlusion can only affect part of those components. In LS-ICA [89], a face image is projected onto a set of local-

Table 2.1: Summary of literature works in occluded face recognition.

Category	Abbreviation	Full name / Brief description
Locality emphasized features/classifiers.	LFA [122]	Local feature analysis.
	AMM [107]	Gaussian mixture modelling of part-based Eigenface.
	SOM-AMM [138]	Self-Organizing map modelling of part-based Eigenface.
	LS-ICA [89]	Local salient - independent component analysis.
	RD-Subspace [62]	Combining reconstructive and discriminative subspace.
	ARG [120]	Attributed relational graph.
	LGBPHS [157]	Local Gabor binary patterns histogram sequence.
	PSVM [84]	Partial support vector machines.
SRC based methods.	SRC [146]	Sparse representation based classification.
	MRF-SRC [160]	Markov random field to enforce spatial continuity in SRC.
	Gabor-SRC [150]	Compressible Gabor feature used in SRC.
	SIFT-SRC [96]	SIFT feature used in SRC.
	RSC [151]	Robust sparse coding.
Explicit occlusion analysis based methods.	Part-PCA [127]	Occlusion analysis + part-based Eigenface.
	S-LNMF [116]	Selective local non-negative matrix factorization.
	KLD-LGBP [156]	Local Gabor binary patterns based on Kullback-leibler divergence.

preserving subspace basis for face representation. In LGBPHS [157], both texture descriptors Gabor wavelet and LBP are summarized locally to construct a robust face representation.

2.3.1.1 Dividing holistic representation into local components

The method AMM [107] demonstrates a way to divide holistic face representation into multiple local components to enhance its robustness to occlusion. The proposed method is based on the original Eigenface [140] (as we have reviewed in Section 2.1.1.1). Indifferent from Eigenface, in order to overcome the occlusion problem, each face image is divided into k local parts ($k = 6$ in [107]):

$$X_k = \{x_{1,k}, x_{2,k}, ..., x_{n,k}\} \tag{2.10}$$

where $x_{i,k}$ is the k-th local region of the ith class in the gallery. Subsequently, k eigenspaces E_k can be trained for each part from the gallery faces. All faces in the gallery are then projected onto the computed k eigenspaces and obtains the new representation $X'_k = \{x'_{1,k}, x'_{2,k}, ..., x'_{n,k}\}$. Because multiple gallery faces are available for each identity ($x_{i,k} = \{x_{i,1,k}, ..., x_{i,r,k}\}$), the new representation is not directly used for matching. Instead, a Gaussian mixture model (GMM) G_i is built for each identity. Considering the local components, for each person there are k GMMs $G_{i,k}$, with the associated mean $\mu_{i,k,g}$ and covariance matrix $\Sigma_{i,k,g}$, where g is the number of components in the GMM. The mixture of Gaussians can

be learned by the EM(Expectation-Maximization) algorithm [50, 108]. The idea of dividing holistic Eigenface into local components can be visually observed in Figure 2.6.

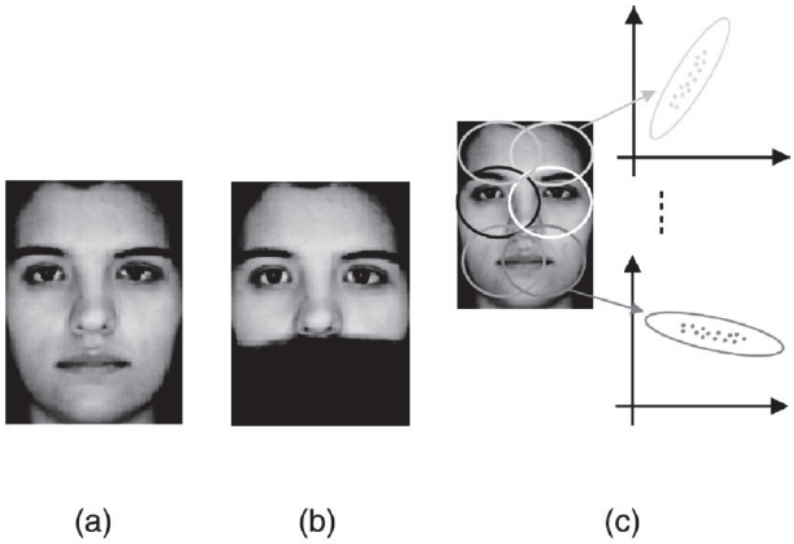


Figure 2.6: Illustration of the AMM algorithm: (a) a clean face; (b) an occluded face; (c) dividing the face image into 6 local components, and each component is independently modelled by a mixture of Gaussian model from a set of training samples. (Image excerpted from [107])

In the recognition part, when a test image t is to be recognized, it is first divided into the same k local parts, and each part t_k is projected onto the corresponding eigenspaces E_k to form the new representation t'_k. With the learned k GMMs, the probability of a local match of the given probe to the ith identity can be calculated as:

$$P_{i,k}^{local} = (t'_k - \mu_{i,k})\Sigma_{i,k}(t'_k - \mu_{i,k}) \tag{2.11}$$

all the local probabilities are then summed up for each match:

$$P_i = \sum_{k=1}^{6} P_{i,k}^{local} \tag{2.12}$$

so as to find the probe's ID:

$$ID = arg\max_i P_i \tag{2.13}$$

by the best fitting.

2.3.1.2 Local-preserving subspace projection

The reason why traditional subspace projection methods (such as Eigenface [140] and Fisherface [21], as we reviewed in Section 2.1.1) do not work in the case of occlusion is because

as a holistic representation each coefficient in the projected subspace is correlated with all original pixels, in other words changes of the original pixel values can affect all components for the face representation. LS-ICA [89] presents an approach which built a subspace with locality preserving basis. The projection onto these basis only correlates parts of the original pixels, and therefore can ameliorate the occlusion problem in face recognition. Similar approaches can be found in LFA [122] and LNMF [94].

ICA [79] is a well known dimensionality reduction tool who has already the locality property to some extent (see Figure 2.7 for a visual example). LS-ICA [89] improves ICA by imposing additional localization constraint in the process of computing ICA basis. Let S be the vector of unknown signals and X is the observed mixtures (gallery faces). A is an unknown mixing matrix , the mixing model is $X = AS$. ICA estimates the independent source signals U by computing the separating matrix W as below:

$$U = WX = WAS \tag{2.14}$$

The kurtosis of U is computed and the separating matrix W is obtained by maximizing the kurtosis. LS-ICA imposes additional localization constraint in the process of the kurtosis maximization. The solution at each iteration step is weighted so that it becomes sparser by only emphasizing large pixel values. As illustrated in Figure 2.7, the basis of LS-ICA is sparser and more locally distributed than the ones of ICA. Such locality is behind the robustness of LS-ICA to partial occlusions.

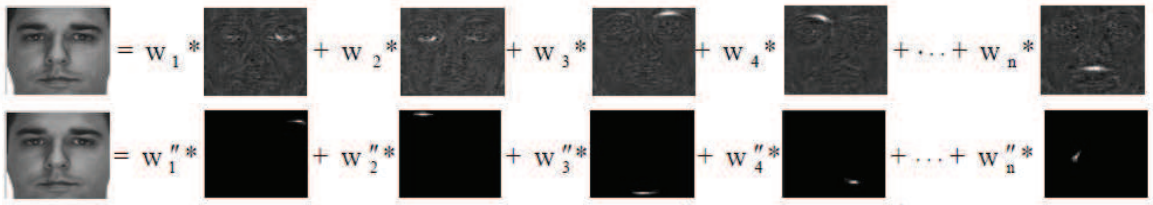


Figure 2.7: Illustration of the ICA basis (upper) and the LS-ICA basis (lower): a face image can be represented as a linear combination of such basis. It is clear that the LS-ICA basis is more locally distributed. (Image excerpted from [89])

2.3.1.3 Local texture descriptor

Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [157] is a variant of LBP based face representation by combining two local texture descriptors Gabor wavelet filtering and LBP. The high order statistics of local texture descriptors have demonstrated good robustness to many facial variations, and in particular to the partial occlusion variation. As illustrated in Figure 2.8, the LGBPHS based face representation can be obtained by the procedures as following: (1) an input face is normalized and filtered by a set of Gabor wavelets to output multiple Gabor Magnitude Pictures (GMPs); (2) Local Binary Patterns are summarized over all GMPs to produce a set of Local Gabor Binary Pattern (LGBP) maps; (3) each LGBP map is divided into k non-overlapping blocks and histograms are computed from these local blocks;

(4) finally all histograms from all local regions of all LGBP maps are concatenated together to form the final face representation. To yield the LGBP feature, following Gabor wavelets are used:

$$\psi_{\mu,\gamma}(z) = \frac{\|k_{\mu,\gamma}\|^2}{\delta^2} e^{(-\|k_{\mu,\gamma}\|^2\|z\|^2/2\delta^2)}[e^{ik_{\mu,\gamma}z} - e^{-\delta^2/2}] \tag{2.15}$$

where 40 Gabor wavelets are generated given five scales $\gamma \in [0, ..., 4]$ and eight orientations $\mu \in [0, ..., 7]$. Since the phase information of the Gabor filtering is time-varying, in LGBPHS, only the magnitude is used. Experiments in [157] showed that the LGBPHS face representation is more robust to partial occlusion in comparison to the original LBP face representation. Its robustness relies on the local patch based feature extraction and the multi-resolution and multi-orientation Gabor wavelet decomposition.

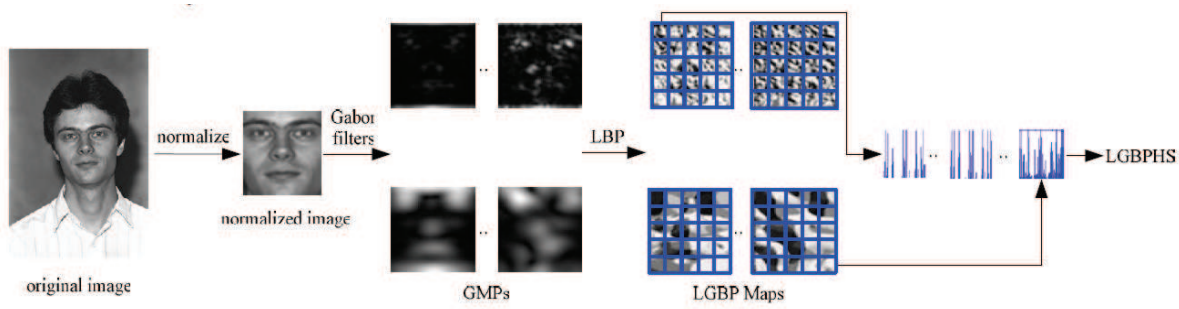


Figure 2.8: Illustration (Image excerpted from [89])

Locality emphasized algorithms can greatly ameliorate the occlusion problem based on the fact that changes in original pixels can only alter parts of the face representation. However, without explicit occlusion analysis and handling, information from the occluded part can still hinder the recognition process and therefore deteriorate the final results.

2.3.2 Sparse Representation based Classification (SRC)

Sparse representation based classification (SRC) [146] is perhaps one of the most impressive techniques for face recognition in recent years. In the SRC framework, face recognition is casted as penalizing the $l1$-norm of the coefficients in the linear combination of an over-complete face dictionary. Sparse representation based classification (SRC) has been demonstrated to be superior to nearest neighbor (NN) and nearest subspace (NS) based classifiers in various subspaces (e.g. random projection, PCA [140] ,LDA [21] or LPP [70]). When applied to face recognition, it can also be efficiently customized to handle errors due to occlusion and corruption. We will first review the regular SRC algorithm for face recognition without occlusion in Section 2.3.2.1, then the method to handle occlusion problem is reviewed in Section 2.3.2.2.

2.3.2.1 SRC for non-occluded faces

Given a training set A consists of the facial images from k classes, where $A = \{A_1, A_2, \ldots, A_k\}$. Giving sufficient training samples of class i, where $A_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,n_i}\} \in R^{m \times n_i}$, a test facial image $y \in R^m$ belongs to the same class could be well approximated by a linear combination of the training samples from A_i as:

$$y = \sum_{j=1}^{n_i} a_{i,j} v_{i,j} \tag{2.16}$$

Since A is the dictionary which includes all the training samples, where $A = \{v_{1,1}, v_{1,2}, \ldots, v_{k,n_i}\}$. Then Equation 2.16 can be rewritten in the form as below:

$$y = Ax_0 \in R^m \tag{2.17}$$

where $x_0 = \{0, \ldots, 0, a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}, 0 \ldots, 0\}^T$ is the coefficient vector in which most coefficients are zero except the ones associated with class i. Straightforwardly, if the correct coefficient vector x_0 is given, the identity of a given probe y is found.

The SRC algorithm relies on the fact that a valid test sample y can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others. In order to find the identity of y, SRC seeks the sparsest solution of Equation 2.17 via the following $l1$-minimization based approximation:

$$\hat{x}_1 = arg\min \|x\|_1 \qquad subject\ to\ \|Ax - y\|_2 \leq \epsilon \tag{2.18}$$

given the resolved sparse coefficient vector, identification can be achieved by finding the least residual among classes. The residual for class i can be computed as:

$$r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2 \tag{2.19}$$

for $i = 1, \ldots, k$, where δ_i is the characteristic function which selects the coefficients associated with the ith class.

The correctness of resolving equation 2.18 is based on the prerequisite that the given face dictionary is over-complete (where the number of training samples N is way larger than the feature length M for $M \ll N$, so that the linear system in 2.18 is underdetermined and therefore has multiple solutions in which the sparsest one can be selected.). However, in practice, a face image has very high dimension (16384 for a 128×128 face image) and limited number of faces are given. Because increasing the number of training samples is a very difficult task, reducing the dimension of features is therefore the alternative in the SRC algorithm. Several subspace methods can be applied prior to the $l1$-minimization (e.g. random projection, PCA [140], LDA [21], LPP [70]) :

$$\tilde{y} \doteq Ry = RAx_0 \in R^d \tag{2.20}$$

where R is the dimensionality reduction function, and $\tilde{y}$ is the approximated representation in the subspace. Experiments in [146, 149] have demonstrated the effectiveness of above approach.

2.3.2.2 SRC for occluded faces

When the occlusion problem occurs, original SRC method we reviewed in Section 2.3.2.1 fails because the dimensionality tools used are all "holistic". To overcome the occlusion problem in SRC based face recognition, Wright et al. [146] argued that the original pixel is the most "local" feature in comparison to feature extraction by any transformation and proposed to model the occlusion in the original pixels' space as a simple linear model:

$$y = y_0 + e_0 = Ax_0 + e_0 \tag{2.21}$$

where $e_0 \in R^m$ is the error vector (caused by occlusion) and the goal for face recognition is to recover the correct x_0. To achieve this goal, a "cross-and-bouquet" model [145] is proposed which co-optimizes the sparse coefficient vector x_0 and the error vector e_0 by adding an identity matrix (I, the "cross") into the original $l1$-minimization formulation as below:

$$y = [A, I][\begin{array}{c} x_0 \\ e_0 \end{array}] \doteq Bw_0 \tag{2.22}$$

where $B = [A, I] \in R^{m \times (n+m)}$ (m is the number of pixels n is the number of training samples), so the system $y = Bw$ is always underdetermined and has multiple solutions for w. Then the recovery of x_0 is caseted to solve the following extended $l1$-minimization problem:

$$\hat{w}_1 = arg\min \|w\|_1 \qquad subject\ to\ \|Bw - y\|_2 \leq \epsilon \tag{2.23}$$

with the recovered $\hat{w}_1 = [\begin{array}{c} x_0 \\ e_0 \end{array}]$ in the sense of $l1$-minimization, the identity of a probe can be found by the residual minimization method we reviewed in Section 2.3.2.1. This approach is equivalent to the visual representation illustrated in Figure 2.9.

Even though SRC demonstrates impressive recognition capability as well as robustness to occlusion, it suffers from the "curse of dimensionality" problem. In many practical scenarios, the number of templates (of each identity) is insufficient to support the recovery of correct sparse coefficients. We will show by experiments in Chapter 3, on a limited dataset, SRC based algorithms yields much inferior results (w.r.t. a simple LBP based method).

2.3.3 Occlusion Analysis + Face Recognition

According to our review in Section 2.2, only Part-PCA [127], S-LNMF [116], and KLD-LGBP [156] incorporates explicit occlusion analysis into the face recognition process. Because Part-

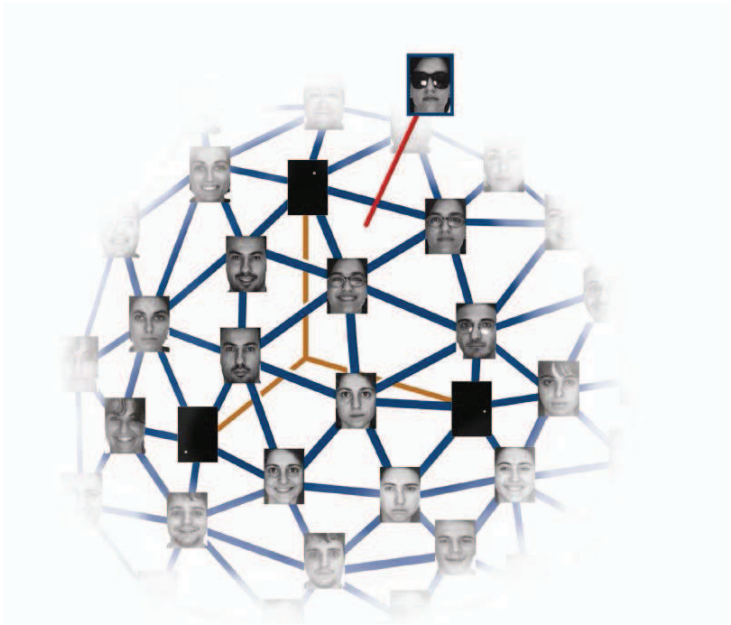


Figure 2.9: Illustration of the SRC algorithm in case of occlusion occurs: an (well aligned) occluded face can be expressed by a linear combination of faces in the gallery plus the measurement of erroneous on each pixel location. (Image excerpted from [146])

PCA uses manual annotation for occlusion detection to improve basic Eigenface algorithm, we do not consider it as a practical system in this thesis. S-LNMF conducts a binary occlusion detection in local patches to improve Local Non-negative Matrix Factorization (LNMF) [94] based face recognition, whereas KLD-LGBP adopts a fuzzy occlusion detection scheme based on Kullback–Leibler divergence to improve LGBP based face recognition. Both methods will be reviewed in Section 2.3.3.1 and Section 2.3.3.2, respectively.

2.3.3.1 S-LNMF

S-LNMF performs occlusion detection explicitly via 1-NN classifier in local regions of a face. The recognition process is performed using the selected LNMF bases from the non-occluded patches. For occlusion detection, a probe face is firstly divided into k non-overlapping local patches (in [116] $k = 6$). Subsequently, k PCA subspaces are trained individually for the k patches, using the all faces (non-occluded) from the gallery. The PCA coefficients of all gallery faces correspond to those k PCA coefficients can be obtained:

$$\Gamma_{i,k} = E_k^T(X_{i,k} - \mu_k) \quad i = 1, 2, ..., N \; and \; k = 1, 2, ..., 6 \tag{2.24}$$

where $X_{i,k}$ is the kth patch of the ith image in the gallery, μ_k is the mean patch of the kth patch and E_k is the trained PCA projection for the kth patch. Similarly, the PCA coefficients for a probe face x can be obtained:

$$\gamma_k = E_k^T(x_k - \mu_k) \quad k = 1, 2, ..., 6 \tag{2.25}$$

The occlusion detection is then achieved by comparing the patch of the probe γ_k with the set of patches of all gallery faces Γ_k in the PCA subspace. In [116], a supervised 1-nearest neighbour threshold classifier is used for occlusion detection (see Algorithm 1).

Algorithm 1 Supervised 1-NN threshold classifier

```
Find the nearest neighbour of the test data;
if the nearest neighbour is an outlier: then
    test data is assigned to the outlier class.
else
    if distance < threshold: then
        test data is assigned to the target class.
    else
        test data is assigned to the outlier class.
    end if
end if
```

With the binary information obtained which indicates if a local patch is occluded, the LNMF bases from the non-occluded patches can be selected (Figure 2.10 shows an example of LNMF bases correspond to a non-occluded local patch). The selection is achieved via the measurement of following occlusion energy:

$$E^i_{Occlusion} = \frac{\sum_{x,y \in W} I_i^2(x,y)}{\sum_{x=1}^{C} \sum_{y=1}^{R} I_i^2(x,y)}, \quad i = 1, 2, ..., N \tag{2.26}$$

where $C \times R$ is the image size, $I_i(x,y)$ is the ith LNMF base at pixel location (x,y), W is the detected occluded region, and N is the number of bases. The occlusion energy $E^i_{Occlusion}$ indicates the ratio of LNMF energy of base i contained in the occluded region to the whole LNMF energy of base i. If this energy is higher than a threshold, it is discarded from the LNMF basis for face representation. Experiments in [116] demonstrated that by selecting the LNMF bases from the non-occluded patches only can significantly improve recognition result in comparison to the original LNMF and a number of locality emphasized algorithms including Part-PCA [127], LFA [122] and AMM [107].

2.3.3.2 KLD-LGBP

LGBPHS [157] is already a robust face recognition method thanks to its locality property and the Gabor and LBP based texture descriptors. KLD-LGBP [156] further improves LGBPHS by explicit occlusion detection based on Kullback–Leibler divergence (KLD) and integration of occlusion information in the matching process by a weighting scheme to weaken the significance of occlusion in the classification. Similar to LGBPHS, the LGBP operator is firstly applied to the face image I:

$$LGBP^{\mu,\gamma}_{P,Q}(I(x,y)) = \sum_{p=0}^{P-1} s(G(I(x,y)_p)^{\mu,\gamma} - G(I(x,y)_c)^{\mu,\gamma}) 2^P \tag{2.27}$$

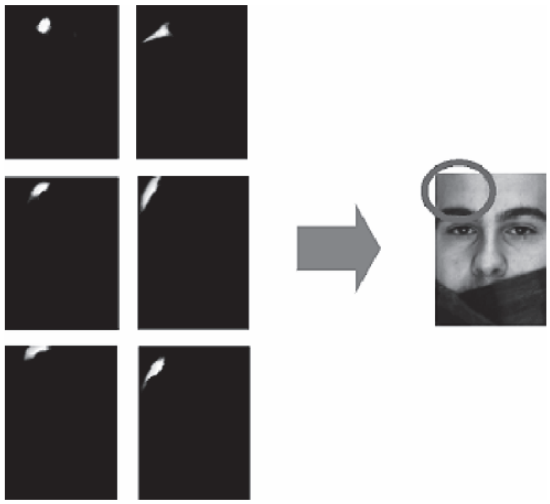


Figure 2.10: Illustration of LBMF bases correspond to a non-occluded local patch (Image excerpted from [116])

where $G(I(x,y))^{\mu,\gamma} = I(x,y) * \psi_{\mu,\gamma}$ denotes the Gabor wavelet filtering, and $\psi_{\mu,\gamma}$ is the Gabor wavelet with μ scales and γ orientations. Then the binary code can be locally calculated by differentiating the central pixel $I(x,y)_c$ from its P equally spaced (at radius Q) neighbours $I(x,y)_p$. Following the procedures we reviewed in Section 2.3.1.3, its LGBPHS representation (global histogram) is then formed.

To estimate the occlusion probability in a local region r of a face image, the mean histogram $\overline{h}_{\mu,\gamma,r} = \frac{1}{N}\sum_{n=0}^{N-1} h^n_{\mu,\gamma,r}$ of the gallery set (N faces without occlusion) is computed as the "normal" distribution. The occlusion probability of a probe local histogram $h'_{\mu,\gamma,r}$ is estimated as the KL-divergence [91] from $\overline{h}_{\mu,\gamma,r}$:

$$P^{RGB}_{\mu,\gamma,r} = \| \sum_{i=0}^{L-1} h'_{\mu,\gamma,r}(i) \log \frac{h'_{\mu,\gamma,r}(i)}{\overline{h}_{\mu,\gamma,r}(i)} \| \tag{2.28}$$

where $\|\cdot\|$ indicates the normalization that filters small deviations, L is the histogram length. By averaging the occlusion probability from all Gabor wavelet decompositions (with different μ and γ), visualization of the estimated occlusion probability for each patch can be viewed in Figure 2.11.

Instead of direct matching of LGBPHS between a probe face and the gallery faces, KLD-LGBP associates a set of weights (which are inversely proportional to the occlusion probabilities) to local histograms for the computing of global similarity between two faces as below:

$$S(H^1, H^2) = \sum_{\mu=0}^{4}\sum_{\gamma=0}^{7}\sum_{r=0}^{R-1}(1 - P^{RGB}_{\mu,\gamma,r}) dist(h^1_{\mu,\gamma,r}, h^2_{\mu,\gamma,r}) \tag{2.29}$$

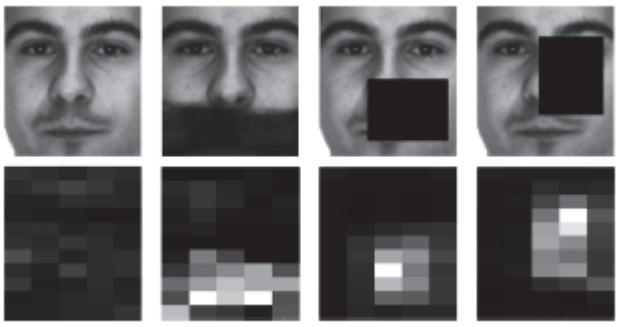


Figure 2.11: Illustration of the occlusion probability estimated using the KLD method (brighter block refers to higher occlusion probability). (Image excerpted from [156])

where $P^{RGB}_{\mu,\gamma,r}$ is the occlusion probability of (μ, γ, r)th local histogram. Experiments in [156] show that KLD-LGBP outperforms the original LGBPHS with a significant margin. However, even though the occlusion weighting strategy can greatly ameliorate the effect of occlusion in the recognition process, we will show in Chapter 3 that the occlusion exclusion strategy (by setting the weight to either 0 or 1) can achieve better results since information from the occluded information can be completely discarded in recognition.

In this section, we have seen that explicit occlusion analysis can further improve locality emphasized algorithms. The works presented in this thesis follow the same philosophy but outperform the state-of-the-art techniques.

2.4 Conclusions

In this chapter, we have given a complete study of face recognition under occlusion conditions in the literature. Starting by reviewing the basic tools and concepts in face biometrics, we presented why the standard techniques such as (Eigenface, Fisherface or LBP) fail when the occlusion problem occurs, followed by the face recognition evaluation metrics (in which our choice of evaluation metric is justified), general face recognition challenges, and the overview of AR face database. Then a complete review of works in the literature is summarized. According to the different manners the literature works approached, we categorize them into three different classes: the locality emphasized features/classifiers, the sparse representation for classification (SRC) based methods and the explicit occlusion analysis based methods. For each of these classes, we reviewed in details the representative state-of-the-art techniques to understand the advantages and the limitations of their algorithm design. At first the locality emphasized algorithms are summarized in three different manners, including: dividing holistic representation into local components, local-preserving subspace projection and local texture descriptor. Then the review of the standard SRC method for face recognition is given. At last, we show the previous works exploit explicit occlusion analysis, which is comprised of S-LNMF and KLD-LGBP.

Review of the state-of-the-art enables us to better understand the occlusion problem in face recognition. In the next chapters, we will present our solutions to the traditional occlusion problems (such as sunglasses and scarf) which outperform the state-of-the-art, as well as identify and address a number of new occlusion challenges (sparse occlusion, dynamic occlusion etc.) which are not considered in the literature. In addition, the emerging new sensor such as Kinect is also exploited, which depicts a promising perspective of handling the occlusion challenge for face recognition.

Chapter 3

Classical Occlusion Handling for Robust Face Recognition

3.1 Introduction

Facial occlusions, due for example to sunglasses, hats, scarf, beards etc., can significantly deteriorate the performance of any face recognition system. Unfortunately, the presence of facial occlusions is quite common in real-world applications especially when the individuals are not cooperative with the system such as in video surveillance scenarios. While there has been an enormous amount of research on face recognition under pose/illumination changes and image degradations, problems caused by occlusions are mostly overlooked. In this chapter, we will present solutions to address the most common facial occlusion caused by typical facial accessories (for example scarf and sunglasses) to achieve robust face recognition.

Facial occlusions may occur for several intentional or undeliberate reasons (see Figure 3.1). For example, facial accessories like sunglasses, scarf, facial make-up and hat/cap are quite common in daily life. Medical mask, hard hat and helmet are required in many restricted environments (e.g. hospital, construction areas). Some other people do wear veils for religious convictions or cultural habits. In addition, facial occlusions are often related to several severe security issues. Football hooligans and ATM criminals tend to wear scarf and/or sunglasses to prevent their faces from being recognized. Bank robbers and shop thieves usually wear a cap when entering in places where they commit illegal actions.

In the face recognition literature, typical occlusion problems are casted as how to correctly identify a person when an external object (typically a scarf or sunglasses) partially occludes the face appearance based on a single intensity image. Almost all state-of-the-art works devise their algorithms based on the AR face database, where the scarf and sunglasses problems are considered in face recognition. In comparison to the more advanced (and thus relatively uncommon) occlusion scenarios we will consider in Chapter 4, we regard the sunglasses and scarf problems as the classical facial occlusion problems. In this chapter, we elaborate two



(a) (b)

Figure 3.1: Illustration of different types of facial occlusions: (a) ordinary facial occlusions in daily life; (b) facial occlusions related to severe security issues (ATM crimes, football hooligans etc.).

proposed algorithms to handle the classical facial occlusion problem in face recognition, and reveal the advantages of the proposed methods over the state-of-the-art works supported by experimental analysis.

Because partial occlusions can greatly change the original appearance of a face image, it can significantly deteriorate performances of holistic representation based face recognition systems (such as [17, 21, 140], since the face representations are largely altered accordingly). To control partial occlusion is a critical issue to achieve robust face recognition. Most of the literature works [62, 84, 89, 96, 107, 120, 122, 138, 146, 150, 151, 157, 160] focus on finding corruption-tolerant features or classifiers to reduce the effect of partial occlusions in face representation. However, information from the occluded parts can still hinder the recognition performance. Recently, researchers [113, 116, 127, 156] demonstrated that prior knowledge about the occlusion (e.g. type, location, size) can be used to exclude the information from occluded parts, so as to greatly improve the recognition rate. Hence, explicit occlusion analysis is an important step in occlusion-robust face recognition.

As depicted in Figure 3.2, the philosophy of our approach is to apply explicit occlusion analysis to exclude the occluded part in local feature based face representations, so as to attain occlusion-robust face recognition. It takes both advantages of information-selectivity from occlusion exclusion and the robustness from local features based face representation. The proposed occlusion analysis approach consists of two main parts, namely the occlusion detection part and occlusion segmentation part. In the occlusion detection part, the presence of occlusion can be discovered in local patches, which is achieved by Gabor wavelet filtering, PCA for dimensionality reduction and SVM based classification. In the occlusion segmentation part, we adopt a generalized Potts model-Markov random field (GPM-MRF) to recover the presence of occlusion in the pixel level. Using the patch based occlusion detection, information from the non-occluded local patches can be preserved for later recognition, whereas the non-occluded pixels can be preserved by adding the occlusion segmentation procedure.



Figure 3.2: Illustration of different types of facial occlusions: (a) ordinary facial occlusions in daily life; (b) facial occlusions related to severe security issues (ATM crimes, football hooligans etc.).

To justify our philosophy, the proposed occlusion analysis is in cooperated into two face recognition algorithms. In the first algorithm, we use the patch-based occlusion detection to select non-occluded blocks to improve basic LBP based face recognition. Experiments verify our assumption that by excluding the occlusion information can significantly improve face recognition under occlusion conditions. We then devise the second algorithm, which is called occlusion assited LGBPHS (OA-LGBPHS). In OA-LGBPHS, both patch-based occlusion detection and pixel-based occlusion segmentation are applied to improve LGBP based face recognition. The proposed algorithm competes a number of state-of-the-art algorithms [17, 140, 151, 156, 157] in various testing scenarios, yielding significant results.

The rest of this chapter is structured as follows. First, the proposed occlusion analysis methods (consisting of the occlusion detection part and occlusion segmentation part) are detailed in Section 3.2. Then, how to exploit the proposed occlusion analysis to improve LBP [17] and LGBP [157] based face recognition are described in Section 3.3.1 and Section 3.3.2 respectively. Finally, we draw conclusions in Section 3.4.

3.2 Occlusion Analysis

Integrating occlusion analysis into face recognition is a relatively new topic in face recognition. In [116], Oh et al. proposed to conduct binary occlusion detection in local patches based on PCA face subspace; then the face part from the non-occluded patches can be used for face recognition. In [156], a fuzzy occlusion detection scheme based on Kullback–Leibler divergence is performed in the LGBP feature space in local patches; then the estimated occlusion probability can be associated to the face matching so as to weaken the significance of the occluded parts in face recognition. We consider the first case as occlusion exclusion and the later one as occlusion weighting (note that occlusion exclusion can be regarded as a special case of occlusion weighting, where the weights are either 0 or 1).

Similar to [116], we propose to conduct occlusion exclusion in face recognition rather than occlusion weighting, because occlusion weighting still preserve some information from the occluded region which can be thoroughly discarded in occlusion exclusion. Experiments in Section 3.3.2.2 also confirms our choice. In this section, we propose an occlusion detection method in local patches based on Gabor, PCA and SVM. On top of the results from occlu-

sion detection in local patches, an occlusion segmentation method based on GPM-MRF is proposed in order to extract more precise occlusion information in pixels. Details of the proposed occlusion analysis approaches are given in the following sections.

3.2.1 Occlusion Detection in Local Patches

Figure 3.3 depicts the procedure of the proposed occlusion detection algorithm. We first divide the face image into different facial components. The number and the shape of the components are determined by the nature of the occlusion. Here we focus on the classical occlusion problems caused by scarf and sunglasses, we accordingly divide the face image into two equal components as shown in Figure 3.3. The upper part is used for analysing the presence of sunglasses while the lower part is used for detecting scarf.

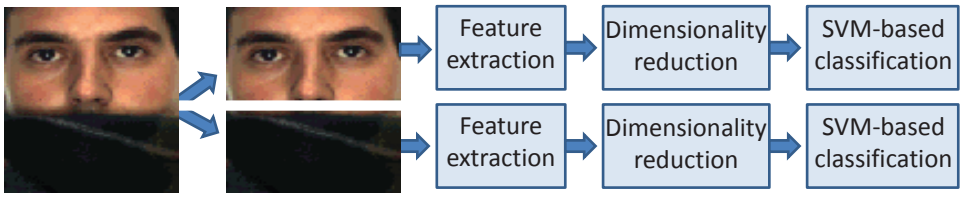


Figure 3.3: Overview of occlusion detection in local patches.

Once the face is divided into facial components, Gabor wavelet features are extracted from each component. The feature vector is then subject to dimensionality reduction. The result is fed to an SVM classifier for determining whether an occlusion is present or not in each facial component. Based on the observation in our experiments [112], in comparison to other methods that exploit information on facial features (e.g. skin color, mouth etc.) [38, 83, 99, 100, 129], our occlusion detection method is more robust against texture variations and it also tolerates better image degradation. Details of each step are described in the following sections.

3.2.1.1 Gabor wavelet based feature extraction

Gabor wavelets are used for extracting features from the potentially occluded regions. The choice of using Gabor wavelets is motivated by their biological relevance, discriminative power and computational properties. A Gabor wavelet consists of a complex sinusoidal carrier and a Gaussian envelope which can be written as:

$$\psi_{\mu,\gamma}(z) = \frac{\|k_{\mu,\gamma}\|^2}{\delta^2} e^{(-\|k_{\mu,\gamma}\|^2\|z\|^2/2\delta^2)}[e^{ik_{\mu,\gamma}z} - e^{-\delta^2/2}] \tag{3.1}$$

where μ and γ are the orientation and scale of the Gabor kernels, $z = (P, Q)$ is the size of the kernel window, $\|\cdot\|$ denotes the norm operator, $k_{\mu,\gamma} = k_\gamma e^{i\phi_\mu}$ is a wave vector, where

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

In our system, we set $z = (20, 20)$, $\delta = 2\pi$, $k_{max} = \pi/2$ and $f = \sqrt{2}$ as also suggested in [157]. Five scales $\gamma \in [0, ..., 4]$ and eight orientations $\mu \in [0, ..., 7]$ are selected to extract the Gabor features. In total, 40 Gabor wavelets are generated. Figure 3.4 shows the real part of the Gabor kernels and the corresponding magnitudes in 5 scales. In the figure, the desirable properties of spatial frequency, spatial locality and orientation selectivity are clearly shown.

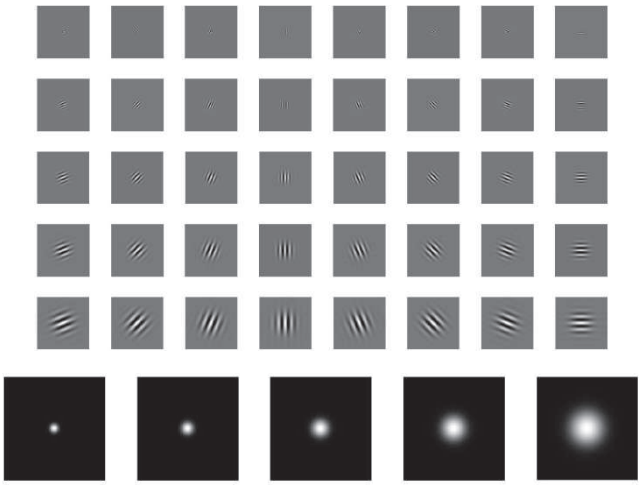


Figure 3.4: Real part of the 40 Gabor wavelets and their magnitudes in five scales.

Once the Gabor wavelets are generated, feature extraction is performed by convolving the wavelets with the face image I:

$$C_{\mu,\gamma}(x, y) = I(x, y) * \psi_{\mu,\gamma}(z) \tag{3.2}$$

Because the phase information of this transform is time varying, we only explore the magnitude information. The computed Gabor magnitude pictures (GMPs) thus form a set $\Omega = \{C_{\mu,\gamma}, \mu \in [0, 7], \gamma \in [0, 4]\}$, in which an augmented feature vector is constructed by concatenating all the GMPs. The obtained feature vector is down-sampled by a factor λ (here $\lambda = 5$) for further processing. Figure 3.5 shows an example image (the lower part of a clean face) and its corresponding GMPs using the above mechanism. In the figure, the filtered images exhibit the properties of the original image in different scales and orientations corresponding to the Gabor wavelets shown in Figure 3.4. Note that GMPs are not only used in occlusion detection, but also used to compute the face representation selective LGBPHS which will be introduced in Section 3.3.2.

3.2.1.2 Dimensionality reduction using PCA

Because the size of extracted Gabor feature is rather big, in order to reduce the dimension of the feature vectors while preserving its discriminative power, we apply principal component



Figure 3.5: An example image and the 40 GMPs filtered by Gabor Wavelets.

analysis (PCA) to maximize the variance in the projected subspace for the Gabor features. To compute the PCA subspace, we consider a training dataset consisting of feature vectors from both occluded and non-occluded image patches. Let us denote the feature vectors from the non-occluded patches as X^c and the feature vectors from the occluded patches as X^s. The training dataset S can be formed as: $S = \{X_1^c, X_2^c, ..., X_{M/2}^c, X_{M/2+1}^s, ..., X_M^2\}$, where M is the size of the training dataset. The eigenvectors associated with the k largest eigenvalues of $(S - \bar{S})(S - \bar{S})^T$ (the covariance matrix of S) are thus computed to describe the eigenspace. The Gabor wavelet based features are then projected onto the computed eigenspace for dimensionality reduction. Figure 3.6 shows the distribution of features in the eigenspace (projected onto first 3 eigenvectors). It is clear that the features from the two classes are roughly separated into two clusters in the eigenspace.

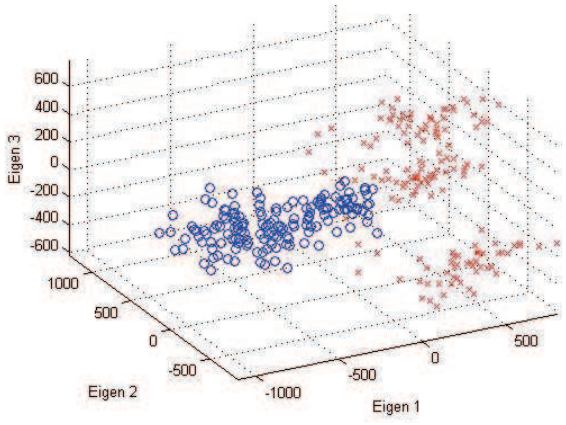


Figure 3.6: Illustration of the distributions of 150 occluded (red crosses) and 150 non-occluded faces (blue circles) in the eigenspace.

3.2.1.3 SVM based occlusion detection

Occlusion detection can be casted as a two-class classification problem. Since non-linear support vector machines (SVM) are proven to be a powerful tool for discriminating 2 classes of high dimensional data, we adopted then a non-linear SVM classifier for occlusion detection. Let us consider a training set consisting of N pairs $\{x_i, y_i\}_{i=1}^{N}$, where x_i refers to a reduced feature vector of a facial component i, and $y_i \in \{-1, 1\}$ is the label which indicates if the sample x_i is occluded or not. SVM finds the optimal separating hyper-plane $\{\alpha_i, i \in [1, N]\}$ by solving a quadratic programming problem [46], and predict the label of an unknown face x by:

$$f(x) = sign(\sum_{j=1}^{N} \alpha_j y_j K(x_j, x) + b) \tag{3.3}$$

where $\{x_j, j \in [1, N]\}$ are the support vectors. Non-linear SVM applies kernels $K(x_i, x_j)$ to fit the maximum-margin hyper-plane in a transformed feature space. In our system, the Radial Basis Function (RBF) kernel is used. The implementation of the non-linear SVM is provided by LIBSVM [41].

3.2.1.4 Results

For occlusion detection, from the AR face database, we randomly selected 150 non-occluded faces, 150 faces occluded with scarf and 150 faces wearing sunglasses for training the PCA space and SVM. The upper parts of the faces with sunglasses are used to train the SVM-based sunglass detector while the lower parts of the faces with scarf are used to train the SVM-based scarf detector. The 150 non-occluded faces are used in the training of both classifiers.

240 non-occluded faces, 240 faces with scarf and 240 face with sunglasses are selected from AR face database to test the proposed algorithm. We show the detection rates on all 720 testing images. Table 3.1 illustrates the results as a confusion matrix. Note that only 2 images (faces with very bushy beard) from the non-occluded faces are wrongly classified as faces with scarf. The correctness of the proposed occlusion detection ensures the reliability to improve later face recognition steps.

Table 3.1: Results of occlusion detection

	no-occlusion	Scarf	Sunglass	Detection Rate
no-occlusion	**238**	2	0	99.17%
Scarf	0	**240**	0	100%
Sunglass	0	0	**240**	100%

To test the proposed occlusion detection algorithm against various occlusion textures, we built our own dataset for testing. The proposed dataset consists of 72 face images taken from 6 lab members in 2 sessions. In each session, there are 3 clean faces with different facial expressions (natural expression, smile and talking) and 3 scarf faces with different scarf appearances for each individual. We select the clean faces with different facial expressions

in order to increase the variation of each individual, so that the images are more suitable to simulate the real conditions from video surveillance system. In sum there are 36 clean faces and 36 scarf faces. Figure 3.7 shows the scarf faces extracted from the proposed dataset. It exhibits the large variation of occlusion appearance.



Figure 3.7: Examples of our own occlusion dataset with different scarf appearances.

Our method has been compared to the state-of-the-art works including: 1.Edge Detection + Geometric Analysis (EDGA) [100], 2. Gabor Band + Geometric Analysis (GBGA) [38], 3. Skin Color Ratio (SCR) [99], 4. PCA + SVM (PS) [129], 5. GM + PCA + SVM (GPS) [83]. The results of our experiments on the proposed dataset are shown in Table 3.2. From the results in the table we can observe that the proposed occlusion detection achieves perfect detection rate even if the occlusion appearance is complex, whereas the other methods all return inferior results.

Table 3.2: The results on the proposed dataset.

Method	FAR	FRR	Detection Rate
EDGA [100]	16.67%	0%	93.33%
GBGA [38]	8.33%	25%	81.67%
SCR [99]	29.17%	8.33%	83.33%
PS [129]	72.22%	16.67%	75%
GPS [83]	0%	5.55%	97.22%
The proposed method	0%	0%	100%

3.2.2 Occlusion Segmentation

Once the presence of occlusion is known in the patch-level by the proposed occlusion detection. One further step can be applied to better understand the occlusion structure on an occluded face. In this section, we introduce an occlusion segmentation algorithm given the prior information from the patch based occlusion detection. This allows us to identify the presence of occlusion in the pixel-level so as to preserve more face information for the later recognition task.

In order to efficiently exploit the information of facial occlusion for face recognition, we generate a binary mask β (1 for occluded pixels and 0 for non-occluded pixels) indicating the location of occluded pixels to facilitate later feature extraction and matching in the recognition phase. This mask generation process is called occlusion segmentation. To generate an accurate occlusion mask (which can remove the occluded part meanwhile preserving as much

as information from the non-occluded part), we adopt a generalized Potts model-Markov random field (GPM-MRF) [31] to enforce structural information (shape) of occlusion, so as to identify a given pixel is occluded or not.

Our occlusion segmentation can be formulated as a typical energy-minimization problem in computer vision. Let us consider the face image (consists of multiple facial patches) as an undirected adjacency graph $G = (V, E)$ where $V = \{v_i, i \in [1, N]\}$ denotes the set of N pixels (vertex) and E denotes the edges between neighbouring pixels. Given a set of observations $O = \{o_1, ..., o_N\}$ corresponding to the set of vertex V, we want to assign a label (occluded: 1, non-occluded: -1) to each vertex. We model the set of labels $L = \{l_i, i \in [1, N]\}$ (discrete random variables taking values in $\Lambda = \{-1, 1\}$) as a first-order Markov random field. The structural prior is incorporated into the MRF by a generalized Potts model. Then our goal is to find the label set $\hat{L}$ that maximizes the posterior probability $P(L|O)$, which can be achieved by the maximum a posteriori (MAP) estimation [66] that maximizes the joint probability $P(O, L)$, where:

$$P(O, L) = P(O|L)P(L) \tag{3.4}$$

$$= \frac{1}{Z}\exp^{(-\frac{U(L)}{T})} \tag{3.5}$$

where Z is the partition function and T is the temperature. $U(L)$ is the sum of potentials from all cliques $C = \{c_i, i \in [1, N]\}$, which can be written as:

$$U(L) = \sum_{c_i \in C} V_c(L) \tag{3.6}$$

$$= \sum_{l_i \in L} \Psi(o_i|l_i) + \omega \sum_{(l_i, l_j) \in E} \Phi(l_i, l_j) \tag{3.7}$$

where ω is a weighting parameter controlling the importance of MRF prior (the choice of ω is based on experiments on a validation set.). The unary potential Ψ is defined by the likelihood function:

$$\Psi(o_i|l_i) = -\ln p(o_i|l_i) \tag{3.8}$$

we approximate the occlusion likelihood ($l = 1$) as follows:

$$p(o|l = 1) = \begin{cases} \mathrm{e}^{-1} & \text{if } o > \tau \\ 1 - \mathrm{e}^{-1} & \text{else} \end{cases} \tag{3.9}$$

where $0 < \tau < 1$ and the face likelihood ($l = -1$) as a constant $c \in [0, 1]$:

$$p(o|l = -1) = c \tag{3.10}$$

Because we have already identified the type of occlusion (obtained by our occlusion detector), we can give a initial guess of observations O (the seed of occlusion mask, see Figure 3.8 (b)) to each type of occlusions. The structural information is enforced into this initial guess via the isotropic MRF prior $P(L)$, where the pairwise potential $\Phi(l_i, l_j)$ has the form of

generalized Potts model as defined in [31]:

$$\Phi(l_i, l_j) = u(i, j) \cdot (1 - \xi(l_i - l_j)) \tag{3.11}$$

where $\xi(\cdot)$ represents the unit impulse function, then $\Phi(l_i, l_j) = 2u(i, j)$ if i and j have different labels ($l_i \neq l_j$) and zero otherwise. The structural information $u(i, j)$ is obtained as the first-order derivative after a Gaussian filtering (with kernel size $(5, 5)$) from the original image. Note that maximizing the joint probability $P(O, L)$ is equivalent to minimizing the cliques potential $U(L)$, and this energy minimization problem can be solved exactly using graph cuts [30, 32, 90] in polynomial time. The obtained label set $\hat{L}$ (see Figure 3.8 (e)) is converted to the segmentation mask $\beta \in [0, 1]$ for later recognition task.

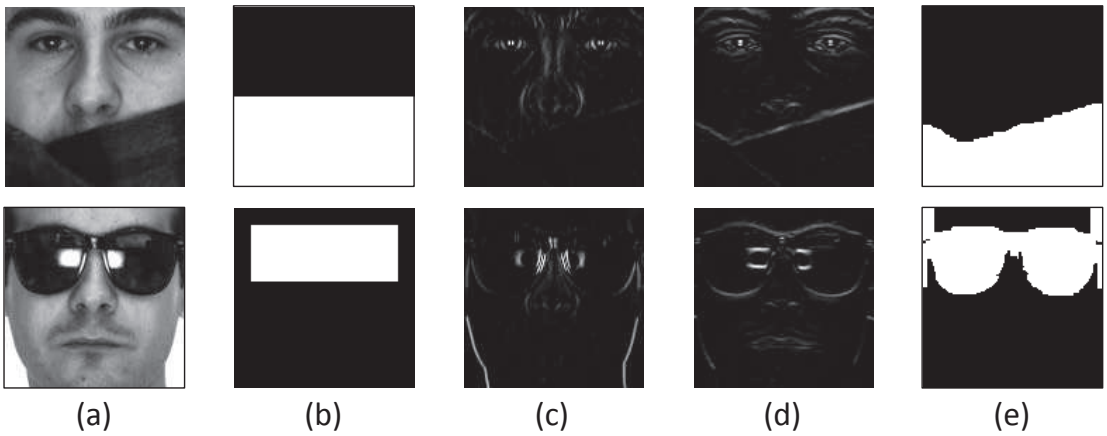


Figure 3.8: Illustration of our occlusion segmentation: (a) examples of faces occluded by scarf and sunglasses; (b) initial guess of the observation set according to the results from our occlusion detector; (c)(d) are the visualization of $u(i, j)$ in horizontal and vertical directions respectively; (e) the generated occlusion masks ($\omega = 150$).

3.3 Face Recognition

To demonstrate how facial occlusion analysis can be used to efficiently improve face recognition under occlusion conditions, two examples are illustrated in this section. In the first example, we demonstrate how the proposed occlusion detection (as described in Section 3.2.1) can be integrated to select the non-occluded blocks in LBP based face recognition [17], where significant improvements are observed. In the second example, both the proposed occlusion detection and occlusion segmentation (as described in Section 3.2.2) are incorporated into the LGBP based face recognition [157]. In comparison to the state-of-the-art approach KLD-LGBPHS [156] based on occlusion weighting, our approach based on occlusion exclusion demonstrates more powerful robustness to different partial occlusions. In addition, the proposed algorithm competes a number of state-of-the-art algorithms [17, 140, 151, 156, 157] (from different categories according to our review in Section 2.2) in various testing scenarios. The proposed two algorithms are described in the following sections, respectively.

3.3.1 Improving LBP based Face Recognition

Based on the occlusion detection scheme described in Section 3.2.1, we improve LBP based face recognition [17] in presence of occlusions such as sunglasses and scarf. The flowchart of the proposed approach is illustrated in Figure 3.9. As shown in the figure, during the training phase, local binary patterns (LBP) are used to efficiently represent the gallery images (face templates), thus obtaining an LBP feature space. The adopted LBP-based facial representation and the motivations behind it are described in more details in Section 3.3.1.1. Then, given a target (i.e. probe) face image (which can be occluded or not) to be recognized, its LBP representation is first computed. The probe image is then divided into a number of facial components for occlusion detection. Each component is individually analyzed by an occlusion detection module, which is described in Section 3.2.1. As a result, potential occluded facial components are identified. Then, the LBP features from only the non-occluded parts are selected and used for recognition. The recognition is performed by comparing the selected LBP features from the probe image against selected LBP features from the corresponding non-occluded components of the template images. The nearest neighbour (NN) classifier and Chi-square (χ^2) distance are adopted for the recognition.

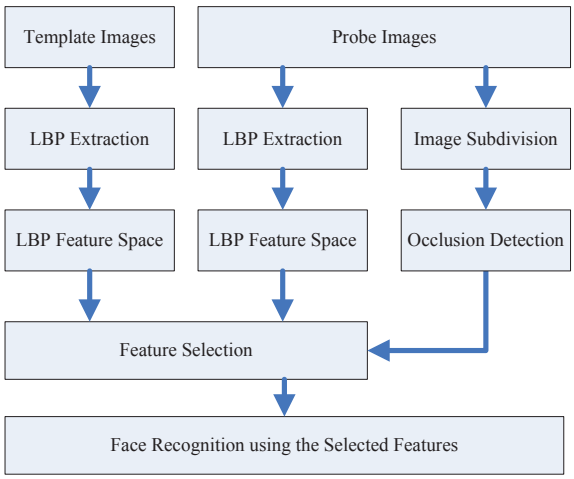


Figure 3.9: Flowchart of applying occlusion detection to improve LBP based face recognition under occlusion conditions.

3.3.1.1 Method

We adopt the local binary patterns for representing the non-occluded facial components and thus recognizing the face. In the LBP approach, a face image is divided into several regions from which the LBP features are extracted and concatenated into an enhanced feature histogram which is used as a face descriptor. LBP provides state-of-the-art results in representing and recognizing face patterns. The success of LBP in face description is due to the discriminative power and computational simplicity of the operator, and its robustness to monotonic gray scale changes caused by, for example, illumination variations. The use of

histograms as features also makes the LBP approach robust to face misalignment and pose variations.

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins including different types of curved edges, spots, flat areas, etc.

The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c,y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^P \tag{3.12}$$

where g_c corresponds to the gray value of the center pixel (x_c,y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities. For an efficient representation, facial images are first divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Figure 3.10 shows an example of an LBP based facial representation for the non-occluded region. In such a description, the face is represented in three different levels of locality: the LBP labels for the histogram contain information about the patterns on a pixel-level, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face. This locality property, in addition to the computational simplicity and tolerance against illumination changes, are behind our choice of adopting LBP for representing the non-occluded facial components for robust face recognition.

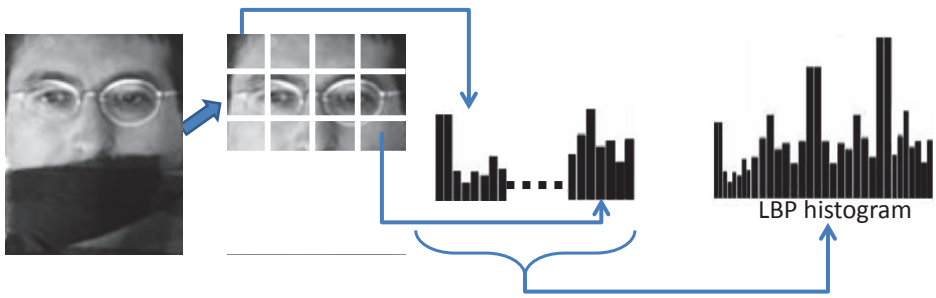


Figure 3.10: Example of extracting the LBP histogram from the non-occluded facial regions.

3.3.1.2 Results

For our experimental analysis, we considered the AR face database [106] which contains a large number of occluded faces. The original image resolution is 768×576 pixels. Using the eye coordinates, we cropped, normalized and down-sampled the original images into 128×128 pixels.

To conduct face recognition, the face images are divided into 64 blocks as shown in Figure 3.11. The size of each block is 16×16 pixels. The LBP histograms are extracted using the operator $LBP^{u2}_{8,2}$ (using only uniform patterns, 8 equally spaced pixels on a circle of radius 2) yielding in feature vector histograms of 3776 bins. In case of non-occluded faces, all these 3776 bins are used for matching using the Chi-square distance (χ^2). For occluded faces, however, the feature vector histograms are extracted only from the non-occluded parts as shown in the example in Figure 3.10. The occluded faces are thus represented with histograms of 1888 bins, corresponding to the 32 non-occluded blocks. This means that when a face is occluded by a scarf, the upper 32 blocks are selected, while the lower 32 blocks are used when the face is occluded by sunglasses. The faces which are occluded with both sunglasses and scarf are naturally rejected by the recognition system.



Figure 3.11: The face images are divided into 64 block for LBP based face recognition.

We first selected 240 non-occluded faces from session 1 of the AR database as the templates images. These non-occluded faces correspond to 80 subjects (40 males and 40 females), with 3 images per subject under neutral expression, smile and anger. For evaluation, we considered the corresponding 240 non-occluded faces from session 2, the 240 faces with sunglasses of session 1 and the 240 faces with scarf of session 1, under three different illuminations conditions.

Figure 3.12 shows the face recognition performance of our approach on three different test sets: non-occluded faces, face occluded with scarf and faces occluded with sunglasses. For comparison, we also report the results of eigenfaces (i.e. PCA) and basic LBP methods. Since PCA and LBP methods do not address occlusion detection, we implemented a third baseline approach for comparison. We thus combined our occlusion detection module with eigenfaces and call this approach FA-PCA (facial accessories robust PCA). In FA-PCA, three eigenspaces are computed during the training stage. The first one is computed using the whole face images, while the second and third eigenspaces are computed using the upper and lower facial regions, respectively. During the recognition phase, the non-occluded com-

ponents are projected into the corresponding eigenspace when partial occlusions are detected.

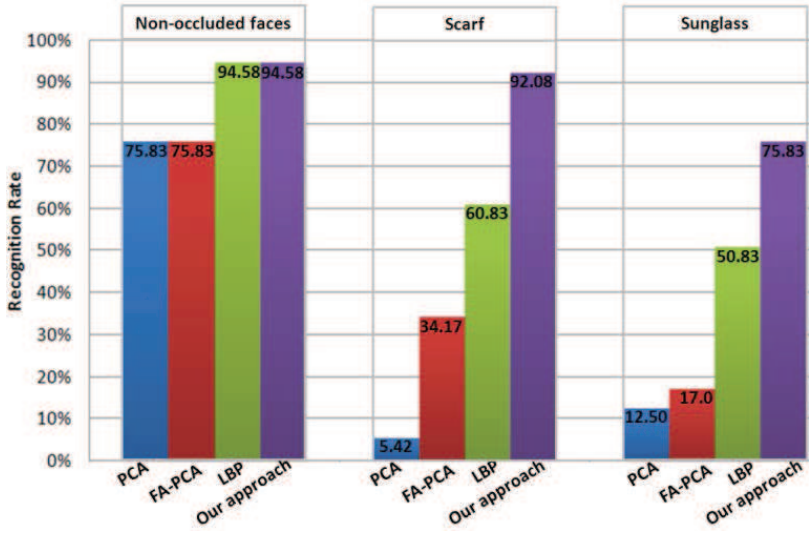


Figure 3.12: Recognition performance of different methods on three test sets: non-occluded faces, face occluded with scarf and faces occluded with sunglasses.

The results in Figure 3.12 clearly show that our proposed approach (it could be denoted by FA-LBP) significantly outperform all other methods. On the non-occluded faces, our approach and LBP yielded equal performance (94.83%) while the Eigenface method (with and without occlusion detection) yielded much lower performance (75.83%). On the test set of faces with scarves, our proposed approach gave best results (92.08%), followed by LBP (60.83%), and then by PCA-based methods (34.17% and 5.42%). Note that LBP performed quite well even under occlusion, thus confirming the earlier findings stating that local feature-based methods are more robust against occlusions than holistic methods. Comparing the results on the test sets of faces with sunglasses and scarves, we notice that most methods are more sensitive to sunglasses than to scarf. This is an interesting conclusion which is in agreement with the psychophysical findings indicating that the eye regions play the most important role in face recognition.

To get insight into our approach, we also considered the recent work of Oh et al. [116] for comparison. In [116], the authors proposed an approach called Selective LNMF (S-LNMF) that detects the presence of occlusions in pre-defined local patches, and then performs face recognition by selecting LNMF bases from the non-occluded patches (details of S-LNMF can be found in Section 2.3.3.1). This approach is closely related to our proposed method. The authors evaluated their method also on the AR face database. The reported results showed that the S-LNMF method outperformed several other techniques including PCA [140], LNMF [94], AMM [107] and LFA [122]. Moreover, the authors also studied the robustness of S-LNMF against drastic facial expression caused by screaming and against illumination changes caused by right-lighting, as well.

We compared the proposed approach against S-LNMF using similar protocol under the more challenging scenario in which the gallery face images are taken from Session 1 of AR database while the test sets are taken from Session 2. Note that the two sessions were taken at time interval of 14 days. The comparative results between our proposed approach and S-LNMF are illustrated in Table 3.3.

Table 3.3: OUR APPROACH VS. S-LNMF [116].

	Sunglass	Scarf	Scream	Right-Light
S-LNMF	49%	55%	27%	51%
Our approach	54.17%	81.25%	52.50%	86.25%

The results in Table 3.3 clearly show that our proposed approach outperforms S-LNMF method in all configurations assessing robustness against sunglasses, scarves, screaming and illumination changes. The robustness of our approach to illumination changes and drastic facial expression is brought by the use of local binary patterns, while the occlusion detection module significantly enhances the recognition of face occluded by sunglasses and scarves.

3.3.2 Improving LGBP based Face Recognition

In this section, we introduce a powerful feature descriptor to handle facial occlusions like sunglasses and scarf in face recognition, namely selective-LGBPHS. In comparison to the LBP based method we presented in Section 3.3.1, we improve face recognition in the following two aspects: (1) a more precise occlusion segmentation strategy is incorporated into the process, which can preserve more information from the non-occluded part, so as to improve the discriminative power of the face representation; (2) a more powerful face descriptor LGBP is used to replace LBP, which has been proven to be robust to pose and partial occlusion in the literature [156, 157].

A comprehensive overview of the proposed approach for face recognition is given in Figure 3.13. Given a target (i.e. probe) face image (which can be occluded or not) to be recognized, the possible presence of occlusion is first analysed. In addition to the occlusion detection procedure described in Section 3.2.1, an occlusion mask is generated by a more precise segmentation approach (see Section 3.2.2) to supervise the feature extraction and matching process. Based on the resulting occlusion mask, its LGBPHS representation is computed using the features extracted from the non-occluded region only, namely selective LGBPHS. The recognition is performed by comparing the selective LGBPHS from the probe image against selective LGBPHS from the template images using the same occlusion mask. The nearest neighbour (NN) classifier and Chi-square (χ^2) distance are adopted for the matching.

3.3.2.1 Method

The proposed face descriptor is based on the original LGBP [157] for face recognition. To construct the face representation, Gabor wavelet filtering is firstly applied. Because we also

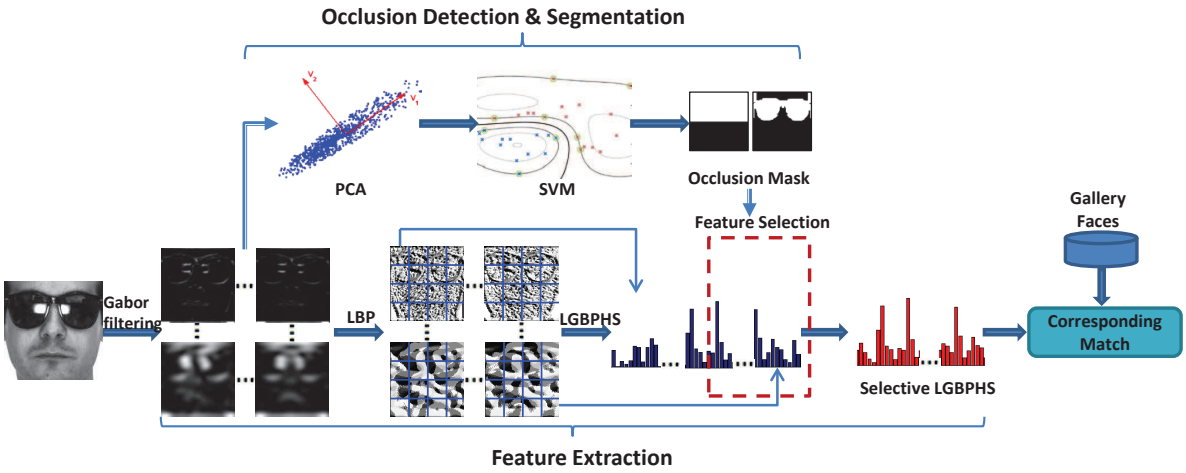


Figure 3.13: Illustration of the proposed selective-LGBPHS approach for face recognition in occluded conditions.

conduct Gabor filtering in the occlusion detection part, the computed GMPs (as described in Section 3.2.1.1) can be directly adopted to calculated the LGBP features.

Given a face image and its Gabor magnitudes pictures (GMPs) $\Omega = \{C_{\mu,\gamma}, \mu \in [0,7], \gamma \in [0,4]\}$ obtained by the method described in Section 3.2.1.1, the GMPs are further encoded by an LBP operator, resulting in a new feature description - local Gabor binary patterns (LGBP). The LBP operator forms labels for the image pixels by thresholding the 3×3 neighbourhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas, etc.

The calculation of LGBP codes are computed in a single scan through each GMP using the LBP operator. The value of the LGBP code of a pixel at position (x_c, y_c) of each scale μ and orientation γ of GMPs is given by:

$$LGBP_{P,R}^{\mu,\gamma} = \sum_{p=0}^{P-1} s(g_p^{\mu,\gamma} - g_c^{\mu,\gamma}) 2^P \tag{3.14}$$

where $g_c^{\mu,\gamma}$ corresponds to the intensity of the center pixel (x_c, y_c) in the GMP $C_{\mu,\gamma}$, $g_p^{\mu,\gamma}$ refers to the intensities of P equally spaced pixels on a circle of radius R. $\mu \times \gamma$ LGBP maps $\{G_{\mu,\gamma}\, \mu \in [0,7], \gamma \in [0,4]\}$ are thus generated via the above procedure. In order to exploit the spatial information, each LGBP map $G_{\mu,\gamma}$ is first divided into r local regions from which histograms are extracted and concatenated into an enhanced histogram $h_{\mu,\gamma} = (h_{\mu,\gamma,1}, ..., h_{\mu,\gamma,r})$. Then the LGBPHS (local Gabor binary pattens histogram sequence) is obtained by concatenating all enhanced histograms $H = (h_{0,0}, ..., h_{4,7})$.

The original LGBPHS summarizes the information from all pixels of a face image. Given an occlusion mask β (generated by the occlusion segmentation as described in Section 3.2.2),

our interest is to extract features from the non-occluded pixels only. Hence, we compute each bin h_i of the histogram representation using a masking strategy as below (please note that $\beta = 1$ for occluded pixels and $\beta = 0$ for non-occluded pixels)):

$$h_i = \sum_{x,y}(1 - \beta(x, y)) \cdot I\{f(x, y) = i\}, \forall i \in [0, 2^P - 1] \tag{3.15}$$

where i is the i-th LGBP code, h_i is the number of non-occluded pixels with code i, and:

$$I\{A\} = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases} \tag{3.16}$$

Then the local histograms ($h_{\mu,\gamma,r}$) extracted from all local regions of all GMPs are concatenated into the final representation, which is named as selective-LGBPHS. During matching, selective-LGBPHS are computed for both probe face and template faces, based on the occlusion mask generated from the probe.

In the selective-LGBPHS description, a face is represented in four different levels of locality: the LBP labels for the histogram contain information about the patterns on a pixel-level; the labels are summed over a small region to produce information on a regional-level; the regional histograms are concatenated to build a description of each GMP; finally histogram from all GMPs are concatenated to build a global description of the face. This locality property, in addition to the information selective capability, are behind the robustness (to facial occlusions) of the proposed descriptor.

3.3.2.2 Results

Similar to the experiments we conducted for LBP based approach (see Section 3.3.1), we cropped, normalized and down-sampled the original images into 128×128 pixels for faces from the AR face database [106]. 240 non-occluded faces are selected from session 1 as the templates images. These non-occluded faces correspond to 80 subjects (40 males and 40 females), with 3 images per subject under neutral expression, smile and anger. The evaluation set is consisting of 240 non-occluded faces from session 2, the 240 faces with sunglasses of session 1 and the 240 faces with scarf of session 1, under three different illuminations conditions.

For face recognition, the face images are then divided into 64 blocks as shown in Figure 3.11. The size of each block is 16×16 pixels. The selective LGBPHS are extracted using the operator $LBP^{u2}_{8,2}$ (using only uniform patterns, 8 equally spaced pixels on a circle of radius 2) on the 40 GMPs, yielding in feature histograms of 151040 bins.

Figure 3.14 shows the face recognition performance of our approach on three different test sets: clean (non-occluded) faces, faces occluded with scarf and faces occluded with sunglasses. For comparison, we also report results of the state-of-the-art algorithms (for the name abbreviations please refer to Table 2.1) for both standard face recognition and occluded

face recognition. Eigenfaces [140] (i.e. PCA), LBP [17] are among the most popular algorithms for standard face recognition. We also tested the approaches which incorporate our occlusion analysis (OA) with the standard Eigenface and LBP, namely OA-PCA and OA-LBP [113]. Similarly, we denote the proposed approach as occlusion analysis assisted LGBPHS (OA-LGBPHS). In order to justify that the proposed method is more appropriated for occluded faces, we also tested the standard LGBPHS [157] and its variant KLD-LGBPHS [156] on the same data set; where LGBPHS, KLD-LGBPHS and OA-LGBPHS apply different preprocessing methods to the same face representation. The method RSC [151] is selected to represent the family of algorithms based on sparse representation [96, 146, 150, 151, 160], in which RSC is one of the most robust algorithm according to the reported results. It should be noticed that, in the pool of selected algorithms, KLD-LGBPHS, OA-LBP and RSC stand for the state-of-the-art algorithms for occluded face recognition in each of the 3 categories as we reviewed in Section 2.2 (see Table 2.1).

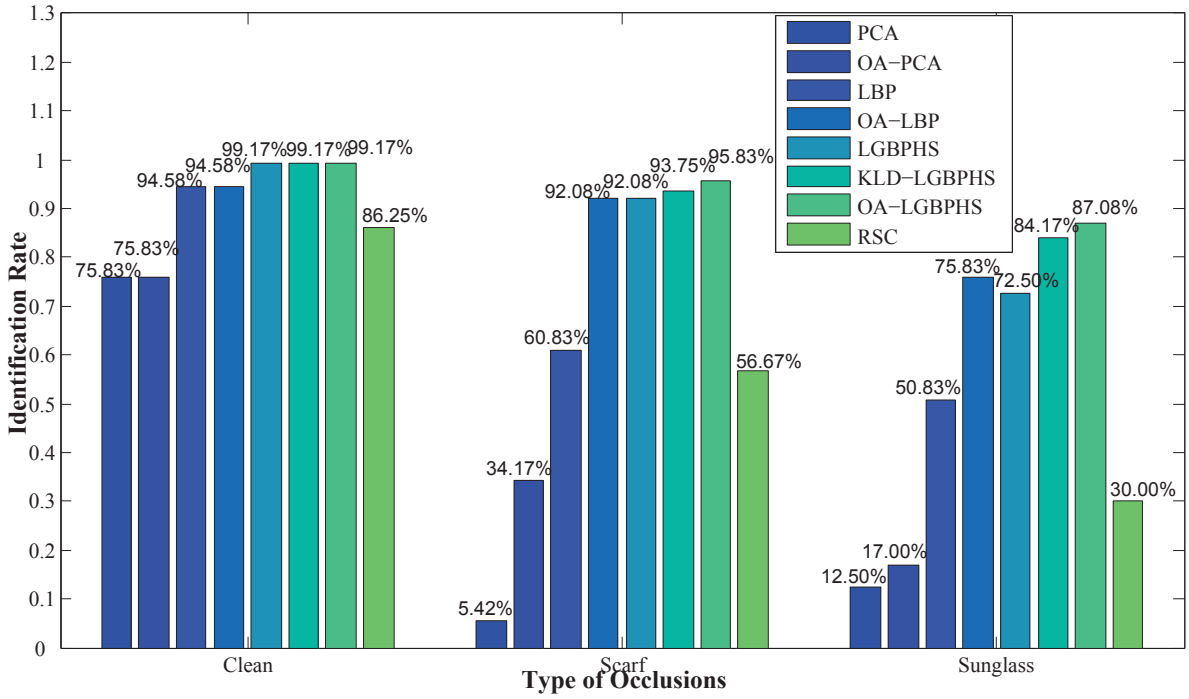


Figure 3.14: Results of PCA, OA-PCA, LBP, OA-LBP, LGBPHS, KLD-LGBPHS, OA-LGBPHS and RSC on three different testing sets (faces are clean and faces occluded by scarf and sunglasses).

In Figure 3.14, it is clear that the proposed approach (OA-LGBPHS) obtains the highest identification rates in all 3 cases (99.17%, 95.83% and 87.08% for clean, scarf and sunglass faces respectively). Without explicit occlusion analysis, facial occlusions such as scarf and sunglasses can greatly deteriorate the recognition results of PCA and LBP; in contrast, OA-PCA and OA-LBP surpass their original algorithms significantly. With a long length feature vector (151040 bins), LGBPHS demonstrates satisfactory robustness to facial occlusions. Without occlusion analysis, LGBPHS can already yield close results to OA-LBP under the occlusion conditions. KLD-LGBPHS improves LGBPHS by associating a weight with each block (which indicates the level of occlusion) to ameliorate the impact from occluded regions, the weight

is measured as a deviation of the target block from the pre-defined mean model based on Kullback-Leibler divergence. Although KLD-LGBPHS greatly increases the results in comparison to LGBPHS (especially for faces occluded by sunglasses), its performance is still inferior to OA-LGBPHS. This result reveals that occlusion exclusion (as in our approach) is more efficient than occlusion weighting (as proposed by KLD-LGBPHS), since distortions due to facial occlusions do not affect the process of recognition when the occluded regions are completely discarded.

Sparse representation based classification (SRC) is well known for its robustness to partial distortions (e.g. noise, occlusion etc.) as well as its discriminative power. However, it also suffers from the "curse of dimensionality" problem, where in many practical cases, the number of templates (of each identity) is insufficient to support the recovery of correct sparse coefficients. On the given data set (240 training faces, with 3 templates for each identity), robust sparse coding (RSC) yields relatively low identification rates (86.25%, 56.67% and 30%).

Comparing the results on the test sets of faces with sunglasses and scarves, we notice that most methods (except for PCA) are more sensitive to sunglasses than to scarf. This is an interesting phenomenon which is in agreement with the psychophysical findings indicating that the eyes/eyebrows region play the most important role in face recognition [133].

We also compared our proposed approach against OA-LBP and S-LNMF [116] using similar protocol under the more challenging scenario in which the gallery face images are taken from Session 1 of AR database while the test sets are taken from Session 2. Note that the two sessions were taken at time interval of 14 days. The comparative results of our approach against OA-LBP and S-LNMF are illustrated in Table 3.4.

Table 3.4: Robustness to different facial variations.

	Sunglass	Scarf	Scream	Right-Light
S-LNMF	49%	55%	27%	51%
OA-LBP	54.17%	81.25%	52.50%	86.25%
OA-LGBPHS	**75%**	**92.08%**	**57.50%**	**96.25%**

The results in Table 3.4 clearly show that our proposed approach outperforms OA-LBP and S-LNMF in all configurations showing robustness against sunglasses, scarves, screaming and illumination changes. The robustness of our approach to illumination changes and drastic facial expression is brought by the use of local Gabor binary patterns, while the occlusion detection module significantly enhances the recognition of faces occluded by sunglasses and scarves even with time elapsing.

Please note that we did not provide the comparative results of our approach to all the literature works (according to our survey in Section 2.2). Instead, we compare our approach to a number of carefully selected methods. Because our method exploits explicit occlusion analysis, KLD-LGBPHS, S-LNMF and OA-LBP which belong to the same category (see Table 2.1) are selected for the comparisons in our experiment. RSC is selected to represent the family of SRC based face recognition. Even though LGBPHS is chosen to stand for the locally

emphasized algorithms without explicit occlusion analysis, our approach could be directly extended to other local feature/classifier based methods for potential improvements.

3.4 Conclusions

In this chapter, we presented our solutions to handle the classical occlusion problem (such as partial occlusions caused by scarves and sunglasses) in face recognition. Our philosophy consisted of first conducting explicit occlusion analysis and then performing face recognition from the non-occluded regions. The salient contributions of presented works are: (i) a novel framework for improving the recognition of occluded faces is proposed; (ii) the new techniques to detect and segment facial occlusion are thoroughly described; and (iii) extensive experimental analysis is conducted, demonstrating significant performance enhancement using the proposed approaches compared to the state-of-the-art methods under various configurations including robustness against sunglasses, scarves, non-occluded faces, screaming and illumination changes.

Although we focused on occlusions caused by sunglasses and scarves, our methodology can be directly extended to other sources of occlusion such as hats, beards, long hairs, etc. As a future work, it is of interest to extend our approach to address face recognition under general occlusions, including not only the most common ones like sunglasses and scarves but also like beards, long hairs, caps, extreme facial make-ups, etc. Automatic face detection under severe occlusion, such as in video surveillance applications, is also far from being a solved problem and thus deserve thorough investigations.

Chapter 4

Advanced Occlusion Handling for Robust Face Recognition

4.1 Introduction

Facial occlusion is a critical issue in many face recognition applications. Existing approaches of face recognition under occlusion conditions mainly focus on the conventional facial accessories (such as sunglasses and scarf). We consider those well studied types of occlusions in the literature as classical occlusions, and proposed our approaches (as presented in Chapter 3) to handle the classical occlusions which outperform the state-of-the-art results. In this chapter, we are going to discuss about facial occlusions in more advanced conditions. Such advanced occlusions are usually not as common as the classical ones, as a consequence, they are omitted in the face recognition literature. However, in specific scenarios (mainly in uncontrolled conditions such as video surveillance), such advanced occlusions can occur and significantly deteriorate the performance of face recognition systems. In this chapter, we will investigate two new types of facial occlusions, namely the sparse occlusion and the dynamic occlusion, and propose corresponding solutions to improve face recognition in presence of such occlusions.

The newly identified facial occlusions – sparse occlusion and dynamic occlusion are in contrast to the classical occlusions with dense and static properties. In order to handle the classical occlusion, works in the literature normally assume the occluded part is dense and contiguous,and identify the presence of occlusion in a single static intensity image. However, due to the wide variety of natural sources which can occlude a human face in uncontrolled environments, methods based on the dense assumption are not robust to thin and randomly distributed occlusions (which is so called sparse occlusions according to our definition). Similarly, to the best of our knowledge, occlusion due to cap (a typical dynamic occlusion scenario) has never been studied in the literature, but the importance of this problem should be emphasized since it is known that bank robbers and football hooligans take advantage of it for hiding their faces.

We present the solutions to address the sparse occlusion and dynamic occlusion problems in the context of face biometrics in video surveillance. For sparse occlusion, we show that the occluded pixels can be detected in the low-rank structure of a canonical face set under the Robust-PCA framework [35]; and the occluded part can be inpainted solely based on the non-occluded part and a Fields-of-Experts prior [128] via spatial inference. Experiments demonstrate that the proposed approach significantly improve various face recognition algorithms in presence of complex sparse occlusions. For dynamic occlusion, we consider a specific case of cap detection in the entrance surveillance scenario. The proposed approach consists of two parts: detection and tracking of occluded faces in complex surveillance videos; detecting the presence of cap by exploiting temporal information. The detection and tracking part is based upon body silhouette [144] and elliptical head tracker [25]. The classification of cap/non-cap faces utilizes dynamic time warping (DTW) [33] and agglomerative hierarchical clustering. The proposed algorithm is evaluated on several surveillance videos and yields good detection rates.

The main contributions of our works in this section can be summarized as follows:

1. We identify two new types (which are not studied in any prior work according to our best knowledge) of facial occlusion (sparse occlusion and dynamic occlusion) which are important issues however omitted in state-of-the-art of occluded face recognition.

2. The idea to detect and then recover sparse occlusions via spatial inference (using inpainting technique) so as to improve face recognition.

3. A complete system consists of detection, tracking and occlusion detection for dynamic occlusion working in real time.

4. Extensive experiments and analysis to justify the proposed solutions on synthetic and real world datasets.

The rest part of this chapter is organized as follows. In Section 4.2 we will first overview the sparse occlusion problem and give a formal definition; then we present the proposed solution supported with experimental results. In Section 4.3, we will present the dynamic occlusion problem and our proposed solution. Finally, we draw the conclusions in Section 4.4.

4.2 Sparse Occlusion

In this section, we present our solution to the newly identified sparse occlusion problem. Following the definition we given in Section 4.2.2, the sparse occlusion problem can be addressed by explicit occlusion modelling using Robust-PCA [35] and inpainting using Field-of-Experts prior [128]. Based on this idea, we built an automatic system to detect and inpaint sparsely occluded faces and demonstrated significant improvements of various face recognition algorithms (we tested PCA [140], SIFT [102] and LBP [17] based face recognition respec-

tively) under different sparse occlusions. Detailed processing steps and results are given in the following sections.

4.2.1 Overview

With the emphasis on real world scenarios (e.g. face recognition in video surveillance), a number of challenges including pose/illumination changes, image degradation as well as partial occlusion is required to be explicitly handled. Facial occlusion, as one of those major challenges, has been extensively studied in the literature. However, almost all previous works [62, 84, 89, 96, 107, 113, 116, 120, 122, 127, 138, 146, 150, 151, 156, 157, 160] focus on facial occlusions which are dense and contiguous (e.g. sunglasses, scarf, beards, hat and hand on face), whereas neglecting the other types of facial occlusions.

De facto, there exists a large variety of facial occlusions in uncontrolled environments. In addition to the well-studied facial occlusions in the literature, in this section, we point out that occlusions caused by facial painting, face dirt, and face behind fence (where the occluded part is often not dense) can also greatly hinder many popular face recognition systems. Inherent from the sparsity/density dichotomy in graph theory [51], we categorize facial occlusions into 2 classes: the dense occlusion and the sparse occlusion (see Figure 4.1, in the figure, classical occlusions such as scarf and sunglasses are considered as dense occlusion; advanced occlusions such as facial painting, face dirt, and face behind fence are considered as sparse occlusion. Definition of sparse/dense occlusion is given in Section 4.2.2.2.). Unlike the traditional studies (such as the works in the literature described in Section 2.2, Chapter 2 and our previous works in Section 3), our goal here is to address the newly identified sparse occlusion problem in face recognition.



Figure 4.1: Examples of various kinds of facial occlusions: (a) densely occluded faces, (b) sparsely occluded faces.

Recently, researchers have revealed that imposing prior knowledge of occlusion can significantly improve results of face recognition under occlusion conditions [113, 116, 156, 160]. Hence, explicit occlusion analysis is an essential step in occluded face recognition. However, since the previous focus is primarily on dense occlusions, the dense assumption is made intentionally or undeliberately in the detection of occluded regions. In [116] and [113], faces are divided into pre-defined local patches for occlusion detection, where the occluded part is supposed to be larger/equal to the patch size and condensed; the occlusion-free patches are then used in local feature based face recognition. Zhou et al. [160] used a Markov Random Fields (MRF) model to incorporate spatial continuity constraints in the modelling of contiguous occlusions (in order to exclude the information from the occluded part) which improves sparse representation based face recognition [146]. Apparently, such methods are inappropriate when dealing with sparsely occluded faces.

Towards the problem of sparse occlusion, we detect occluded pixels with emphasis on the sparsity by explicit occlusion modelling. In addition, inspired by the work of image inpainting [23], instead of simply excluding information from the occluded part (as suggested by [113, 116]), we propose to further recover the occluded part from the non-occluded part via spatial inference. The detection and inpainting of sparse occlusion are then achieved based on the methods of Robust Principal Component Analysis (Robust-PCA) [35] and Fields-of-Experts (FoE) [128], respectively.

In the following parts, we will first outline the problem by giving the definition of dense/sparse occlusion. Then the proposed algorithm (which consists of a sparse occlusion detection module and a occlusion inpainting module) is described in details in Section 4.2.3. Section 4.2.4 presents the experimental results and analysis using the proposed approach on different face recognition algorithms (PCA [140], LBP [17] and SIFT [102]).

4.2.2 Problem Statement

To formally define the term sparse/dense occlusion, we first review the definition of dense/sparse graph in graph theory. Based on this mathematical definition, we then can distinguish dense occlusion and sparse occlusion in particular scenarios. As a consequence any given facial occlusion can be identified accordingly.

4.2.2.1 Dense Graph vs. Sparse Graph

According to the definitions in graph theory [51], the term dense graph and sparse graph can be defined as follows:

Dense Graph [26]: A graph in which the number of edges is close to the possible number of edges.

Sparse Graph [27]: A graph in which the number of edges is much less than the possible number of edges.

In regards to the general definition of graph, a directed (complete) graph can have at most $n(n-1)$ edges, where n is the number of vertices; on the other hand, an undirected (complete) graph can have at most $n(n-1)/2$ edges. Please note that there is no strict distinction between sparse and dense graphs. People usually set a threshold for the number of edges to classify the dense graph and sparse graph.

However, image is a special graph in which only the adjacent vertices (pixels) are connected, namely an undirected adjacency graph. In this case, considering a $M \times N$ image, it has at most $(M-1) \times (N-1)$ connected edges. The threshold to distinguish dense and sparse graph in our case is thus set according to this number of edges. According to the definitions above, we can give the definition of dense and sparse occlusion in the next section.

4.2.2.2 Definition of Sparse/Dense Occlusion

Let us consider an image as an undirected adjacency graph $G = (V, E)$ in which the pixels represent the vertices and the pixel-neighbourhoods represent the edges. Given the occluded part of a face image as an induced subgraph $G' = (V', E')$ of the entire graph G, the occlusion is called dense when the number of edges $|E'|$ in G' is close to the maximal number of edges in an adjacency graph with $|V'|$ vertices and vice versa. By this definition, facial occlusions like sunglasses, scarf, and hat (Figure 4.1 (a)) are regarded as dense occlusions; whereas examples like facial painting, face dirt and face behind fence (Figure 4.1 (b)) belong to the sparse occlusion category.

4.2.3 Method

In this section, we propose a solution to handle sparse occlusions in face recognition. We assume that occlusions are large deviations from a low-dimensional face space. A probe face (well-aligned) is thus represented as the lowest rank reconstruction from a canonical face set (where the faces are well-aligned and non-occluded) added with a sparse error vector under the context of Robust-PCA [35]. Based upon the computed sparse error vector, we can discriminate the large error entries from the small facial appearance agitations so as to detect the occluded pixels on a face image. For the inpainting of occluded pixels, we adopt a generic FoE prior [128] to infer the missing part, which demonstrates significant improvements in both visual quality and recognition results for faces with sparse occlusions. Finally, the recovered face is utilized as input to the face recognition system. A high-level work flow diagram of the proposed method is shown in Figure 4.2.

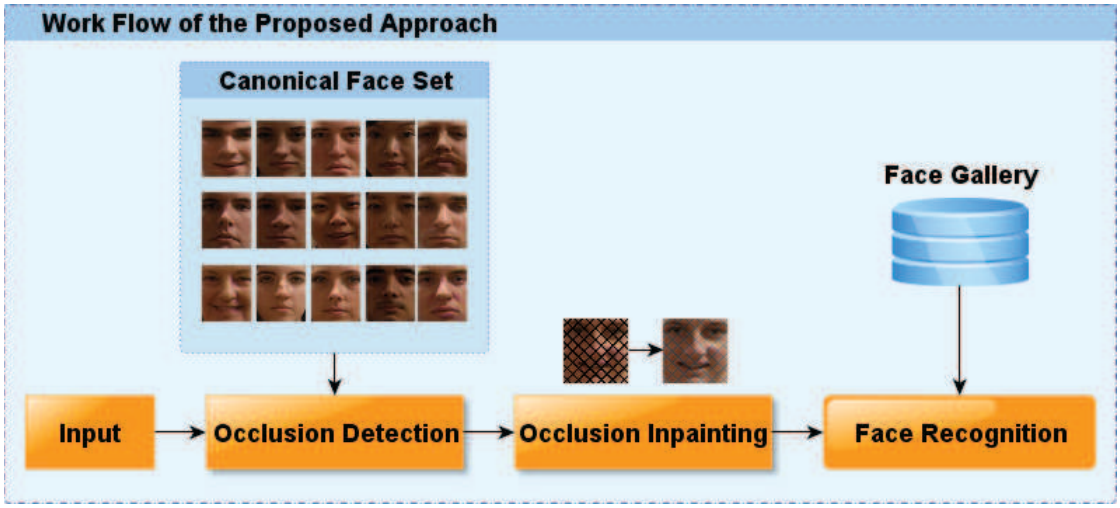


Figure 4.2: A high-level work-flow of the proposed method.

4.2.3.1 Sparse Occlusion Detection

We are given a set of K well-aligned and non-occluded faces $C = \{c_1, c_2, \cdots, c_K\}$ represented by pixel-vectors $c_i \in R^m$, where m is the feature dimension. Given a probe face $y \in R^m$, the occlusion modelling method suggested by [146, 160] (details of SRC based methods for facial occlusions can be found in Section 2.3.2, Chapter 2) is to find a sparse error vector $e \in R^m$ by solving:

$$arg\ min_{(x,y)}\ \|x\|_1 + \|e\|_1 \qquad subj.\ to\ Cx + e = y \tag{4.1}$$

where x is a sparse coefficients vector. (Please notice that here the term 'sparse' is different from the term we used in the dense/sparse occlusion dichotomy. In a vector/matrix, the term 'sparse' indicates the small number of non-zero entries, but not refer to the sparse definition in graph theory.) The prerequisite to correctly find e relies on a sufficient number of well-aligned training samples $\{a_1, a_2, \cdots, a_k\} \in C$ from the same subject of y, so that y can be linearly approximated in a low-dimensional space by:

$$y \cong \sum_{i=1}^{k} a_i \tag{4.2}$$

However, in many practical face recognition scenarios, the training samples of each subject in C are often insufficient (the "curse of the dimensionality" [52] problem, in the extreme case only one template face per subject is available) to correctly resolve equation 4.2. Therefore we loosen the prerequisite by only assuming faces in C are well-aligned and non-occluded. In this sense, we model occlusions as large deviations from a low-dimensional face space derived by the canonical set C (such a set can be different from the gallery set, preferably a smaller set to accelerate computation).

To do so, we first integrate the probe face y with set C to build an observation matrix $C^+ = \{y, c_1, c_2, \cdots, c_K\}$, where $C^+ \in R^{m\times(K+1)}$ was generated by a low-rank matrix $A \in R^{m\times(K+1)}$ with large corruption (occlusion) on C_1^+ and minor errors (small facial appearance agitations) on $C_{2\sim K+1}^+$. The corruptions are represented by an additive matrix $E \in R^{m\times(K+1)}$, where $C^+ = A + E$. Our goal is thus to recover the correct corruptions E', more specifically E_1'. Because facial occlusion affects only a portion of the entire face, thanks to the "blessing of the dimensionality" [52], the sparse error E can be efficiently and exactly separated from the low-rank structure of C^+. This problem formulation can be effectively resolved by the Robust-PCA framework [35] (see Figure for a visual interpretation) via the following optimization relaxation:

$$arg\ min_{(A,E)}\ \|A\|_* + \gamma\|E\|_1 \qquad subj.\ to\ A + E = C^+ \tag{4.3}$$

where $\|\cdot\|_*$ is the nuclear norm (also known as the trace norm, defined as $\|A\|_* = trace(\sqrt{A * A})$) which pursues the lowest rank A that aims to regenerate the observations; and $\|\cdot\|_1$ is the l1-norm which pursues the sparsity of errors. We adopt the inexact Augmented Lagrange Multiplier [101] (ALM) method to solve equation 4.3 due to its reported accuracy and efficiency.

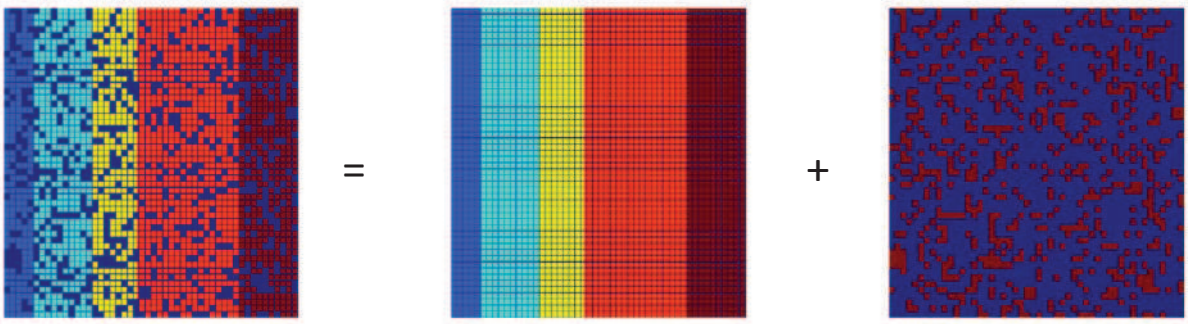


Figure 4.3: A visual interpretation of the robust-PCA (excerpt from [14]). Left: matrix of corrupted observations, in our case, the first column is the occluded faces, all other columns are the canonical faces. Middle: the underling low-rank structure of the clean face subspace. Right: the reconstructed sparse error matrix, in our case, the first column contains large errors caused by occlusion.

Once the sparse error vector E_1' is computed, we exploit it to discriminate the large error entries (regarded as the occluded pixels) from the small facial appearance agitations by giving a pre-defined threshold:

$$M\{i\} = \begin{cases} 1 & |E_1'(i)| > \tau, \forall i \in [1, m] \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

τ is selected empirically to minimize the detection error. M is the indicator of occlusion (i.e. $M(i) = 1$ if pixel i is occluded), where M will be served as the mask which supervises the sparse occlusion inpainting approach introduced in the next section.

4.2.3.2 Inpainting of Sparse Occlusion

In this paper, we apply an image inpainting method based on the Fields-of-Experts (FoE) model [128] to infer the sparsely occluded pixels. The FoE model learns a generic image prior P_{FoE} (a high-order Markov Random Field model) from a large number of nature image patches. P_{FoE} models the local image structures over extended image neighbourhoods (i.e. the local spatial properties), and therefore can be used to predict the missing part from existing observations via probabilistic inference.

In a common image inpainting setting, ground truth mask of the region to be inpainted is known. In contrast, we are facing a more challenging scenario where the part to be inpainted (occluded pixels) is unknown. Hence, we supply the mask of pixels which should be inpainted by our automatic occlusion detection ($M(i), i \in [1, m]$, given in Section 4.2.3.2). Let M be the mask used in the inpainting process, given the learned image prior P_{FoE}, the image inpainting method based on the FoE prior described in [128] can be casted as the following gradient ascent-based process:

$$x^{(t+1)} = x^{(t)} + \eta * M * [\nabla_{x^{(t)}} \log P_{FoE}(x^{(t)})] \tag{4.5}$$

where t is the iteration index and η is the update rate; the mask M sets the gradient to zero for all pixels outside the masked region.

Figure 4.4 shows the inpainting results of various sparse occlusions of the same face. In the figure, it is clear that all four faces have large distortions where their PSNR are all below 20 dB. If the ground truth masks are given, using FoE prior based image inpainting, the PSNR of inpainted faces improve significantly (up to 39.12 dB). When using the masks returned by our automatic occlusion detection, the recovered images also achieve good visual quality improvements (where their PSNR are all above 26 dB); although some occluded pixels are not detected and thus not inpainted in the eyes and eyebrows region due to the similar appearances.

When the input face is inpainted by the proposed approach, it is then fed into the system for face recognition. Our experiments demonstrate that the proposed method cannot only improves the visual quality of sparsely occluded faces but also significantly improves the results of face recognition systems.

4.2.4 Results

To assess the performance of our proposed approach, we performed a series of experiments on AR face database [106], with different types of artificially generated sparse occlusions. The dataset and detailed configurations of our experiments are firstly introduced. Then we will illustrate that the recognition results of face recognition algorithms accompanied with our proposed occlusion detection and inpainting approach which can significantly surpass

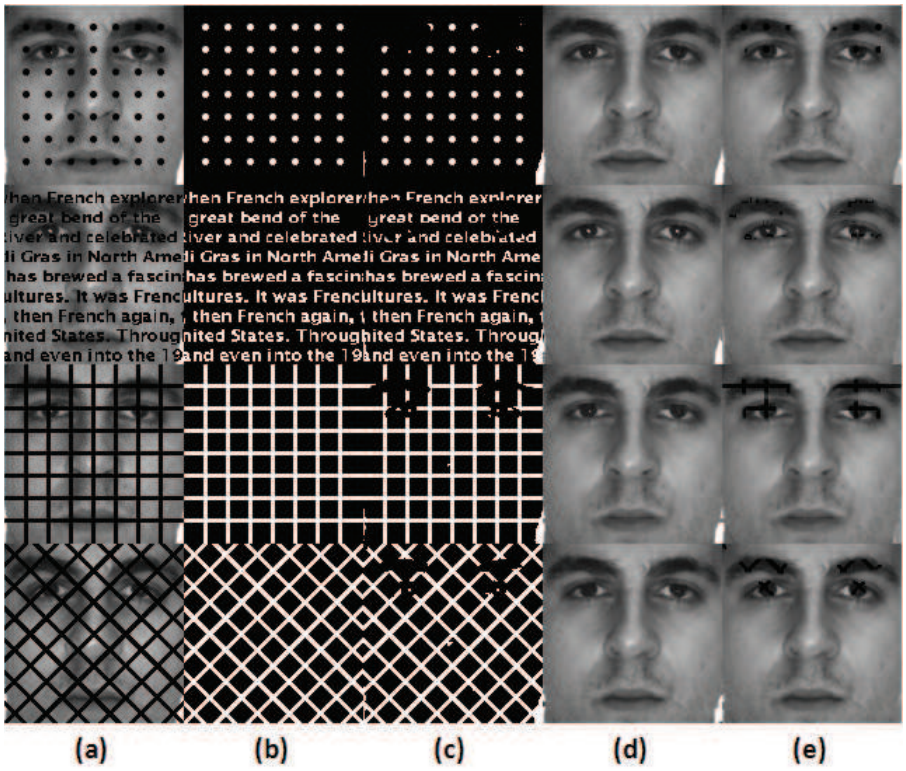


Figure 4.4: Illustration of our sparse occlusion inpainting: (a) faces with different sparse occlusions (stain, text, orthogonal grid, and diagonal grid), PSNR=19.12 dB, 13.92 dB, 13.25 dB, 12.81 dB ; (b) ground truth masks of the sparse occlusions; (c) results of our sparse occlusion detection (τ = 0.004); (d) faces after inpainting using the masks in (b), PSNR=39.12 dB, 34.05 dB, 33.26 dB, 32.51 dB; (e) faces after inpainting using the masks in (c), PSNR=30.43 dB, 28.30 dB, 26.05 dB, 26.50 dB.

the results from the standard algorithms (Eigenface [140], SIFT [102] and LBP [17]) in presence of sparse occlusions.

The AR face database contains more than 4000 face images of 126 subjects (70 men and 56 women) with different facial expressions, illumination conditions, and occlusions. For each subject, 26 pictures were taken in two separate sessions (two weeks interval between the two sessions). The original image resolution is 768x576 pixels. Using the eye coordinates, we cropped, normalized and down-sampled the face region into 128×128 pixels. In our experiments, 300 non-occluded faces (with facial expression and illumination variations) are randomly selected to form the canonical face set C. For face recognition, 100 subjects (half of male and half of female) are selected. For each subject, we chose 14 images with different illumination conditions and facial expressions: 7 images from session 1 as the template faces and 7 images from session 2 as the probe faces. The probe faces are imposed by 4 kinds of artificially generated sparse occlusions (stain, text, orthogonal grid and diagonal grid as shown in Figure 4.4) to simulate the real-world scenarios.

We tested 3 different face recognition algorithms on the proposed dataset, namely PCA, SIFT and LBP based face recognition, with and without the proposed occlusion detection and

inpainting, respectively. For PCA based method, template faces (occlusion-free) are used to train the Eigenspace for both template and probe faces representation. For SIFT and LBP based method, features are extracted from both template and probe faces for the Nearest-Neighbor (NN) based classification. The SIFT feature extraction is adopted from [102], and the LBP operator $LBP^{u2}_{8,2}$ is used in our experiment. Other settings of the experiments are listed here: $\gamma = 1/\sqrt[2]{m}$, $\tau = 0.004$; the inpainting process are consists of 2 steps: a rough step with $t = 500$ and $\eta = 10$, and a refined step to "clean up" the image with $t = 250$ and $\eta = 0.01$ as suggested by [128].

Figure 4.5, Figure 4.6, and Figure 4.7 shows the recognition rates of PCA, SIFT and LBP based algorithms on the clean face set and the faces with different types of sparse occlusions, respectively. It is clear that without explicit treatment, sparse occlusions can greatly deteriorate the results of those face recognition algorithms (the results marked as "Original"). In the figure, SIFT based method achieves very accurate recognition rate (97.57%) for non-occluded faces, however it is very sensitive to the sparse occlusion distortions (less than 25% in all cases, since the descriptor summarizes the edge-like features which correspond to the sparse occlusions located on the probe faces). LBP based method are somewhat robust to certain types of sparse occlusions (stain, text and orthogonal grid, because those occluded parts are located at the boarder of 8×8 LBP blocks); however when the occluded part is located inside the LBP blocks (e.g. the diagonal grid case) its recognition rate decreases drastically (down to 28.57%). Those results illustrate that even if local-feature based methods are known to be somewhat robust to conventional partial occlusions (such as scarf and sunglasses); they are still fragile to sparse occlusions.

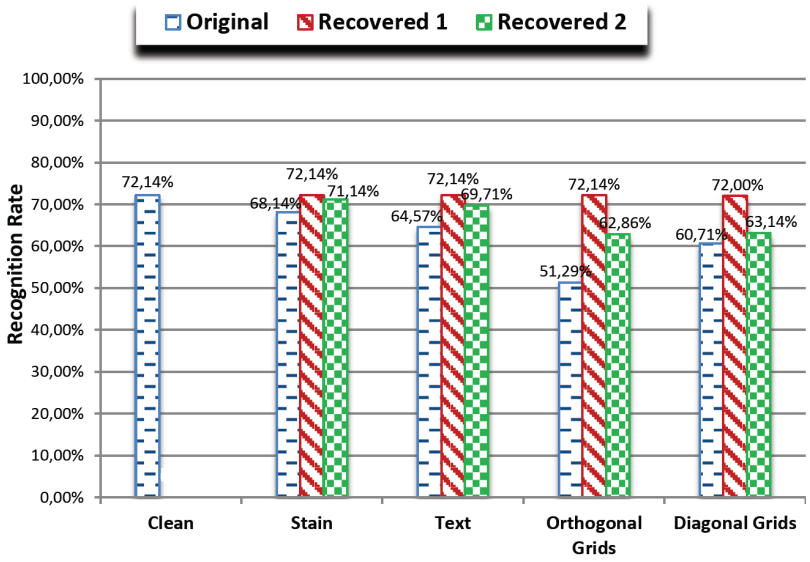


Figure 4.5: Face recognition results based on PCA.

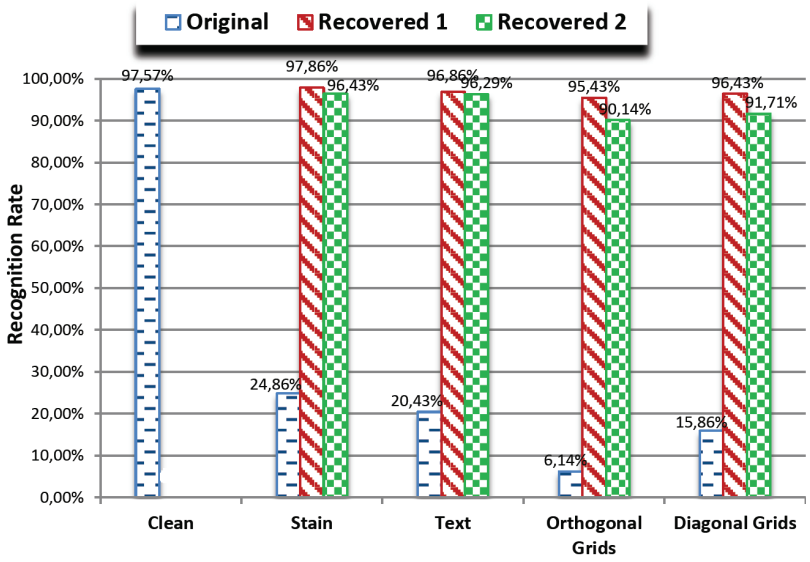


Figure 4.6: Face recognition results based on SIFT.

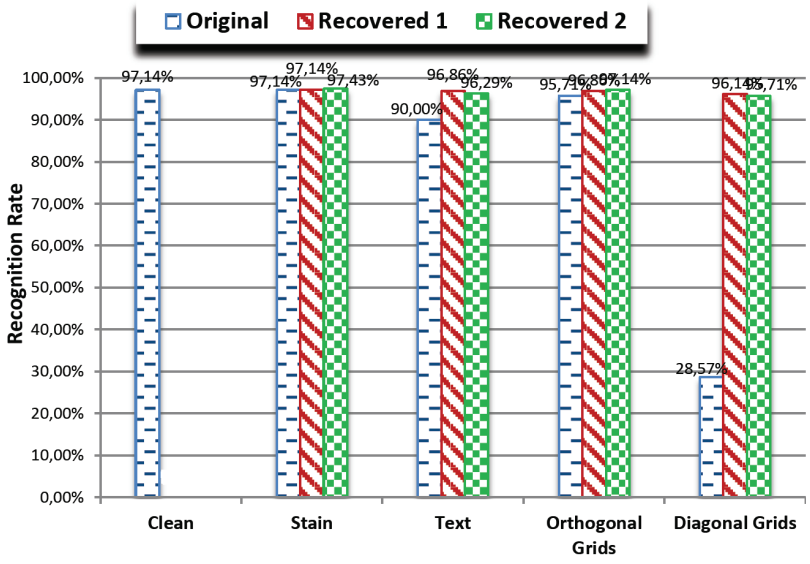


Figure 4.7: Face recognition results based on LBP.

In the figure, the inpainted faces using FoE prior based on the ground truth masks ("Recovered 1") can achieve recognition rates as good as the non-occluded faces (with the deviations less than 2%). In addition, using the proposed sparse occlusion detection and inpainting ("Recovered 2") can significantly promote the recognition rates in comparison to the results from the occluded faces. For LBP based method, the proposed approach can even achieve

recognition results very close to the non-occluded faces (97.14%), because the small distortions (due to the detection errors) in few LBP blocks do not impair the discriminative power of the overall representation.

From the experimental results, we can conclude that explicit occlusion detection and inpainting can greatly improve the results of those face recognition algorithms when dealing with sparse occlusions.

4.3 Dynamic Occlusion

In this section, we present our solution to the newly identified dynamic occlusion problem. For dynamic occlusion, we consider a specific case – cap detection in entrance surveillance. The presence and importance of this problem is firstly outlined in Section 4.3.1. Then we present the challenges of the target scenario and the innovations of our proposed system in Section 4.3.2. The proposed dynamic occlusion detection system is comprised of many steps ranging from detection, tracking, feature extraction and classification. We therefore detail each step in Section 4.3.3. To justify the proposed system, it is tested on a real world dataset of 40 video sequences, yields significant detection rates.

4.3.1 Overview

It should be noted that the state-of-the-art of occluded face recognition addresses the time-invariant occlusions only (e.g. scarf, sunglasses or artificially generated occlusions) from a single image. Time-invariant indicates that the occlusion is not changing without external forces within a certain period of time. Here we identify a new type of occlusion: the time-variant (dynamic) occlusion, a specific case can be often observed is the occlusion due to cap in entrance surveillance. Nowadays, Closed-Circuit Television (CCTV) cameras are deployed everywhere for security purpose. The most common deployment of such cameras is at the room ceiling in order to monitor the entrances of various places (e.g. banks, ticket machine of subway station, supermarkets, libraries, and stadium) (see Figure 4.8). With videos captured by such cameras, the identities of people inside the restricted place can be recorded and later recognized by automatic face recognition (AFR) systems or human observers. However, according to many recent police reports, bank robbers, shop thieves and football hooligans usually wear a cap when entering the places where they commit crimes. Due to the viewing angle problem, a close-distance face captured by a CCTV camera is often occluded by the cap visor, whereas a far-distance face is not occluded but it has a low resolution. The occlusion caused by cap with a visor is therefore varying along with people moving. Because of the trade-off between occlusion and image quality, obtained face images are usually unrecognisable from the recorded videos. Thus, it is imperative to equip automatic occlusion detection systems in entrance surveillance, which cannot only detect suspicious persons, but also provide prior knowledge of dynamic occlusion to improve the face recognition of criminals.



Figure 4.8: Illustration of Entrance surveillance scenarios.

We address then the occlusion detection due to cap for entrance surveillance. The proposed approach consists of first detecting and tracking the face region and then detecting the presence of cap within the tracked face. The detection and tracking part is based upon silhouette analysis [144] and elliptical head tracker [25]. The classification of cap/non-cap faces adopts dynamic time warping (DTW) [33] and agglomerative hierarchical clustering.

In the following parts, we will first illustrate the challenges in the proposed cap detection (dynamic occlusion) scenario as well as the innovations of the proposed system. Then each step of the proposed algorithm is discussed in details in Section 4.3.3. Finally, we show the results of our cap detection system on a real world dataset to demonstrate its efficiency.

4.3.2 Challenges and Innovations

There are many challenges for the proposed cap detection system. First of all, in entrance surveillance, even a close-distance face is relatively far away from the camera (in comparison with the traditional biometric scenario). In addition, various variations due to occlusion, rotation, low resolution and complex backgrounds make the correct detection of face region a difficult problem. In the targeted entrance surveillance scenario, the most common approach – Viola-Jones algorithm [143] implemented in OpenCV cannot even find a non-occluded face correctly (mainly due to the resolution problem). For this reason, we implemented a customized method to detect and track the face region, which exploits more robust features such as body silhouette and head geometry.

Secondly, occlusion due to cap is time-variant. Unlike the time-invariant occlusions (such as scarf and sunglasses), the occlusion caused by cap is varying along time. When a face is approaching, the occluded area on the face increases accordingly. The changing of occluded area mainly depends on the speed and pause of people walking. Since the walking habit

is non-homogenous and largely different among individuals, the occlusion situation is also diverse for different people or even the same people in different videos. In addition, the occluded area is also changing because of the rigid head movements (e.g. head nodding and looking way). For above reasons, occlusion detection using a single image is very unlikely to get accurate results. In our approach, we exploit all the frames in the video to reduce the effects of different variations and tracking errors; furthermore, the temporal information is preserved by using Dynamic Time Warping (DTW) to distinguish the characteristics of occlusion varying from the occluded and non-occluded faces.

Last but not least, the proposed system is supposed to work in real time, since the computational capability is limited in distributed sensor networks. The flowchart of the proposed system is given in Figure 4.9. The simplicity of each of the steps ensures the overall performance of our system.

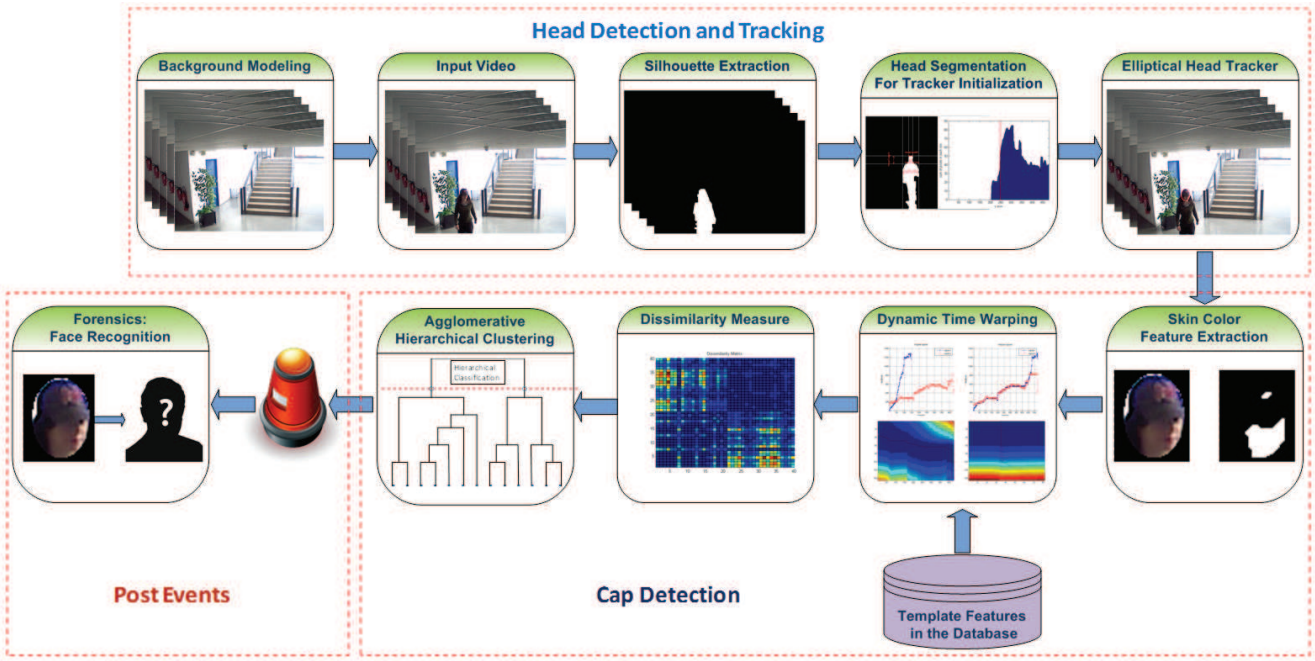


Figure 4.9: Illustration of Entrance surveillance scenarios.

4.3.3 Method

The proposed cap detection system is comprised of two major modules: the head detection and tracking module and the dynamic occlusion detection module. To tell whether a people is occluded by a cap in a short video sequence, the head part of the people should be firstly located and tracked in each frame. Based on the information extracted from the region of interest (ROI) in each frame, we apply a classification strategy to determine whether or not the face region is partially occluded.

4.3.3.1 Head Detection and Tracking

The customized head detection and tracking is based on silhouette image and elliptical head tracker. The accuracy of detection and tracking thus largely depends on the quality of extracted body silhouette. Following the well established silhouette analysis in gait recognition [144]; our system is able to achieve good detection and tracking of non-optimal faces in textured background.

Background modelling: to extract the body silhouette, the first step is to construct a reliable background model for silhouette extraction. Here, the LMedS (Least Median of Squares) method [152] is applied to build the background from a small portion of video sequences, thanks to its power of filtering moving objects. Let I represent a video sequence including N frames, the resulting background can be computed as follows:

$$b_{xy} = \min_{n} med_t(I^t_{xy} - p)^2 \tag{4.6}$$

where p is the background intensity to be determined for the pixel location (x, y), med is the median value, and t is the frame index of the video sequence (with N frames). As suggested by [152], $N > 60$ is sufficient to retrieve a stable background. Figure 4.10 shows the example of modelling the background from 60 frames using LMedS.



Figure 4.10: Background estimated from 60 frames using LMedS.

Silhouette extraction: once the background image is correctly constructed, the foreground detection can be achieved by differencing and thresholding between the background image and the current frame. Because the background subtraction step may introduce noise and distortions in the segmented foreground, post-processing including morphological operations and connected component selection are applied in order to obtain a clear body silhouette.

Head segmentation: correct segmentation of head from the binary body silhouette is critical to initialize a head tracker. The bounding box of walking body is naturally obtained from the silhouette. To segment the head from the other part of the body, we apply a horizontal projection to the silhouette image (accumulating the pixel values in each row). Figure 4.11,

shows the horizontal projection of a given silhouette image. Then the head-shoulder can be selected by the steep rise in the projection histogram. With the upper the shoulder image, we can easily find the geometric centroid of the largest connected component, so as to locate the head region.

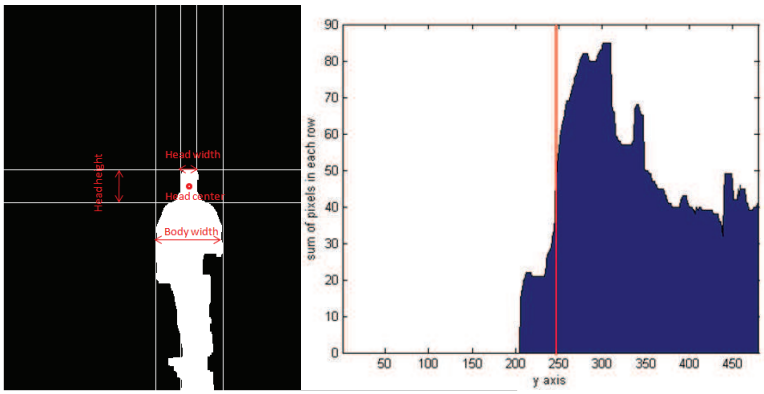


Figure 4.11: The proposed head segmentation method. Left: a silhouette image. Right: its horizontal projection.

Scalable elliptical head tracker: The tracking part is directly adopted from [25]. The choice of the elliptical head tracker is based on the fact that it is robust to occlusion, rotation, tilting and textured background to some extent. In addition, the merits of scalability, reacquisition and low complexity encourage us to integrate it into our system. However, instead of applying the elliptical head tracker to a gradient map, we fit the ellipse to the body contour which is extracted by a Canny edge detector from the silhouette image. The reason is due to the gradient map of a low resolution head is indistinguishable from the textured background. Given the Canny edge map of the silhouette image, the customized elliptical head tracker can be formed to maximize the dot product of the gradient magnitude and ellipse norm as below:

$$s^* = arg\max_{s \in S}\{\frac{1}{N_b}\sum_{i=1}^{N_b}|n_b(i)\cdot g_s(i)|\} \tag{4.7}$$

where $s = \{x, y, b\}$ stands for the head statues, in which (x, y) is the estimated ellipse center, and b is the length of minor axis; $g_s(i)$ is the intensity value at perimeter pixel i at location s, $n_b(i)$ is the unit vector normal to the ellipse at pixel i, and N_b is the number of pixels on the perimeter of an ellipse with size b. In [25], the tracking is achieved by predicting the location (x, y) of the ellipse center based on the one in the previous frames with fixed b. In our system, we predict also the ellipse size (based on b) to maximize the energy function 4.7 for a current observation. This is because in our entrance surveillance scenario, the head size is growing along time, which shall be considered deliberately in tracking.

In the prediction part, the monotonic increment assumption is made (where b is increasing monotonically). This assumption can accelerate the system speed as well as automatically exclude the person who is leaving the entrance (in which case the face is invisible and also re-

dundant, it was recorded during entering). Examples of our tracking results can be observed in Figure 4.12

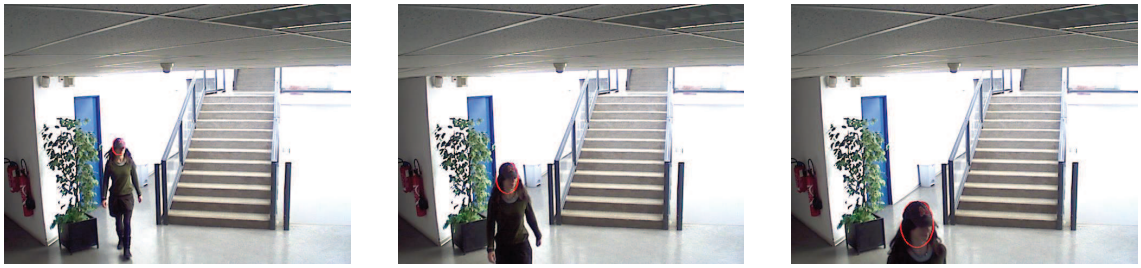


Figure 4.12: The scalable elliptical head tracker: the head size is growing along time according to our monotonic increment assumption.

In our experiments, the elliptical head tracker correctly tracked the head in 84% of all the frames (of 40 video clips) for 10 different people with and without cap under 2 illumination conditions (morning and evening). Errors are due to incorrect ellipse fitting (23%) or lost tracking (77%).

4.3.3.2 Dynamic Occlusion Detection

Because facial components, such as eyes, nose and mouth, are almost undetectable in the entrance surveillance scenario, the only reliable feature for occlusion classification is the skin color. In this section, a novel Spatial-Temporal (ST) feature representation based on skin color is proposed. Inspired by the work from [33], DTW and agglomerative hierarchical clustering are employed to classify the unsynchronized ST features.

Skin color based feature extraction: a traditional way [36] to extract skin pixels which is fast and stable is applied. Skin pixels are differentiated from the non-skin pixels in the $YCbCr$ color space by pre-defined thresholds (for any pixel p_i whose $p_i^{Cb} \in [77\ 127]$ and $p_i^{Cr} \in [133\ 173]$ is detected as a skin pixel).

Based on the extracted skin pixels, we tested two types of features: 1. the number of skin pixels in the region of interest (ROI); 2. the ratio of number of skin pixels to number of all pixels in the ROI. Then the extracted features from all frames of one video clip are concatenated to construct a ST representation. The perceptual classifications between features extracted from the occluded and non-occluded video clips are shown in Figure 4.13. From the figure, we can observe that the non-occluded class and the occluded class are well separated by the proposed ST feature.

Dynamic time warping: dynamic time warping, also called curve registration, is widely used in various applications (e.g. speech recognition, musical information retrieval and bioinformatics), thanks to its capability of measuring the similarity between two sequences varying in time or speed. The optimal match between two feature vectors is found by stretching and compression under the constraint of monotonicity. First, a difference ma-

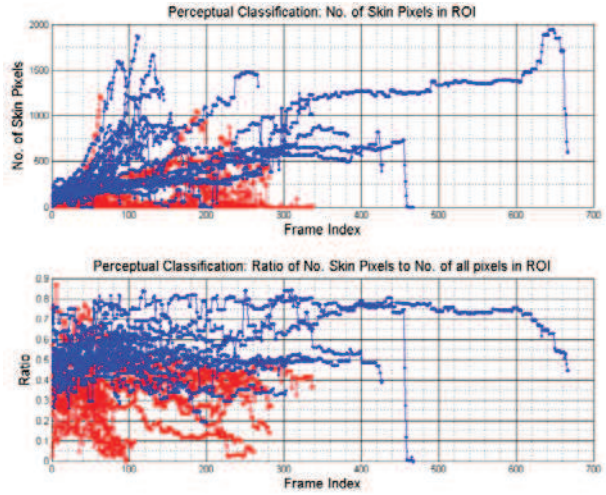


Figure 4.13: Perceptual classification of the proposed features (Blue: faces without cap, Red: faces with cap).

trix $D[i,j] = |Q[i] - T[j]|$ of the query feature vector $Q[i]$ and the template feature vector $T[j]$ is calculated. Then a cost matrix M is recorded by keeping a running tab on the dissimilarities of elements while summing up to a minimum accumulated cost measure via dynamic programming [33]. The formation of M can be written as:

$$M[i,j] = \min \begin{pmatrix} M[i-1,j-1] \\ M[i-1,j] \\ M[i,j-1] \end{pmatrix} + D[i,j] \tag{4.8}$$

With matrix M, a minimum cost path can be retraced along diagonal. The final dissimilarity between the query and template is thus the summation of all costs on the path, normalized by the length of the longer feature vector between Q and T. Because our features extracted from different video sequences are unsynchronized, DTW is an ideal tool for dissimilarity measure in our classification process.

Dissimilarity matrices: in our cap detection system, a number of features extracted from video clips of occluded and non-occluded faces serve as the templates. When a query video comes in, the dissimilarities between the query and all templates are computed via DTW. To reduce computation, the dissimilarities between all pairs of templates are pre-calculated and stored as dissimilarity matrices in the database. The system also support online updating by integrating the dissimilarities computed from the query into the matrices. A visual presentation of the dissimilarity matrices is shown in Figure 4.14.

Agglomerative hierarchical clustering: knowing the dissimilarity matrices, the agglomerative hierarchical clustering algorithm is applied for classification. A dendrogram is returned by the clustering algorithm. In the dendrogram we choose the level with two clusters for decision making. The class label of each cluster is decided by the number of template features within the same cluster; as an example, a cluster belongs to the cap class when there

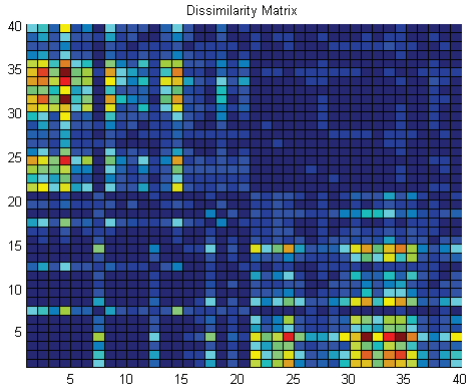


Figure 4.14: Dissimilarity Matrices of 40 video clips (1-20 without cap, 21-40 with cap, cold color indicates low dissimilarity).

are more template features from cap videos in the cluster and vice versa. A query feature is thus classified according to the cluster's class label. Ward's method is chosen empirically for the clustering task.

4.3.3.3 Post Events

Here we briefly discuss the post events after cap detection. In the entrance surveillance scenario, the presence of cap is undesirable for security management. In the cooperative case (e.g. in airport security check), the system can provide information so that guards can kindly ask the passenger to remove his/her cap. In the non-cooperative case (e.g. in bank), it can identify the suspicious person so that increase the security level or draw the attention from a human observer.

To improve face recognition, the information of dynamic occlusion we obtained here can be naturally followed by our proposed method in Section 3 for static occlusions. Once the presence of dynamic occlusion (e.g. a cap) is presented in a video sequence, our static occlusion detection method and local feature based face recognition method can be applied in each frame, so as to achieve robust identity classification.

4.3.4 Results

To justify the proposed approach, we built a rather challenging dataset to simulate the environment of entrance surveillance in real-world. 40 videos (length varying from 60 to 666 frames) have been recorded using an Axis P1343 IP camera fixed at the room ceiling in an indoor terrasse (with sunlight during the daytime and artificial lights in the evening). 10 people were asked to walk toward to the camera and enter the door of office below the camera. Each person had twice recording, one time wearing a cap and another time without wear-

ing a cap. The caps have different colors and textures in order to maximize the variations. The experiments are organized in 2 sessions, one session in the morning and another one in the evening. In sum, 40 videos are comprised of 20 with cap and 20 without cap, under 2 different illumination conditions.

In our experiment, we tested 2 different features within 3 different ROI. The features are: 1. number of skin pixels (Skin) in ROI, 2. ratio of number of skin pixels to number of all pixels (Ratio) in ROI. Three ROI are selected automatically according to the minor diameter of ellipse returned by the head tracker. Illustrations of different ROI are given in Figure 5.



Figure 4.15: Pre-defined ROI 1-3 (from left to right: the whole ellipse, the face region, the upper-eyebrow region).

Table 4.1 shows the results of our experiment. Here we define that a people without wearing cap is "accepted" and the one wearing cap is "rejected". It is clear that Ratio has better classification accuracy than Skin. Ratio_ROI2 and Ratio_ROI3 lead to the best result (up to 87.5%). For Ratio_ROI2, 1 person is wrongly "accepted" and 4 people are wrongly "rejected", whereas nobody is wrongly "accepted" using Ratio_ROI3.

Table 4.1: Results of the proposed system.

Feature	FAR	FRR	Classification Accuracy
Skin_ROI1	15%	40%	72.5%
Skin_ROI2	15%	40%	72.5%
Skin_ROI3	35%	5%	80%
Ratio_ROI1	5%	30%	82.5%
Ratio_ROI2	5%	20%	87.5%
Ratio_ROI3	0%	25%	87.5%

The proposed dynamic occlusion detection algorithm isolates the occluded faces from the non-occluded faces. After the occluded faces are correctly detected, the static occlusion detection described in Section 3 can be applied to single frame of the video in order to further determine which part of the face is occluded. Then the non-occluded part of a face can be used in face recognition.

4.4 Conclusions

We present the first solutions to the newly identified sparse occlusion problem and dynamic occlusion problem in the context of face biometrics in video surveillance. The proposed approaches exploit several advanced image analysis and processing techniques in order to achieve the complicated goal. For sparse occlusion, the proposed system automatically detects sparse occlusion on faces (using R-PCA) and then inpaints the occluded part (using FoE prior) to improve face recognition. We have demonstrated the significant improvements of various face recognition systems based on the proposed approach via extensive experiments. For dynamic occlusion, the proposed system can be applied to a wide range of applications for security management in nowadays video surveillance. In particular, it can provide prior information of occlusion to face recognition system of identifying suspicious persons. It overcomes several challenges due to sensor and human behaviours in the imperfect world. In addition, it exploits the temporal information to analyse the characteristics of occlusion varying.

As the future works, it is of interests to detect dynamic occlusion (such as cap) in crowed scenes. In addition, to identify new but overlooked facial occlusion problem is very important. The solutions to such new occlusion types may be helpful to improve face recognition in many practical cases. One potential work could be address the occlusion problem caused by the reflection of windscreen of cars, so as to improve face recognition to identify the car pilots for many security management and access control purposes.

Chapter 5

Exploiting New Sensor for Robust Face Recognition

5.1 Introduction

As one of the most successful computer vision technologies in recent years, Kinect sensor [2] depicts a broad prospects of 3-Dimensional data based computer applications. Taking advantages of its efficiency (i.e. real-time processing), good accuracy, low-cost, ease of RGB-D mapping and multiple-modalities, the designing of novel algorithms and applications for Kinect sensor has received vast amounts of attentions from various research communities [158] not only in computer vision [131], but also in computer graphics [72, 80], augmented reality (AR) [57], human-computer-interaction (HCI) [85], instrument measurement [88] and robotics [59]. For the biometrics community, directly inheriting from the power of body parts segmentation and tracking using Kinect [131], a number of works has been presented for gait recognition [65, 126, 135] and body anthropometric analysis [19, 64, 141].

However, the adoption of this powerful new sensor for face recognition has been mostly overlooked (although we introduce a pioneering work of 3D face recognition specific to Kinect in [111], details can be found in Chapter 6). The reason is due to the lack of standard testing dataset that limits the deployment of new algorithms and applications in this very dynamic research domain. Therefore, it is of great importance to provide a standard database for researchers to design and experiment 3D and multi-modal (i.e. 2D+3D) based face recognition using Kinect data, so as to establish the connection between Kinect and face recognition research communities.

For above reasons, in this chapter, we present the first publicly available face database (KinectFaceDB[1]) based on Kinect sensor for face recognition. The database consists of different data modalities (well-aligned and processed 2D, 2.5D, and 3D based face data) and

[1] online at http://rgb-d.eurecom.fr

multiple facial variations. We conduct benchmark evaluations on the proposed database using standard face recognition techniques (including PCA [140], LBP [17] and TPS warping parameters [58]), and demonstrate the performance gain from Depth to RGB via score-level fusion. We also report the performance comparisons between Kinect images (KinectFaceDB) and traditional high quality 3D scans (FRGC database [123]) in the context of face biometrics, demonstrates interesting results.

In previous chapters, we have already seen that partial occlusion is a very challenging problem which belongs to a number of different classes (dense/static/sparse/dynamic) in face recognition, and thus should be explicitly handled to achieve robust results. With embodiment of the emerging Kinect sensor, it is of our great interest to understand whether Kinect can be helpful to handle the partial occlusion problem in face recognition. Thanks to the developed KinectFaceDB, we propose in this chapter a new approach to address the partial occlusion problem for face recognition. The key idea is to exploit heterogeneous cues (both depth and RGB) from Kinect acquisition for different tasks: namely the occlusion analysis and face recognition. In the proposed solution, we first conduct occlusion analysis based on depth image, then use this information to improve LGBP based face recognition [157] (using intensity/RGB image) via a weighting strategy. Results on KinectFaceDB show that improvements are achieved in comparison to LGBP and KLD-LGBP [156] in various occlusion conditions. Our experiments demonstrate that, in comparison to the traditional 2D sensors, Kinect do improve face recognition in presence of partial occlusions.

The main contributions of our works in this chapter can be summarized as follows:

- A complete multi-modal (including well-aligned and processed 2D, 2.5D, and 3D based face data) face database based on Kinect sensor is built and thoroughly described.
- Benchmark evaluations and fusion on this database are conducted.
- Face recognition results on both KinectFaceDB and FRGC are compared following the same protocol, in order to demonstrate the data quality effects of Kinect under the biometric context.
- A novel solution based on RGB-D Kinect sensing is proposed, which can significantly improve face recognition results under occluded conditions.

The rest of this chapter is structured as follows. We first present our face database (KinectFaceDB) based on Kinect sensor in details in Section 5.2, in which the 3D databases in the literature are also reviewed in Section 5.2.2. Then our solution to exploit both RGB and depth cue for robust face recognition using Kinect is presented in Section 5.3, which is supported by experimental results on KinectFaceDB. Finally, we draw conclusions in Section 5.4.

5.2 Kinect Face Database

In order to fill the gap between traditional face recognition research and the emerging Kinect technology, as well as providing a standard medium to develop robust face recognition algorithms, in this section, we build the publicly available EURECOM Kinect face database. We first give an overview to explain why such a database based on Kinect is necessary in Section 5.2.1. Then a literature review of the existing 3D face databases is conducted, in order to illustrate the difference between the new database and the existing ones. In Section 5.2.3, details of the proposed database and how the data acquisition is achieved are given. Finally, extensive experiments are performed on both the Kinect face database and FRGC [123], demonstrating interesting results in Section 5.2.4.

5.2.1 Overview

Recent surveys [16, 28] have suggested that face recognition exploiting 3D shape cues (either in the format of 3D curvature description or 2.5D depth feature) demonstrates the superiority over typical intensity image based methods. It is true that 3D face recognition inherits intrinsic merits per se. For instance, 3D shape information is invariant to illumination variations; it supplies complementary information with respect to 2D textures; and viewpoint variations are readily addressed by rigid surface registrations [24, 40]. However, because most of the literature works report their results on the face database with high quality 3D data (e.g. 3D faces in FRGC, which are obtained by a digital laser scanner [3], with depth resolution of $0.1\ mm$ within the typical sensing range), there is an unbalanced matching between the 2D and 3D data in terms of both efficiency and accuracy. With respect to efficiency, the acquisition of a high resolution RGB image normally takes $< 0.05s$, whereas the laser scanning of a face depth map consumes $9s$ in average [123] (and hence with high user cooperation, which is conflicting with the non-cooperative property of face recognition). Regarding accuracy, the measurement of an object with $10\ cm$ depth along the z axis needs $10\ bits$ representation, whereas all intensity information are represented by $8\ bits$. Due to the above imbalance, we argue that it is not perfectly fair to compare 2D and 3D face recognition using such data, and it impedes the use of 3D and multi-modal (2D+3D) face recognition in practical scenarios.

Fortunately, the Kinect sensor reduces above problems by providing both 2D and 3D information simultaneously at interactive rates, where the practical 3D and 2D+3D face recognition systems are feasible for real-time and online processing [111]. Even though the overall sensing quality of Kinect is known better than the sensors using stereo vision or ToF, it is still much inferior than the laser scanning quality which is usually considered in 3D face recognition researches. In addition, it suffers from a few particular problems including data missing in "blind points" [158], relatively low depth resolution and accuracy, big noises with large depth transitions (i.e. at boundaries) and spatially calibrating and mapping of RGB and depth images [73, 74]. Therefore, the understanding of how such data degradation (in comparison to laser scanning data) can affect the performance of face recognition systems in the context of biometric metrics is an important issue for face recognition researchers.

In order to exploit the capability of Kinect to improve face recognition under occlusion conditions, in this section, we build a face database based on Kinect sensor. The proposed database consists of 936 shots of well-aligned 2D, 2.5D and 3D face data from 52 individuals taken by the Kinect sensor. It includes 9 different facial variations (including expressions, illuminations, occlusions, poses) and 2 sessions. We provide the benchmark evaluations using baseline algorithms (Eigenface and LBP for 2D and 2.5D recognition, and our 3D face recognition based on TPS warping parameters [58]). Score-level fusion of RGB and depth data are conducted and demonstrates significantly improved results. The comparative results between the proposed Kinect database and high quality 3D database (i.e. FRGC [123]) are also provided. In addition to its standard usage for face recognition, the proposed database can also be applied in many research tasks (such as facial demographic analysis and 3D face modelling).

In the following sections, we will first review the 3D face databases in the literature in section 5.2.2. Details of the organization and the acquisition of our Kinect face database are described in section 5.2.3. In section 5.2.4, we conduct a series of experiments on the proposed dataset, and thus study the performance of using Kinect data for face recognition.

5.2.2 Review of 3D Face Database

This section gives an up-to-date review of publicly available 3D face databases. A list of more generic face databases can be found at [4]. In comparison to the large number of published 2D face databases, the number of available 3D face databases is relatively small. TABLE 5.1 gives an overview of off-the-shelf 3D face databases with statistics of different properties. In the table, it is clear that most of existing databases (FRGC, ND-2006, GavabDB, BJUT-3D, UMD-DB and 3DTEC) adopts high quality laser scanners for face data acquisition. The 3D face acquisitions of BU-3DFE, BU-4DFE, XM2VTSDB, Texas 3DFRD, Bosphorus and UHDB11 are achieved by high quality stereo imaging systems, which can yield similar data accuracy in comparison to the data obtained by laser scanners. Although those high quality scanning systems can provide accurate facial details (e.g. the wrinkles and eyelids) for analysis, the capturing procedure is rather long and demands for careful user cooperation. On the other hand, only one 3D face database is captured by relatively low-quality 3D inference scheme. In 3D-RMA, 4000 points of each identity is obtained by structured light. This scanning scheme is similar to the one proposed by Freedman et al. [18] which is used in Kinect but developed in 1999. For a long time the 3D-RMA database has been the only publicly available database, although its quality is rather low and no texture mapping is provided. Description of the sensor used in U-York face database is not available.

A face database should include enough variations in order to test the robustness of face recognition algorithms. Facial variations such as facial expressions, illuminations, partial occlusions and pose variations are usually considered in the typical face databases [105, 123]. Because facial expression is a major challenge in 3D face recognition, most of the 3D face databases include this variation (except 3D-RMA, XM2VTSDB and UHDB11). The second

Table 5.1: Summary of Off-the-Shelf 3D Face Databases.

Database	Year of Publication	DB Size	No. of Subjects	3D Sensor	2D Texture	Exp.	Ill.	Occlusion	Pose	Video
FRGC (Ver. 2.0) [123]	2005	121589	466	Minolta Vivid 900/910 [3]	✓	✓	✓			
ND-2006 [60]	2006	13450	888	Minolta Vivid 910	✓	✓				
GavabDB [114]	2004	549	61	Minolta VI-700 digitizer	✓	✓			✓	
3D-RMA [5]	1998	360	120	Structured Light					✓	
XM2VTSDB [110]	1999	1180	295	Stereo Camera	✓					✓
U-York [6]	n.a.	5250	350	n.a.	✓	✓	✓		✓	
BU-3DFE [154]	2006	2500	100	3DMD digitizer [7]	✓	✓			✓	
BU-4DFE [153]	2011	606	101	Di3D digitizer [12]	✓	✓				✓
BJUT-3D [11]	2005	n.a.	500	CyberWare3030RGB/PS [8]	✓	✓				
Texas 3DFRD [69]	2010	1149	118	MU-2 stereo imaging system [119]	✓	✓				
UMB-DB [44]	2011	1473	143	Minolta Vivid 900	✓	✓		✓		
Bosphorus [130]	2007	3545	81	InSpeck 3D Digitizer [13]	✓	✓		✓	✓	
3DTEC [124]	2011	428	214	Minolta Vivid 910	✓	✓				
UHDB11 [15]	2011	1656	23	3DMD digitizer	✓		✓		✓	

commonly used variation in 3D face databases are different poses. The pose variation is relatively easier for 3D face recognition than 2D face recognition because the alignment of two 3D surfaces by rigid transformation usually via Iterative Closest Points method (ICP) [24, 40]). In FRGC, U-York and UHDB11, different illumination conditions are also considered, even if illumination usually does not affect 3D face recognition. In addition to the commonly considered variations, partial occlusion is a very challenging problem for both 2D and 3D face recognition but only few databases contains occluded faces [45, 105, 130]. The UMB-DB [45] and Bosphorus database in our review contain different types of occlusions for the evaluation of 3D face recognition under occluded conditions.

Almost all 3D face databases are also multi-modal face databases, since 2D texture images are also provided (except for 3D-RMA). For the databases using stereo vision technologies (BU-3DFE, BU-4DFE, XM2VTSDB, Texas 3DFRD, Bosphorus and UHDB11), 3D structure is inferred from two or more RGB cameras, thus the 2D/3D texture mapping is straightforward. However, for the databases based on laser scanners, the 2D texture images are usually taken by an external RGB camera. Therefore the 2D texture and 3D data are by default not aligned. The alignment can be achieved using additional facial landmarks and warping algorithms such as Thin Plate Spline (TPS) method [55]. Only XM2VTSDB and BU-4DFE in our review provide also video data. In XM2VTSDB, the video sequence is only for 2D texture sequences. In BU-4DFE, the 3D stream is captured in real time, so both the 2D and 3D video are provided. Nevertheless, it is worth to mention that most off-the-shelf 3D scanners are not able to provide 2.5D depth map or 3D points could in real time. Accordingly, most of the 3D face recognition studies are restricted to still 3D faces.

5.2.3 KinectFaceDB

In this section we will first describe in details the proposed database KinectFaceDB. Then we will explain how the Kinect sensor can be used to capture face data, so as to construct our database. Notable issues such as RGB and Depth images alignment are handled carefully following the procedures we suggested in Section 5.2.3.2.

5.2.3.1 Organization

There are totally 52 volunteers in the database, including 38 males and 14 females. Those participants are born between 1974 and 1987, come from different countries and thus belong to different ethnicities. We classify their ethnicities into the following categories (with the number of participants in the parentheses): Caucasian (21), Middle East/Maghreb (11), East Asian (10), Indian (4), African American (3) and Hispanic (3). We further distinguish the person with or without eye glasses, since this variation is of interest for glass removal/reconstruction task for both 2D and 3D faces [42, 121, 147]. The participants are asked to attend two sessions for the database recording. There are 5 to 14 days interval between session the two sessions, for which the same recording protocol is applied. This allows people to study the effects of time-elapsing with respect to different algorithms. A meta-data file including the information of gender / year of born / ethnicity / with or without glasses / capturing time for session 1&2 is associated with each identity. The demographic classification of the proposed database is shown in Figure 5.1.

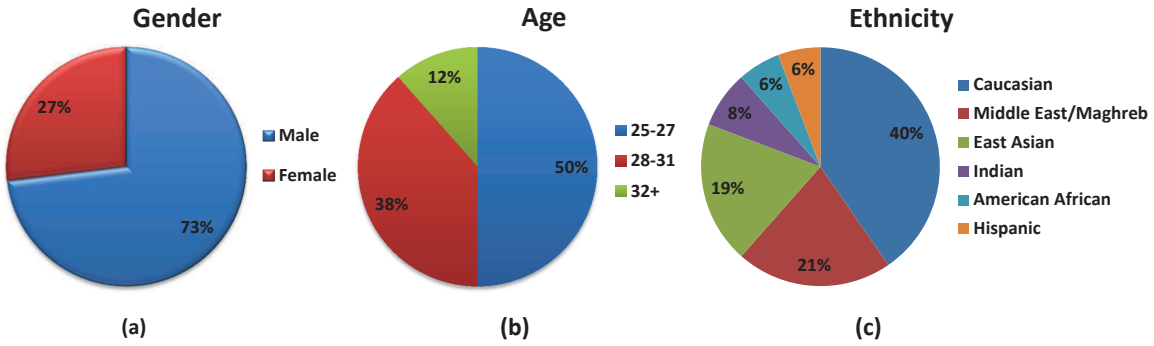


Figure 5.1: Demographics of KinectFaceDB validation partition by: (a) gender, (b) age, and (c) ethnicity.

In each session, three types of face data are captured: (1) 2D RGB image, (2) 2.5D depth map, and (3) 3D points cloud. We carefully selected 9 types of facial variations in both sessions: neutral face, smiling, mouth open, strong illumination, occlusion by sunglasses, occlusion by hand, occlusion by paper, right face profile and left face profile. Examples with above variations are illustrated in Figure 5.2. The images were taken under controlled conditions but no restraints on wearing (clothes, glasses, etc.), make-up, hair style etc. were imposed to participants.

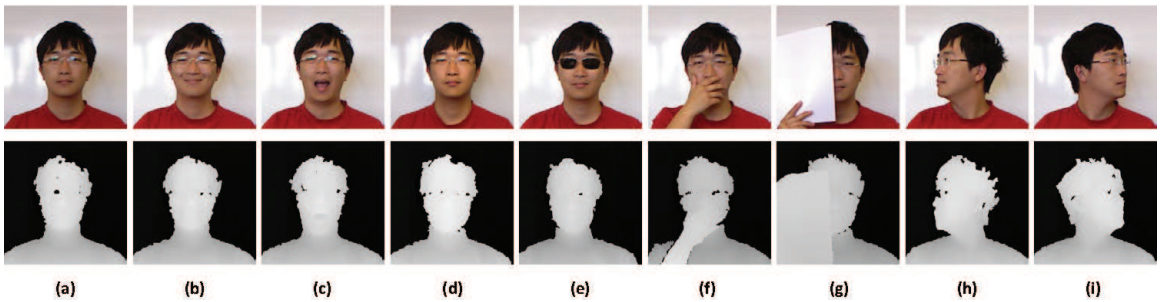


Figure 5.2: Illustration of different facial variations acquired in our database: (a) neutral face; (b) smiling; (c) mouth open; (d) strong illumination; (e) occlusion by sunglasses; (f) occlusion by hand; (g) occlusion by paper; (h) right face profile and (i) left face profile. (Upper: the RGB images. Lower: the depth maps aligned with above RGB images.)

Since we focus on the face region, we crop the captured RGB image and depth map into a pre-defined ROI (with resolution of 256×256, see images in Figure 5.2). The pre-cropping scheme ensures the captured faces have a simple/uniform background (the white board only, and therefore easy to segment); also, it minimizes the differences between the RGB image and Depth map after the alignment (see section 5.2.3.2).

To obtain the proposed database, we set up a "regular" indoor environment (nature light source at daytime, with normal indoor LED diffusions) for the database capturing. A Kinect is mounted and stabilized (where its lens are in parallel to ground floor by adjusting its tilt) on top of a laptop. The participants[2] were asked to sit in front of the Kinect sensor at a distance (ranging from $0.7m - 0.9m$) and to follow the pre-defined acquisition protocol (in compliance with the proposed database structure). A white board is placed behind the participants with the distance to Kinect at $1.25m$, so as to produce a simple and easily filtered background. A LED lamp is set in front of the participants to produce strong illumination variation. Three pairs of sunglasses and a white paper are used to produce the occlusion variations. A human operator is required to sit in front of the laptop (in the opposite position of the participants) to monitor and control the acquisition process. The acquisition environment is illustrated in Figure 5.3.



Figure 5.3: Acquisition environment for the Kinect face database.

[2] 52 PhD students from EURECOM: http://www.eurecom.fr/

5.2.3.2 Facial Images Acquisition using Kinect

Microsoft Kinect was primarily designed for multimedia entertainment purposes, where it requires an integration of different sensing/display/processing functionalities (for the data source ranging from video/audio to depth) for the targeted ambient intelligence. Since our goal focuses on the capturing of face data in RGB-D, in this section, we summarize the imaging procedure of Kinect, how we can obtain the 3D data, as well as the alignment of RGB and depth images so as to obtain the final output.

RGB and depth imaging from Kinect: As illustrated in Figure 5.4, a Kinect sensor includes 3 main components for RGB-D data capturing: an infrared (IR) laser emitter, a RGB camera and an IR camera. The RGB camera captures RGB images I_{RGB} directly, whereas the laser emitter and IR camera act together as an active depth sensor to retrieve the distance information from the scene.

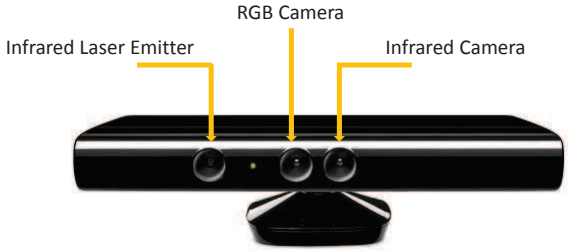


Figure 5.4: Architecture of a Kinect sensor for RGB-D sensing.

Freedman et al. introduced the triangulation process of Kinect for the depth measurement based on the IR laser emitter and the IR camera in [18]. In the system, a pattern of spots (created by transillumination through a raster) are projected to the world scene by the IR laser emitter and the reflection of such pattern is captured by the IR camera. The captured pattern is then compared with a reference pattern (a pre-captured plane at a known distance) so as to produce a disparity map $I_{Disparity}$ of the spots displacement d. From the obtained disparity map $I_{Disparity}$ it is straightforward to deduce a depth map via simple triangulation process. A simplified model to describe such triangulation is suggested in [88]:

$$z_{world}^{-1} = (\frac{m}{f \times b}) \times d' + (Z_0^{-1} + \frac{n}{f \times b}) \tag{5.1}$$

where z_{world} is the distance between Kinect and object (namely the depth, in the unit of mm), $d = m * d' + n$ since the raw disparity values are re-normalized between 0 and 2047, b and f are the base length and focal length respectively, and Z_0 is the distance from the reference pattern. The calibration parameters including b, f and Z_0 are estimated and provided by the device vendor (Microsoft).

With above procedures, Kinect outputs a RGB image ($I_{RGB}(x, y) = [v_R\ v_G\ v_B]$) and a depth map ($I_{Depth}(x, y) = z_{world}$) simultaneously, with the resolution of 640×480. On top of the obtained RGB and depth images, the 3D face data and aligned RGB-D face data can be generated with later processes.

Converting to 3D face data: thanks to the embodiment of Kinect, the calculation of 3D points representing the surface of a face is straightforward. Knowing the depth map $I_{Depth}(x, y) = z_{world}$, the 3D coordinates of each point $\{x_{world},\ y_{world},\ z_{world}\}$ (in the unit of $\{mm,\ mm,\ mm\}$) can be calculated from its image coordinates $\{x, y\}$ as below:

$$x_{world} = -\frac{z_{world}}{f}(x - x_0 + \delta x) \tag{5.2}$$

and

$$y_{world} = -\frac{z_{world}}{f}(y - y_0 + \delta y) \tag{5.3}$$

where x_0 and y_0 are the principal point of the image coordinates, and δx and δx represents the corrections for lens distortions. Those parameters are easily accessible from the device outputs.

Based on above projections, we compute the 3D coordinates (a 3D points cloud) of each pre-cropped face depth image. Illustration of the resulting 3D points cloud is shown in Figure 5.5.



(a) (b)

Figure 5.5: Illustration of the 3D face data: (a) visualization (rescaling to displayable range $[0, 255]$) of a depth map from the pre-cropped region; (b) the 3D points cloud retrieved from (a).

Alignment of RGB and depth images: in order to simultaneously exploit both RGB and depth information for facial image analysis, we need to know the correspondence between the RGB value and depth/3D values from the same location on a face. However, due to the natural instincts of Kinect's imaging facilities (RGB and depth are sampled separately from two different cameras with a displacement), the RGB image and depth map captured by Kinect are not well aligned. Therefore, we further project the depth value from the IR camera to the RGB camera plane. From the depth map, we have already estimated a 3D points cloud $\{x_{world},\ y_{world},\ z_{world}\}$, then the projection can be done by the traditional Tsai's camera model [139]:

$$\begin{bmatrix} x_{RGB'} \\ y_{RGB'} \\ NA \\ NA \end{bmatrix} = VD1/zPTR_zR_yR_x \begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \\ 1 \end{bmatrix} \tag{5.4}$$

where the depth value z_{world} are mapped to a new image location $(x_{RGB'},\ y_{RGB'})$. We then can denote the RGB-D representation of the Kinect data as $I_{RGB-D}(x_{RGB'}, y_{RGB'}) = [v_R\ v_G\ v_B\ z_{world}]$. Some previous works made additional efforts to calibrate the intrinsic and

extrinsic parameters of the RGB camera using chessboard based on an extra RGB camera [74] or the embedded IR camera [73]. In our database, we directly adopt the Kinect factory calibration parameters which can produce accurate enough alignment results. Illustration of the RGB-D alignment is shown in Figure 5.6. In the figure, one can observe the geometrical distortion (in addition to the shift at the image boarder) when re-mapping the depth values to the RGB image.



Figure 5.6: Illustration of the RGB-D alignment: the depth map (left) is aligned with the RGB image (right) captured by Kinect at the same time.

Because we apply a pre-cropping scheme (described in section 5.2.3.1), the large information loss and distortion at the image boarder of the aligned depth map are not included in our final output. Once we found the correspondence between RGB and depth values, it is straightforward to map the RGB color to the vertexes of the 3D points cloud. Visualization of 3D face points cloud with color information is given in Figure 5.7 (with background removal using a threshold τ).



Figure 5.7: Illustration of color mapping on 3D points cloud of a given face: from left to right views.

Facial landmarking: to facilitate the facial region extraction and normalization in face recognition, we defined 6 anchor points on a face (namely left eye center, right eye center, nose-tip, left mouth corner, right mouth corner and chin, see Figure 5.8.). Those anchor points are manually annotated for the RGB images. Then the corresponding locations in the depth images and 3D points clouds are also computed. Please note that the two face profiles are not annotated because only one side of the face is available. Even though the anchor points of occluded faces (faces occluded by sunglasses, hand and paper respectively) are not all visible, we estimated their positions so as to provide the full annotations on those occluded faces (similar to the annotations one can find in AR face database [105]).



Figure 5.8: The 6 facial anchor points: 1. left eye center, 2. right eye center, 3. nose-tip, 4. left mouth corner, 5. right mouth corner and 6. chin.

5.2.4 Experiments

To exploit the capability of Kinect for robust face recognition, we conduct a series of experiments on the proposed Kinect face database using several benchmark techniques (i.e. PCA, LBP and TPS warping parameters). The experiments are mainly comprised of 2 parts: (1) baseline evaluations and (2) comparing with FRGC [123]. In the baseline evaluations, we report the recognition (identification/verification) results for all sessions and variations in 2D, 2.5D and 3D modes. In the comparison part, we assess the 2.5D/3D data quality difference between the Kinect sensor and high quality laser scanner (faces in FRGC) in the context of face biometrics. Results in this section can help us to properly design face recognition algorithms based on the Kinect sensor, especially when the occlusion present (please refer to Section 5.3).

5.2.4.1 Baseline Evaluations

The evaluations on 2D, 2.5D and 3D face recognition are performed using several standard techniques to show the recognition performances of standard techniques on the proposed Kinect face database. The evaluations are done with three techniques (PCA [140], LBP [17] and TPS warping parameters [58] (which also serves as the baseline method for 3D face recognition evaluation in the European project Tabula Rasa (EU FP7) [9])) which use the 3D data (depth map and clouds point) and two techniques (PCA [140] and LBP [17]) which use the 2D data (texture images). Because the techniques which use the depth maps actually exploit the 2.5D shape information, we consider those techniques as 2.5D face recognition. Therefore the evaluation results are given in three groups: the evaluation on 2D, the evaluation on 2.5D, and the evaluation on 3D face recognition.

Pre-processing: because the RGB-D data in the database are already aligned, given the facial landmarks, face cropping and normalization are straightforward. Using the eye coordinates, we cropped, normalized and down-sampled the 2D and 2.5D faces (from intensity image and depth map, respectively) into 96×96 pixels.

The 3D face cropping is achieved by preserving the vertices in a sphere with radius $100mm$, centred $20mm$ away from the nose tip in $+z$ direction. Next, spikes are removed by thresholding and a hole filling procedure is applied (the holes and spikes are interpolated linearly to form a complete mesh, the values filling holes and spikes are estimated by taking the mean of valid pixels in 5×5 neighbourhoods). Finally, a bilateral smoothing filter is employed to remove white noise while preserving edges.

Evaluation protocol: we consider both identification and verification in our baseline evaluation (since our focus here is not on occlusion in particular). In both modes, we use the neutral faces from session 1 as the gallery. Recognition results of each variation (except left/right profiles, since sophisticated face alignment is required in 2D and 2.5D recognition of large pose variations) from both session 1 and session 2 are obtained separately. Then the overall identification/verification rates are reported for all the faces in both sessions. In our evaluation, the rank-1 identification rate and the verification rate where false acceptance rate (FAR) equals to 0.001 are reported.

Results of 2D face recognition: Table 5.2 and 5.3 show the face recognition results on 2D face data using PCA and LBP respectively. In the table 5.2, we observe that PCA based approach is not robust to large local distortions (e.g. extreme facial expressions and partial occlusions), since the local distortion can affect the whole representation in the face space. In contrast, as a locally-based description, LBP based method is robust to such variations to some extent. In all different variations, LBP yields better results than PCA, which corresponds to the results reported in the literature [17].

Table 5.2: 2D Face Recognition using PCA.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	96.15%	78.85%	90.38%	38.46%	73.08%	7.69%	64.10%
	Session 2	82.69%	78.85%	67.31%	73.08%	19.23%	51.92%	1.92%	53.57%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	96.15%	76.92%	84.62%	34.62%	51.92%	0%	58.01%
	Session 2	73.08%	61.54%	51.92%	55.77%	7.69%	26.92%	1.92%	40.11%

Table 5.3: 2D Face Recognition using LBP.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	100%	96.15%	100%	90.38%	100%	67.31%	92.31%
	Session 2	100%	98.08%	94.23%	100%	92.31%	94.23%	57.69%	90.93%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	96.15%	90.38%	96.15%	88.46%	92.31%	40.38%	75.96%
	Session 2	92.31%	82.69%	73.08%	88.46%	65.38%	67.31%	17.31%	59.34%

Results of 2.5D face recognition: Table 5.4 and 5.5 illustrate the evaluation results on 2.5D face data. Although many previous studies (according to [29]) suggested the use of PCA on 2.5D range images, in our experiment, we found that LBP yields better recognition rates than PCA in all different conditions (even if LBP was primarily designed for texture description). Therefore, our results suggest that local feature descriptors (such as LBP and its depth-specific variants [77, 78]) are more appropriate to represent and discriminate depth

face patterns. It should be noted that the use of 2.5D depth map gives worse results than the use of 2D intensity images (for both PCA and LBP). This is because the depth scanning quality by Kinect is relatively low. Nevertheless, the depth map supplies complementary information with respect to the intensity images, and therefore can be integrated into a multimodal face recognition system to generate higher recognition rates (one example is given in the following part for the fusion of RGB and Depth images of Kinect).

Table 5.4: 2.5D Face Recognition using PCA.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	84.62%	57.69%	76.92%	36.54%	7.69%	5.77%	44.87%
	Session 2	46.15%	42.31%	36.54%	30.77%	13.46%	5.77%	0%	25%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	67.31%	38.46%	48.08%	15.38%	0%	0%	17.95%
	Session 2	21.15%	15.38%	15.38%	17.31%	7.69%	0%	0%	8.52%

Table 5.5: 2.5D Face Recognition using LBP.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	94.23%	84.62%	96.15%	84.62%	65.38%	19.23%	74.04%
	Session 2	92.31%	73.08%	80.77%	94.23%	73.08%	38.46%	5.77%	65.38%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	75%	75%	78.85%	34.62%	19.23%	1.92%	41.03%
	Session 2	55.77%	34.62%	15.38%	34.62%	23.08%	11.54%	5.77%	26.92%

Results of 3D face recognition: in addition to the direct utilization of 2.5D range images, the adoption of parameters from rigid/non-rigid 3D surface registration algorithms (ICP or TPS) is another popular approach for 3D face recognition [16, 28]. Table 5.6 shows the 3D face recognition results using TPS warping parameters. In the table, we can observe that face recognition based on 3D points clouds generates inferior results than 2.5D face recognition based on LBP. The results demonstrate that although 3D face registration provides more accurate face alignment in recognition, it is inappropriate for low-quality 3D data, especially the faces with large local distortions. We argue that high-order features (e.g. LBP or curvature descriptors) could be potentially helpful to achieve more robust face recognition systems (before/after the 3D face registration) for low quality 3D data captured from Kinect.

Table 5.6: 3D Face Recognition using TPS warping parameters.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Overall
Rank-1 Identification Rate	Session 1	N/A	59.62%	47.06%	71.15%	38.46%	4.08%	44.53%
	Session 2	78.85%	46.15%	38.46%	67.31%	53.85%	0%	48.70%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	19.23%	23.53%	28.85%	7.69%	0%	12.11%
	Session 2	32.69%	17.31%	15.38%	28.85%	17.31%	0%	19.81%

Fusion of RGB and depth face data: to justify that Kinect is more helpful than sole-RGB based cameras for face recognition, we conduct an additional fusion step to combine both the RGB (2D) and Depth (2.5D) face information from Kinect in face recognition. The weighted sum fusion strategy is thus adopted for this purpose, where the z-score normalization is first

applied to the dissimilarity scores of both matchers. The weights in the fusion step are empirically selected based on a validation set. Table 5.7 and 5.8 illustrate the fusion results from RGB and Depth using PCA and LBP, respectively. From the results, it is clear that the fusion process significantly improve the results from sole RGB based face recognition (64.10% to 76.92% and 92.32% to 97.12% for the overall rank-1 identification in Session 1 for PCA and LBP based methods, respectively). This experiment demonstrate that 3D information provided by Kinect can greatly reinforce the recognition performance from its 2D images.

Table 5.7: Fusion of RGB and Depth for Face Recognition using PCA.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	96.15%	88.46%	100%	59.62%	78.85%	38.46%	76.92%
	Session 2	82.69%	82.69%	71.15%	90.38%	46.15	57.69%	30.77%	65.93%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	96.15%	84.62%	90.38%	30.77%	51.92%	0%	58.65%
	Session 2	75%	71.15%	46.15%	71.15%	5.77%	19.23%	1.92%	41.48%

Table 5.8: Fusion of RGB and Depth for Face Recognition using LBP.

Mode	Session	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	Overall
Rank-1 Identification Rate	Session 1	N/A	100%	98.08%	100%	92.31%	98.08%	94.23%	97.12%
	Session 2	100%	100%	98.08%	98.08%	94.23	94.23%	69.23%	93.41%
Verification Rate ($FAR = 0.001$)	Session 1	N/A	100%	92.31%	98.08%	90.38%	92.31%	15.38%	83.65%
	Session 2	96.15%	80.77%	71.15%	92.31%	63.46%	69.23%	19.23%	73.63%

5.2.4.2 Data Quality Assessment of KinectFaceDB and FRGC

It is straightforward to visually observe the 3D data quality differences between Kinect and a high quality laser scanner (e.g. Minolta, we give an example in Figure 5.9). The scanners' parameters also specify their accuracy and resolution, which indicate the significant data quality difference between Kinect and Minolta. Recently, additional efforts have been made to better understand Kinect's accuracy [88]. However, it is not obvious to understand the data quality differences in the context of face biometrics. Nevertheless, it is an essential issue to evaluate the identification/verification differences of 2.5D/3D faces captured by Kinect and Minolta, in order to allow the deployment of practical face recognition systems using Kinect by the state-of-the-art 2.5D/3D face recognition algorithms (whose results were reported on high quality laser scanners in the literature, notably on FRGC).

Following the same protocol as described in section 5.2.4.1, we tested various 2.5D/3D face recognition algorithms (PCA and LBP using 2.5D depth maps, TPS warping parameters (WPs) using 3D points clouds) on both KinectFaceDB and FRGC. For KinectFaceDB, we use all neutral faces in session 1 as the gallery faces and the corresponding neutral faces from session 2 as the probe faces. Similarly, we select 2 neutral faces from 2 different sessions of 198 subjects (from FRGC ver.1) to form the gallery and probe set of FRGC respectively. In both databases, the rank-1 identification rates and verification rates (FAR=0.001) are reported for comparison.

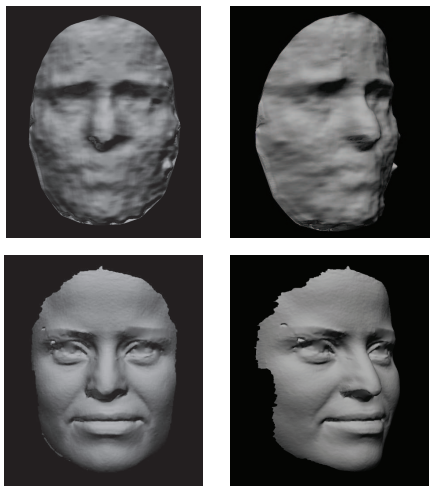


Figure 5.9: Cropped and smoothed 3D face captured by Kinect (upper row) and Minolta (lower row) of the same person, frontal view (left column) and side view (right column). The 3D face from Minolta keeps more details (wrinkles, eyelids etc.).

Table 5.9 shows the comparative results on KinectFaceDB and FRGC. In the figure, results on FRGC significantly outperform the results on KinectFaceDB for all three different approaches. It should be noticed that the results decreasing of Kinect in verification mode is much larger than the ones in identification mode. This suggests that Kinect is more appropriated for the non-cooperative face identification, whereas high quality laser scanner is more suitable for the cooperative verification (also due to the fact that the slow speed of laser scanner is not a significant problem in the cooperative case). For FRGC, the 3D WPs yields better recognition rates than 2.5D LBP and 2.5D PCA. However, for KinectFaceDB, 2.5D LBP achieves much better verification rate than 3D WPs. This is because data quality degradation also affects the sophisticated face registration procedures in 3D WPs, so that amplifies the result decreasing in the final recognition. This phenomena suggests that a simple yet efficient depth descriptors using 2.5D depth images is preferred for Kinect based face recognition in comparison to the methods based on registering of 3D points clouds.

Table 5.9: KinectFaceDB vs. FRGC

Mode	DB	2.5D PCA	2.5D LBP	3D WPs
Rank-1 Identification Rate	KinectFaceDB	46.15%	76.92%	78.85%
	FRGC	68.18%	93.94%	96.97%
Verification Rate ($FAR = 0.001$)	KinectFaceDB	21.15%	38.46%	32.69%
	FRGC	53.54%	81.82%	87.37%

Although the 2.5D/3D face recognition capability of Kinect is inferior than the ones of a high quality laser scanner, its intrinsic advantages make it a very competitive sensor for the face recognition task. We have identified its merits in our 3D face recognition work [111] (details can be found in Chapter 6), which can be summarized as follows: (1) it works in real time, which allows online face enrolment in non-cooperative scenarios; (2) its 3D data pro-

vides complementary information to 2D and can be easily integrated with the intensity based methods for more accurate multi-modal face recognition; (3) with a short period of capturing (a few seconds), 3D face recognition in video is feasible and yields significantly improved recognition rates comparing with still 3D face recognition. With the increasing interests on the Kinect sensor, we argue that newly developed 2.5D/3D or 2D+3D face recognition algorithms should not only be tested on databases with high quality scans (e.g. FRGC), but also on the more challenging KinectFaceDB, so as to make their methods more robust and reliable in more practical scenarios.

5.3 Depth Assisted 2D Face Recognition Under Partial Occlusions

Based on the results we have obtained in the last section (identification/verification rates on KinectFaceDB), we have a few interesting observations as following: (1) Intensity/RGB images yield much better results than depth/3D data for face recognition based on Kinect; (2) partial occlusion is a very challenging problem which can significantly deteriorate face recognition results (much more than the other facial variations); (3) simple and efficient texture descriptors (such as LBP) using 2.5D depth images is more appropriated for Kinect face data.

From above observations we know that depth from Kinect has insufficient discriminative power for robust face recognition. However, even if the depth information captured from Kinect is relatively noisy and incomplete, we argue that it could be very helpful for facial occlusion analysis. In this section, we present a new approach to address the partial occlusion problem in face recognition using the Kinect sensor. Traditional methods applying occlusion analysis to improve face recognition exploit homogeneous information in both steps (2D→2D or 3D→3D). Instead, we use heterogeneous cues (depth→RGB) to improve face recognition in presence of occlusions. The proposed approach first conducts occlusion analysis based on depth image, then use this information to improve LGBP based face recognition [157] (using intensity image) via a weighting strategy. Experiments on the proposed Kinect-FaceDB demonstrate that significant improvements are achieved in comparison to LGBP and KLD-LGBP [156] in various occlusion conditions.

5.3.1 Overview

Recently, explicit occlusion analysis is found that it can further improve the local representation based algorithms, because information from the occluded part of a face can be reduced/excluded. In [116] and [113], binary occlusion detection are performed in local patches to improve LNMF and LBP based face recognition respectively. A fuzzy occlusion detection scheme based on Kullback–Leibler divergence is incorporated in LGBP based face

recognition for occluded faces, namely KLD-LGBP [156]. Significant improvements have been observed when explicit occlusion analysis is integrated into face recognition process.

Most of the existing facial occlusion analysis algorithms rely on the assumption that the intensity distributions of face appearance and potential occlusions are largely different. However, this assumption may not stand when the appearance of occluding object is similar to the face appearance (for example, hand on a face). Instead, structural information (depth or 3D points) can be a useful cue to analyse the presence of occlusion on a face surface. In [43], Colombo et al. proposed to detect and restore partial occlusions in the principal face subspace based on depth maps to improve 3D based face recognition.

Nevertheless, it remains unclear whether 3D structure can help more than 2D appearance to detect occlusions to improve face recognition, or vice versa. The main difficulty relies on the simultaneous acquisition of both 2D and 3D images from a face, due to the costly processing of the traditional 3D imaging/sensing devices. Thanks to the recent success of low-cost RGB-D cameras, such as Kinect/PrimeSensor, RGB and depth face images can be accessed at interactive rates.

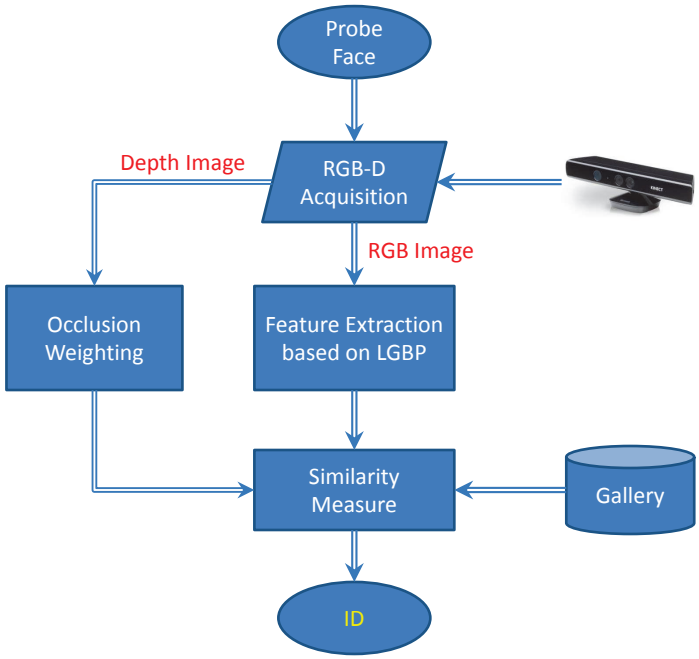


Figure 5.10: Overview of the proposed solution for face recognition under occlusion conditions based on Kinect sensor.

We propose an approach to estimate the occlusion probability of each region of a face using the 3D information (depth map[3]) to improve LGBP based face recognition (using RGB image) based on Kinect. The overview of the proposed solution is shown in Figure 5.10. In the figure, a probe face is captured by the Kinect sensor, then its RGB component is used

[3]Although depth map is 2.5D representation, in this section, we consider it as 3D information to avoid ambiguity.

for feature extraction and its depth component is used for occlusion analysis. By combining the estimated occlusion probability in the matching step, the identification can be achieved by reducing the effect of occlusion based on features extracted from the RGB/intensity component of a face. In comparison to KLD-LGBP which exploits the intensity information for occlusion analysis to improve LGBP based face recognition [156], our method yields better recognition results. Comparative results on KinectFaceDB also reveal the superiority of 3D cue to 2D cue for occlusions analysis in face recognition, in various testing scenarios.

In the following parts, we will give a more detailed look (in regarding to our brief review in Section 2.3.1.3) onto the literature work KLD-LGBP in Section 5.3.2, which exploits RGB/intensity information for occlusion analysis to improve face recognition. Then the proposed approach which uses depth cue for occlusion analysis to improve face recognition is presented in Section 5.3.3. Finally, experimental results and analysis are given in Section 5.3.4.

5.3.2 RGB based Occlusion Analysis to Improve Face Recognition

Almost all works in the literature [113, 116, 127, 156] who use explicit occlusion analysis to improve occluded face recognition (please refer to our review in Section 2.2, Chapter 2) are based on RGB/intensity image (which is also used for face recognition). The proposed approach is a variant of LGBP based face recognition, and similar to the KLD-LGBP method. For fair comparison, we firstly review the method KLD-LGBP in this section.

Local Gabor binary patterns (LGBP) has been proven to be an efficient feature descriptor for face recognition, especially for faces with pose and occlusion variations [157]. Explicit occlusion analysis based on Kullback–Leibler divergence (KLD) can be incorporate into the matching process by a weighting scheme to weaken the significance of occlusion in the classification, which is called KLD-LGBP [156]. Details of both methods are described in the following parts.

5.3.2.1 LGBP for face representation

The LGBP descriptor consists of first applying Gabor wavelet decomposition (more detailed information can be found in Section 3.2.1.1, Chapter 3) and then computing LBP codes from the filtered image. Given a target face image I, its LGBP codes can be computed as below:

$$LGBP^{\mu,\gamma}_{P,Q}(I(x,y)) = \sum_{p=0}^{P-1} s(G(I(x,y)_p)^{\mu,\gamma} - G(I(x,y)_c)^{\mu,\gamma})2^P \tag{5.5}$$

where $G(I(x,y))^{\mu,\gamma} = I(x,y) * \psi_{\mu,\gamma}$ denotes the filtering, and $\psi_{\mu,\gamma}$ is the Gabor wavelet with μ scales and γ orientations. As suggested in [156, 157], five scales and eight orientations are used. Then the binary code can be locally calculated by differentiating the central pixel $I(x,y)_c$ from its P equally spaced (at radius Q) neighbours $I(x,y)_p$.

The LGBP operator generates $\mu * \gamma$ LGBP images from a single input. To form the final face representation, each LGBP image is firstly divided into R local regions, from which local histograms are extracted to summarize the LGBP codes. Finally, all local histograms of different LGBP images are concatenated into one feature vector to build the global representation H.

5.3.2.2 Occlusion weighting based on LGBP representation

Since the LGBP histogram actually represents the distribution of local textures on a face, when a region r is occluded, its LGBP histogram should largely deviated from the ones statistically estimated from the same region in face images without occlusion. Considering the mean histogram $\overline{h}_{\mu,\gamma,r} = \frac{1}{N}\sum_{n=0}^{N-1} h^n_{\mu,\gamma,r}$ of the gallery set (N faces without occlusion) as the "normal" distribution, the occlusion probability of a probe histogram $h'_{\mu,\gamma,r}$ is estimated as the KL-divergence [91] from $\overline{h}_{\mu,\gamma,r}$:

$$P^{RGB}_{\mu,\gamma,r} = \| \sum_{i=0}^{L-1} h'_{\mu,\gamma,r}(i) \log \frac{h'_{\mu,\gamma,r}(i)}{\overline{h}_{\mu,\gamma,r}(i)} \| \tag{5.6}$$

where $\|\cdot\|$ indicates the normalization that filters small deviations, L is the histogram length. Since the occlusion probability are estimated from RGB/intensity image, we denote it as P^{RGB}.

5.3.2.3 Face recognition

Once the LGBP representation are extracted for both probe and gallery faces, the normal way to conduct face recognition (identification) is to measure the distances between the LGBP histograms from a probe and all gallery faces. However, direct matching may include the information from the undesirable part when partial occlusion is in presence on the probe face.

Instead of direct matching of LGBP feature vectors between a probe face and the gallery faces, KLD-LGBP associates a set of weights (which are inversely proportional to the occlusion probabilities) to local histograms for the computing of global similarity between two faces as below:

$$S(H^1, H^2) = \sum_{\mu=0}^{4}\sum_{\gamma=0}^{7}\sum_{r=0}^{R-1}(1 - P^{RGB}_{\mu,\gamma,r}) dist(h^1_{\mu,\gamma,r}, h^2_{\mu,\gamma,r}) \tag{5.7}$$

where $P^{RGB}_{\mu,\gamma,r}$ is the occlusion probability of (μ,γ,r)th local histogram. Face recognition can be then conducted based on this similarity measure. Experiments in [156] has demonstrated that this weighting strategy can efficiently improve the recognition results for faces with partial occlusions.

5.3.3 Depth based Occlusion Analysis to Improve Face Recognition

Although the occlusion probability estimated from the intensity/RGB image of a face can already improve face recognition in occlusion conditions, we believe that the structural cue (when available) can be more helpful to reveal the presence of occlusion so as to further increase the recognition rates. Thanks to the RGB-D sensing capability of Kinect, we propose a method to estimate the occlusion probability from the depth image and facilitate face recognition based on LGBP features extracted from the RGB image.

5.3.3.1 RGB-D alignment

To use the depth information for occlusion analysis to facilitate RGB based face recognition, we need to know the correspondences between the RGB values and depth values from the same location on a face. One notable problem to use the RGB and depth images from Kinect is that they are not aligned because of the difference in positions of the IR camera and RGB camera. We then first align the depth image with the corresponding RGB image via the following projections:

$$I'_{Depth} = F_{3D \rightarrow RGB}(F_{Depth \rightarrow 3D}(I_{Depth})) \tag{5.8}$$

where $F_{Depth \rightarrow 3D}$ converts the depth image into 3D points cloud, and $F_{3D \rightarrow RGB}$ denotes the perspective projection from 3D points cloud to the RGB camera plane; the intrinsic and extrinsic calibration parameters are provided by the device vendor (more details of Kinect acquisition of RGB-D data can be found in Section).

5.3.3.2 Occlusion weighting based on depth

Because Gabor wavelet and LBP are both texture descriptors, it is inappropriate to directly use LGBP to estimate the occlusion probability for depth images (as described in Section 5.3.2.3). We thus propose an occlusion estimation method based on the observation that a face surface is behind its occluding object. In other words, an occluded depth pixel has higher value than the one without occlusion at the same location of the same face. However the depth value is unknown for the face surface which is occluded. We therefore assume that the depth value of an occluded pixel is statistically deviated from the ones without occlusion. To support this assumption, we first align the faces to have the same depth at their nose tip (assuming the pixel at nose tip is not occluded). Then a mean model (similar to the one used in KLD-LGBP) is built to act as the "normal" depth value at the pixel location (x, y):

$$\overline{I^D}(x, y) = \frac{1}{N} \sum_{n=0}^{N-1} I^{n,D}(x, y) \tag{5.9}$$

where $\overline{I^D}$ is the mean depth map of the gallery set (faces without occlusion). Figure 5.11 shows an example of the mean depth model for our occlusion estimation.



Figure 5.11: The mean depth model for occlusion estimation.

Given the depth map of a probe face I^D, we compute its deviation from the mean depth model as below:

$$\Phi = \begin{cases} T_{upper} & I^D - \overline{I^D} \geq T_{upper} \\ I^D - \overline{I^D} & \\ T_{lower} & I^D - \overline{I^D} \leq T_{lower} \end{cases} \tag{5.10}$$

where T_{lower} is the lower bound to exclude small facial structure variations and T_{upper} is the upper bound can prevent the confusing with background. These thresholds are empirically estimated using a validation set. Φ is regarded as an occlusion map which indicates the deviation of a probe face from the "normal" facial structure, so as to reflect the structure of occlusion on the probe. Examples of estimated occlusion map can be seen in Figure 5.12 (c). (As also observed in the figure, IR camera is well known to has limitations to sensing glass [132].)

Based on the occlusion map, the occlusion probability for local region r is obtained by simple summation:

$$P_r^D = \| \sum_{k=0}^{K-1} \Phi_r(k) \|_1 \tag{5.11}$$

where $\| \cdot \|_1$ indicates the l1 normalization, K is the number of pixels in a local region r. In contrast to P^{RGB} in KLD-LGBP, we denote the estimated occlusion probability using depth as P^D. Figure 5.12 (d) shows the occlusion probability estimated using our method.

5.3.3.3 Face Recognition

Similar to KLD-LGBP, we incorporate the occlusion probability P^D into the similarity measure as follows:

$$S'(H^1, H^2) = \sum_{r=0}^{R-1} (1 - P_r^D) \sum_{\mu=0}^{4} \sum_{\gamma=0}^{7} dist(h^1_{\mu,\gamma,r}, h^2_{\mu,\gamma,r}) \tag{5.12}$$

Please note that the LGBP face representation H (obtained by the method described in Section 5.3.2.1) used in our approach is extracted from the intensity images from the RGB camera of Kinect. This is because the depth image of Kinect has less discriminative power than the intensity image, based on our experiments in Section 5.2.4. On the other hand, the depth



Figure 5.12: Illustration of the proposed occlusion probability estimation: (a) the intensity image; (b) the depth image; (c) the computed occlusion map from (b); (d) visualization of the occlusion probabilities for different local regions. Row 1 to row 3 corresponds to 3 different types of occlusions (sunglasses, hand and paper).

information is more suitable to detect occluding objects, even if it is noisy and incomplete as in our case.

5.3.4 Results

To evaluate the proposed approach, we performed a series of experiments on EURECOM KinectFaceDB. The dataset and detailed configurations of our experiments are introduced in Section 5.3.4.1. In Section 5.13, we will illustrate recognition results for LGBP [157], KLD-LGBP [156] and our method in different scenarios, including occlusion analysis in 2D and 3D, together with face recognition using either 2D or 3D information for faces with different types of occlusion. The proposed approach yields the best results in most cases. The obtained results also support that 3D information has more advantages over 2D for facial occlusion analysis in face recognition.

5.3.4.1 Dataset and Configurations

We conduct experiments on the built EURECOM KinectFaceDB (see Section 5.2 for details). KinectFaceDB consists of well aligned multi-modal facial images of 52 people (14 females, 38 males) obtained by Kinect. The data is captured in two separate sessions with interval of about half month. Nine different types of variations are associated with each people, including different facial expressions, poses, illuminations and partial occlusions. Using the

eyes and nose coordinates, we cropped, normalized and down-sampled the face region (from both RGB and depth images) into 96×96 pixels. In our experiment, 156 non-occluded faces (52 people with 3 expression variations) of session 1 are selected as gallery faces; 52×3 faces occluded by sunglasses, hand and paper (see Figure 5.12 (a)(b)) from the same session are selected to form the probe set.

For face recognition, the face image is divided into 8×8 local regions. $P = 8$ neighbours at radius $Q = 2$ are used to calculate the LGBP codes. Chi-square distance (χ^2) is adopted in the similarity measure for LGBP representations.

5.3.4.2 Results

In this section, we report the results of LGBP, KLD-LGBP and our proposed method on the proposed dataset. Since the occlusion probability estimation used in KLD-LGBP is based RGB image, we denote the method as PRGB-LGBP. Similarly, because we use depth for occlusion analysis, the proposed method is denoted as PD-LGBP. The three different approaches (LGBP, PRGB-LGBP and PD-LGBP) are tested on both RGB and depth images for face recognition.

Figure 5.13 shows the recognition rates of LGBP, PRGB-LGBP and PD-LGBP for 2D based face recognition. Without explicit occlusion analysis, it is clear that occlusion by a piece of paper is the most difficult occlusion for face recognition, since it occludes the largest amount of face appearance comparing to sunglasses and hand. For PRGB-LGBP, we can observe that it can significantly increase the recognition rates for faces occluded by sunglasses and paper, since their appearances are very different from the face appearance. However, it cannot handle faces occluded by hand, because skin color of hand and face are similar to each other. PD-LGBP returns the best overall recognition rates (from all 3 occlusion types), but worst result for faces with sunglasses. This is because the depth sensing of Kinect is inaccurate for reflectance layers such as sunglasses. We further report the face recognition results based

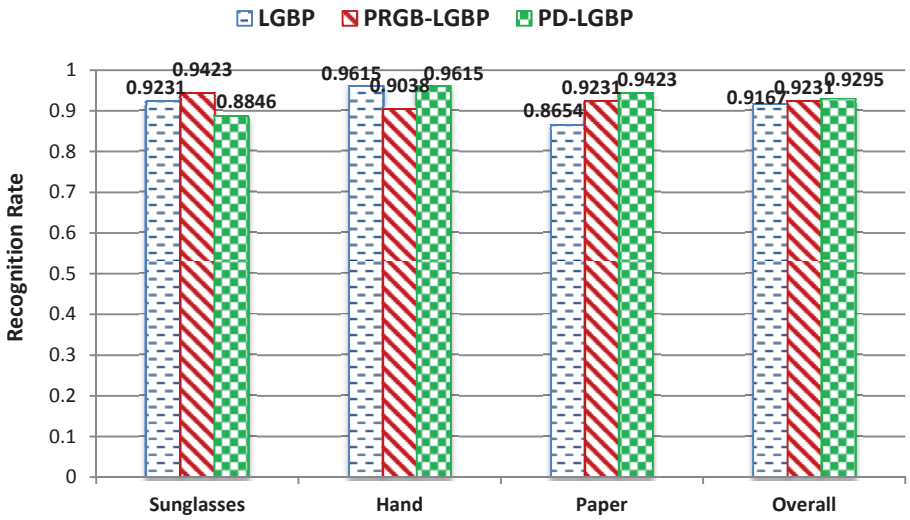


Figure 5.13: Recognition rates of LGBP, PRGB-LGBP and PD-LGBP on RGB images.

on 3D information (LGBP extracted from the depth image) in Figure 5.14. In the figure, PD-LGBP achieves the best results for all types of occlusions. The upgrade of recognition rates for PD-LGBP is more obvious, since LGBP yields much lower rates on depth images. This is because the sensed depth by Kinect has relatively low quality, but it is still very helpful to detect the presence of occlusions.

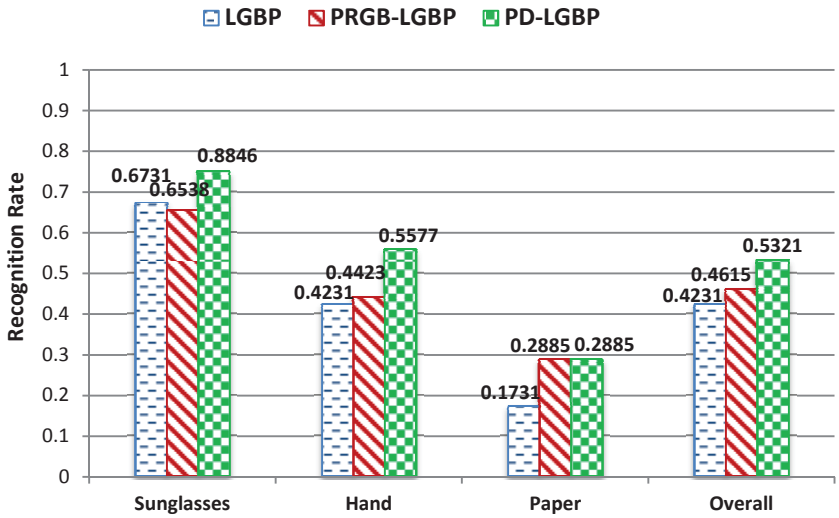


Figure 5.14: Recognition rates of LGBP, PRGB-LGBP and PD-LGBP on depth images.

From above observations in both cases, we can conclude that even with certain limitation (i.e. sensing error for sunglasses), the proposed occlusion analysis based on depth is more powerful than the one based on intensity of KLD-LGBP for face recognition. In certain cases, 3D information has more advantages over 2D information to reveal the presence of occlusion. As a sensor provides both RGB and depth images, Kinect is more helpful to address the occlusion problem in face recognition in comparison to the traditional 2D cameras.

5.4 Conclusions

In this chapter, we revealed the capability of Kinect for face recognition, especially for the scenarios when partial occlusion exists. In order to achieve this goal, a complete multi-modal (including well-aligned 2D, 2.5D, and 3D based face data) face database based on Kinect sensor is constructed. The processes of how to obtain the well aligned and processed 2D, 2.5D, and 3D face data are thoroughly introduced. We highlight the advantages of the proposed KinectFaceDB (as well as Kinect based face recognition) via the review of existing 3D face databases and extensive experimental evaluations. Standard face recognition techniques (including PCA, LBP and TPS warping parameters) are applied on different data modalities (including 2D, 2.5D and 3D based face data) so as to provide the benchmark evaluation for subsequent researches. Comparative study (in the context of biometrics) of KinectFaceDB and the state-of-the-art FRGC database is provided, which can guide the deployment of existing algorithms and the development of new face recognition methods towards more prac-

tical systems. To conclude, the proposed KinectFaceDB supplies a standard medium to fill the gap between traditional face recognition and the emerging Kinect technology.

Based on the proposed KinectFaceDB, we also presented our solution to address the partial occlusion problem in face recognition based on Kinect sensor. Instead of detecting the presence of occlusion using intensity cue, we proposed to estimate the occlusion probability based on depth cue, so as to improve face recognition using intensity images. We demonstrated via experiments that depth cue is more suitable to uncover the occlusion structure on a face than intensity cue in most cases.

As a future work, it is necessary to revisit the literature 3D and 2D+3D face recognition algorithms (which were mostly developed from high quality 3D face data) using the proposed KinectFaceDB for achieving reliable, robust and more practical face recognition system using Kinect. The design of new algorithms and new facial descriptors for the low-quality 3D data is another important topic to investigate. In addition, for facial occlusion analysis, both depth cue and intensity cue from Kinect can be simultaneously exploited, so as to find an optimal strategy which can improve face recognition in presence of all types of partial occlusions. The idea of using depth for facial analysis to improve RGB based face recognition can be directly extend to other structural facial variations, such as facial expressions.

Chapter 6

Improving Baseline Face Recognition Methods

6.1 Introduction

In previous chapters, we have already proven that explicit facial occlusion analysis and processing (based on state-of-the-art computer vision and image processing techniques) can significantly improve the recognition for faces in presence of different types of facial occlusions (dense/static/sparse/dynamic). In Chapter 3, Chapter 4, and Chapter 5, we mainly focused on the occlusion analysis and processing part, and illustrated how such methods can be efficiently incorporated into standard face recognition algorithms (e.g. LBP [17], LGBP [157], PCA [140], SIFT [102]) to obtain the robustness to occlusion. Inherent from the natural instinct of this framework, if we can further improve the standard face recognition algorithms, we can obtain not only higher recognition rate for normal face recognition but also improved results for faces under occlusion conditions (when our explicit occlusion analysis & processing is included).

In this chapter, we present 2 new methods to improve the baseline face recognition based on 2D (LBP and SRC [146]) and 3D (ICP [109]), respectively. The first method improves 2D face recognition by efficiently combining two most popular (in recent years) face recognition techniques, namely Local Binary Patterns and Spares Representation based Classification. In the second method, we improve the standard 3D face recognition technique based on Iterative Closest Points by incorporating an face registration architecture via multiple intermediate references. Undoubtedly, improving of standard face recognition algorithms as in our case can also improve the robustness of a system to occlusion when explicit occlusion analysis & processing is considered.

The choice of LBP and SRC for improving 2D face recognition is based on the fact that they have become both eminent techniques in face recognition. Preliminary techniques of combining LBP and SRC have been proposed in the literature. However, the state-of-art method

[37] suffers from the "curse of dimensionality" for real world scenarios. In this chapter, we propose a novel face recognition algorithm of combining LBP with SRC; in which the dimensionality problem is resolved by divide-and-conquer and the discriminative power is strengthen via its pyramid architecture. The proposed face recognition method is evaluated on AR Face Database and yields impressive results. In addition, thanks to the locally emphasized pyramidal architecture, occlusion detection as we described in Chapter 3 can be easily combined in the occlusion conditions.

According to the recent review of Spreeuwers [134], recognition based on ICP is the basic yet the most popular technique for 3D faces. The essential step of such face recognition algorithms is the face registration step using ICP (which minimizes of the distance between 2 points clouds via rigid transformation). However, the ICP process is known to be computational expensive. Instead of registering a probe to all instances in the database, we propose to only register it with several intermediate references, which considerably reduces processing, while preserving the recognition rate. To validate the proposed face registration architecture, we implemented a real-time 3D face identification system using a consumer level depth camera (Kinect/PrimeSensor). Our system takes a noisy sequence as input and produces reliable identification. The presented system routinely achieves 100% identification rate when matching a (0.5-4 seconds) video sequence, and 97.9% for single frame recognition. These numbers refer to a real-world dataset of 20 people. The methodology extends directly to very large datasets. The process runs at 20 fps on an off the shelf laptop. The scalability (to occlusion) of the proposed system relies on the direct matching in the original depth pixel space, which is readily incorporated with the occlusion weighting strategy we presented in Chapter 5.

The rest of this chapter is structured as follows. The improved 2D face recognition algorithm based on the combination of LBP and SRC is presented in Section 6.2. Then the proposed face registration method to improve 3D face recognition is introduced in Section 6.3. After showing the improvements in both the 2D and 3D cases based on extensive experiments in each section, we draw conclusions in Section 6.4.

6.2 Improving 2D Face Recognition via Combination of LBP and SRC

In this section, we present the proposed 2D face recognition algorithm based on the combination of LBP and SRC. We first motivate and briefly introduce the key idea of the proposed method in Section 6.2.1. Then tools used in the proposed approach (namely LBP and SRC) are reviewed in Section 6.2.2. The proposed algorithm is detailed in Section 6.2.3. And Section 6.2.4 demonstrates the efficiency of the proposed approach via experimental results and analysis. Results show that the proposed algorithm outperforms both the original LBP and SRC based methods, as well as the other combination options.

6.2.1 Overview

LBP is known to be a powerful feature extractor for face representation [17]. The success of LBP in face description is due to the discriminative power and computational simplicity of the operator, and its robustness to monotonic gray scale changes caused by, for example, illumination variations. The use of histograms to collect features also makes the LBP approach robust to face misalignment and pose variations. So far many extensions of LBP for face recognition have been proposed, for example (not an exhaustive list): boosting LBP [17], LGBP [157], Multi-Scale LBP [98], MB-LBP [97], CLBP [68] etc. All those extensions target on improving the robustness and accuracy in non-optimal face recognition scenarios.

Recently, Wright et al. [146] introduced a framework for robust face recognition via sparse representation. Here face recognition is casted as penalizing the $l1$-norm of the coefficients in the linear combination of an overcomplete face dictionary. Sparse representation based classification (SRC) has been demonstrated to be superior to nearest neighbour (NN) and nearest subspace (NS) based classifiers in various subspaces (e.g. PCA or LDA). When applied to face recognition, it can also be efficiently customized to handle errors due to occlusion and corruption. Following Wright et al.'s work, in the past year, several extensions of SRC based face recognition were proposed. In [160], Zhou et al. applied a Markov Random Field model to SRC based face recognition for improving performances under severe contiguous occlusion. Yang and Zhang [150] used image Gabor-features for SRC in order to reduce the cost in coding occluded faces meanwhile improving accuracy. In [148], Yang et al. reviewed five representative l1-minimization methods in the context of SRC based face recognition.

Lately, a preliminary tentative of combining LBP based features with SRC for face recognition is presented by Chan and Kittler [37]. The authors illustrated that histogram descriptors, such as LBP, local phase quantization (LPQ) and Gabor phase pattern (GPP), are more robust to misalignment and illuminations than the holistic features used in SRC. Their approach returns impressive results on the combination of Yale Face Database B and extended Yale Face Database B [92] (38 people under 64 different illumination conditions). Nevertheless, such an approach is infeasible for most real world datasets where only few samples are available for each subject. The reason is due to the "Curse of Dimensionality". In SRC, the ability of discrimination relies on computing the correct sparse solution of an underdetermined system of linear equations. But the solution is non-sparse when the number of non-zero entries in the coefficients vector beyond the equivalence breakdown point (EBP) [53]. In other words, the correct sparse solution can only be recovered when the number of training samples is sufficiently larger than the number of features. Unfortunately, the dimension of LBP histogram for face representation is generally huge (16384 in an ordinary 8×8 block division). Thus the direct combination of LBP and SRC is non-realistic. (In [37], the authors tested only for 2×2 sub-blocks but with 722 training images.)

One possible solution is to reduce the dimension of extracted features before performing SRC by applying dimension reduction tools (e.g. PCA or LDA as suggested by Yang et al. [149]). It ensures the desired solution is sparse and thus the "curse of dimensionality" is no longer a

problem. Such an approach can slightly improve the performance comparing to the baseline algorithm [17] according to our experiments.

Here we propose a more powerful approach to combine LBP with SRC for face recognition. In our approach, the dimensionality problem is resolved by the divide-and-conquer scheme [136], where the SRC is performed on LBP histogram extracted from a single sub-block. The obtained sparse coefficients vectors (SCV) from all sub-blocks are then fused together to yield the final output. Besides, our approach has a pyramid architecture which is able to incorporate information from different levels of descriptions. The architecture thus can improve the robustness to various localized variations (inspired by Modular Eigenface [15]). In the experiments, we built a more realistic dataset (comparing to [37]) with different variations (including facial expressions, illumination conditions and time elapses) with fewer training samples using the AR face database [106]. On this dataset, the proposed algorithm is compared with the baseline method and the methods using dimension reduction tools. Our approach yields the best recognition rate (up to 96%).

6.2.2 Background and Related Algorithms

In this section, necessary knowledge of the tools used in our algorithm is given. First, the LBP based face representation presented in [17] is reviewed in Section 6.2.2.1. Then the algorithm of face recognition using sparse representation [146] is reviewed in Section 6.2.2.2.

6.2.2.1 Local Binary Patterns based Face Representation

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighbourhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas etc. The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{6.1}$$

where g_c corresponds to the gray value of the center pixel (x_c,y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities.

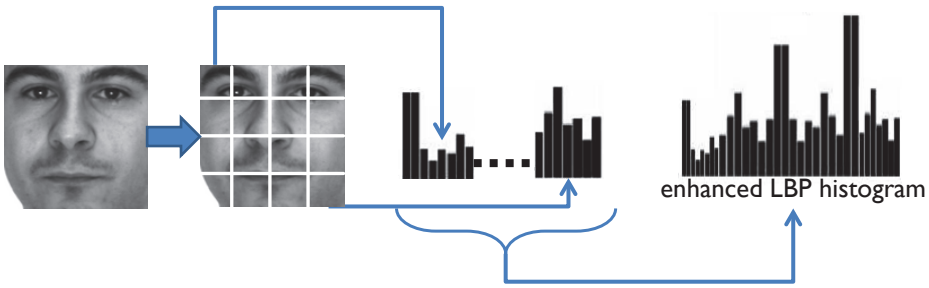


Figure 6.1: The procedure of block based face representation.

As suggested in [17], in order to retain the spatial information, a facial image is divided into K non-overlapping regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Figure 6.1 visualizes the procedure of how to compute the block based face representation. When a probe face is input into the face recognition system, such an enhanced LBP histogram is computed, and then the histogram similarity between the probe face and all template faces are measured through Chi-square distance (χ^2).

6.2.2.2 Sparse Representation based Face Classification

Supposing a training set A consists of the facial images from k classes, where $A = \{A_1, A_2, \ldots, A_k\}$. Ideally, giving sufficient training samples of class i, where $A_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,n_i}\} \in R^{m \times n_i}$, a test facial image $y \in R^m$ belongs to the same class could be well approximated by a linear combination of the training samples from A_i, which can be written as:

$$y = \sum_{j=1}^{n_i} a_{i,j} v_{i,j} \tag{6.3}$$

Since A is the dictionary which includes all the training samples, where $A = \{v_{1,1}, v_{1,2}, \ldots, v_{k,n_i}\}$. Then Equation 6.3 can be rewritten in the form as below:

$$y = Ax_0 \in R_m \tag{6.4}$$

where $x_0 = \{0, \ldots, 0, a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}, 0 \ldots, 0\}^T$ is the coefficient vector in which most coefficients are zero except the ones associated with class i.

Due to the fact that a valid test sample y can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others, to find the identity of y then equals to find the sparsest solution of Equation 6.4. This is the same as solving the following optimization problem ($l0$-minimization):

$$\hat{x}_1 = arg\min \|x\|_0 \qquad subject\ to\ Ax = y \tag{6.5}$$

However, solving the $l0$-minimization of an under-determined system of linear equations is NP-hard. In the case for large number of training samples, it equals to find the minimal

$l1$-norm solution [53]. Therefore, the SRC procedure presented in [146] is shown as below (see Algorithm 2).

Algorithm 2 The SRC Algorithm

1: Normalize the columns of A to have unit $l2$-norm;
2: Solve the l1-minimization problem:

$$\hat{x}_1 = arg\min \|x\|_1 \qquad subject\ to\ \|Ax - y\|_2 \leq \epsilon \tag{6.6}$$

3: Compute the residuals by:

$$r_i(y) = \|y - A\delta_i(\hat{x}_1)\|_2 \tag{6.7}$$

for $i = 1, \ldots, k$, where δ_i is the characteristic function which selects the coefficients associated with the $i - th$ class.
4: Output the identities by:

$$identity(y) = arg\min_i r_i(y) \tag{6.8}$$

6.2.3 Method

Based upon LBP based face representation and sparse representation based classification we reviewed in Section 6.2.2, in this section, we propose a new face recognition algorithm that combines LBP features with SRC. In the first part, we outline the architecture of the proposed algorithm. Then we state the motivation of using the divide-and-conquer scheme which encodes LBP features into sparse coefficients vector (SCV) via efficient $l1$-minimization. Finally, the classification methodology based on the computed representation (augmented sparse coefficients vector) is discussed.

6.2.3.1 Architecture of the Proposed Algorithm

Figure 6.2 illustrates the proposed architecture of combining LBP features with SRC for face recognition. A facial image is first divided into several division levels (2×2, 4×4 and 8×8 in our approach). In each division level, the LBP histograms are summarized over all sub-blocks. Instead of constructing an enhanced LBP histogram for each facial image, we treat the LBP histogram of each sub-block individually. The original SRC takes the entire feature space extracted from the whole image as a dictionary component. In our approach the dictionary component is the feature vector extracted from one sub-block. In this way, the SCVs are computed for all sub-blocks of the input image via $l1$-minimization (Equation 6.6). Then the SCVs within the same division level are combined to build an enhanced SCV (E-SCV) using elementary-wise summation. In order to balance the classification impacts from all division levels, the E-SCVs are normalized to have the unit l2-norm. Finally, the normalized E-SCVs from all division levels are summed to construct an augmented SCV (A-SCV) in order to make the classification decision.

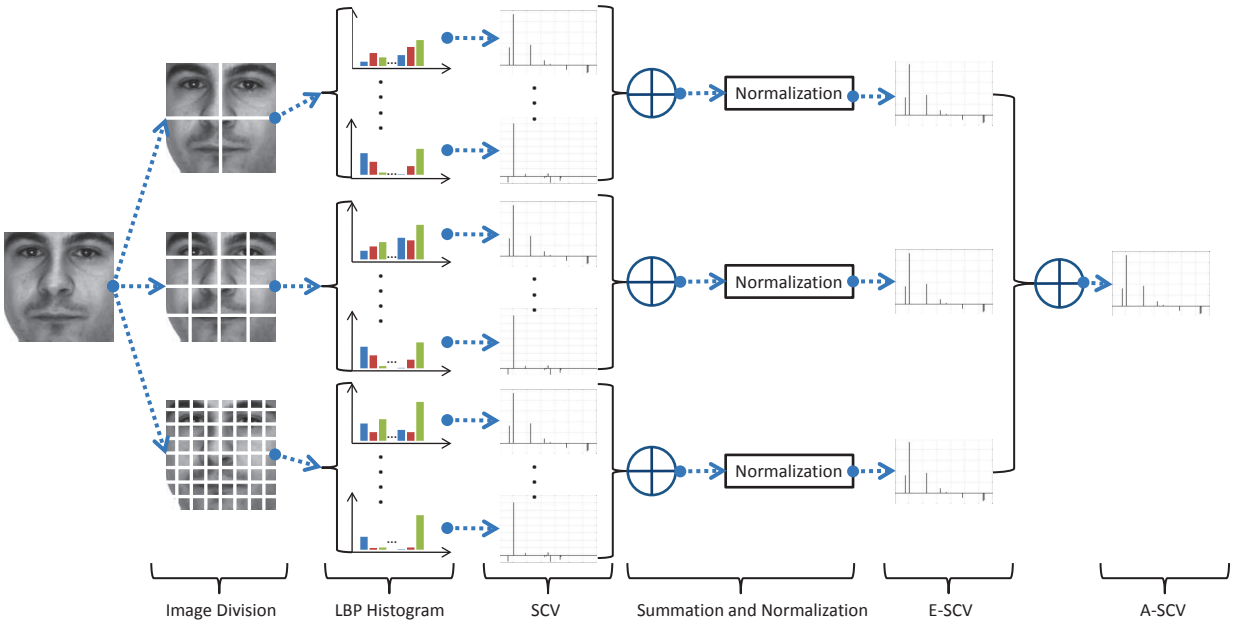


Figure 6.2: Architecture of the proposed approach.

The proposed algorithm is highly hierarchical. In the feature extraction phase: the labels for the LBP histogram include information about the patterns on a pixel-level; then the labels are summed over a small region to produce information on a regional level. In the classification phase: the SCV is recovered for each sub-block on a block-level; the E-SCV is computed for each type of division on a division-level; and the A-SCV is produced at the global-level. Such a pyramidal multi-level architecture ensures the accuracy and robustness of the proposed algorithm.

6.2.3.2 Motivating Divide-and-Conquer for Combining LBP Features with SRC

In this section, we discuss different options for efficiently combining high dimensional features (such as LBP based face representation) with the SRC algorithm. We will first see that using uniform LBP can already reduce the original LBP for efficient classification. Then we will motivate Divide-and-Conquer strategy rather than dimensionality reduction for the combination.

Uniform LBP: since the main difficulty of combining LBP features with SRC is due to the dimensionality problem, it encourages us to exploit more parsimonious features instead of the original LBP histogram. According to [117], most of the texture information is contained in a small subset of LBP patterns. Those patterns are called uniform patterns, which contain at most two bitwise transitions (0 to 1 or 1 to 0). By using uniform LBP histogram, the feature dimension thus decreases from 256 to 59. In our approach, the operator we used is $LBP^{u2}_{8,2}$ (which means uniform patterns, 8 equally spaced pixels on a circle of radius 2). Since the main purpose of our paper is to introduce a more appropriate way to integrate high dimensional features (like LBP) into SRC, we thus focus on the basic LBP descriptor.

Nevertheless, other features such as the extensions of LBP (e.g. LGBP [157], Multi-Scale LBP [98]) and other histogram descriptors (LPQ [118], GPP [155], HOG [47] etc.) could also be adopted into the proposed approach and might achieve higher recognition rates.

Dimensionality Reduction: a possible option to further reduce the feature dimension is to apply the tools like PCA or LDA to project the features from its original space to a reduced space [149]. Denoting the projection function as $R^{d\times m}$ with $d \ll m$, applying R to both sides of Equation 6.4 yields:

$$\tilde{y} \doteq Ry = RAx_0 \in R^d \tag{6.9}$$

The projection guarantees that the system of Equations 6.9 is under-determined. In addition, by selecting properly the reduced dimension, it can also ensure the system of equations has the unique sparsest solution. According to the experimental results we obtained, methods applying LDA can slightly improve the performance comparing to the baseline algorithm. However, it is still less accurate than the proposed method.

Divide-and-Conquer: the definition of divide-and-conquer motivates us to handle the "curse of dimensionality" using such a strategy:

"To solve a large instance of a problem, break it into smaller instances of the same problem, and use the solutions of these to the original problem."

As supported by [136], such a strategy is very useful in high-level image processing, and it can be effectively performed using parallel computing. By applying divide-and-conquer to an input image, the computation of $l1$-minimization becomes feasible for LBP based features in the constrained scenarios (e.g. limited number of training samples available for each class).

In addition, the divide-and-conquer algorithm enables us to incorporate information from multi-levels of descriptions by adding a pyramid structure. The composition of diverse representations from different division levels significantly improves the recognition performance.

6.2.3.3 Classification based on the Augmented Sparse Coefficients Vector

Once the augmented sparse coefficients vector (A-SCV) is computed, classification is conducted based on it. In Wright et al.'s method [146], the probe face is approximated using only the coefficients associated with the $i-th$ class. Then the classification is based on minimizing the residual between the probe image and the approximations. According to Equation 6.8, the identity is assigned by the class which has the least residual.

However, directly adopting Wright et al.'s method to A-SCV is inappropriate. Noticing that A-SCV is derived from SCVs which are computed from sub-blocks of the input image, it cannot be used to approximate the probe face by multiplication with the basis vector which is comprised of the whole facial images. Since a valid probe face can be sufficiently represented

only using the training samples from the same class, its SCV is naturally discriminative. A-SCV is actually an enhanced version of SCV. Supposing the values in A-SCV of a valid test sample y can be written as: $p_y(i,j)$,where $i \in [1,k]$, k is the number of classes and $j \in [1, J(i)]$, $J(i)$ is the number of faces in the $i-th$ class. We define a function $\Phi_i(y)$ as below:

$$\Phi_i(y) = \sum_{j=1}^{J(i)} p_y(i,j)/J(i) \tag{6.10}$$

which returns the normalized summation of sparse coefficients within the same class. Then the identity is assigned by:

$$identity(y) = arg \max_i \Phi_i(y) \tag{6.11}$$

The identity is assigned to the class which maximizes the normalized summation of the associated sparse coefficients.

6.2.4 Results

To assess the performance of our proposed approach, we performed a set of experiments on AR face database [106]. We will here illustrate the result of the proposed approach (D&C+LBP+SRC with the pyramid architecture), as well as four other methods we discussed in previous sections (LBP+NN [17], LBP+PCA+SRC, LBP+LDA+SRC and D&C+LBP+SRC). Results show that our proposed approach outperforms all the others. In addition, following the same configuration on the same database, we obtained better results than the best one reported in [146].

6.2.4.1 Experimental Data and Setup

The AR face database is a standard testing dataset in face recognition research. It contains more than 4000 face images of 126 subjects (70 men and 56 women) with different facial expressions, illumination conditions, and occlusions. For each subject, 26 pictures were taken in two separate sessions (two weeks interval between the two sessions). As Wright et al. did in [146] we configure the same dataset for testing. In the experiment, 100 subjects (half of male and half of female) are selected. For each subject, we chose 14 images with different illumination conditions and facial expressions: 7 images from session 1 for training and 7 images from session 2 for testing. The original image resolution is 768×576 pixels. Using the eye coordinates, we cropped, normalized and down-sampled the original images into 128×128 pixels.

Our dataset is quite challenging since it involves different variations including facial expressions, illumination conditions and the time elapse for 2 weeks. In addition, comparing to the dataset configured in [37], our dataset is more realistic. Instead of using 38 subjects with 19 training samples per subject, in our dataset, the number of subjects is 100 and there are only 7 training samples for each subject. The increased number of identities as well as the reduced

number of training samples increases the difficulty for face recognition. Directly applying SRC to LBP histogram extracted from the whole facial image is infeasible in this scenario.

Other settings of the experiments are listed here: the LBP operator used is $LBP^{u2}_{8,2}$; the ϵ in Equation 6.6 is 0.05 (as suggested in [149]); and the $l1$-minimization algorithm is implemented by $l1$-magic [10].

6.2.4.2 Experimental Results and Analysis

In this part, several LBP based algorithms are tested in order to demonstrate the superiority of the proposed approach. Conventionally, selecting the image division strategy for LBP based approaches is heuristical or empirical. Results in previous publications are often obtained by division strategies which maximize the recognition rates. Here we tested various algorithms with different division strategies. In our test, an image is divided into 2×2, 4×4 and 8×8 sub-blocks respectively.

Firstly, four previously discussed approaches (LBP+NN, LBP+PCA+SRC, LBP+LDA+SRC and D&C+LBP+SRC) are examined. LBP+NN indicates the original LBP based face recognition by Ahonen et al. [17]. It uses a nearest neighbour classifier and its similarity measure is based on the Chi-square distance (χ^2). LBP+PCA+SRC and LBP+LDA+SRC refer to the methods which apply the dimension reduction tools (Principal Component Analysis and Linear Discriminate Analysis, respectively) to reduce the dimension of extracted LBP features before Sparse Representation based Classification. D&C+LBP+SRC combines LBP features with SRC by using the proposed divide-and-conquer strategy but without the pyramid architecture. For fair comparison, the feature dimension is reduced to 59 in both PCA and LDA based approach, which is same as the length of LBP histogram extracted from one sub-block.

Figure 6.3 shows the recognition rates of above four methods on AR face dataset. First of all, it shows that by deploying more precise sub-blocks division the recognition rate increases correspondingly. The reason is because face representations obtained from more precise division strategies contain more spatial information. We consider LBP+NN as the baseline algorithm. In the figure, LBP+PCA+SRC outperforms LBP+NN for 2×2 sub-blocks; but it returns worse results than LBP+NN on 4×4 and 8×8 division strategies. LBP+LDA+PCA yields better results than both LBP+NN and LBP+PCA+SRC on 4×4 and 8×8 division strategies. It corresponds to the fact that by explicitly including the label information of the data constructs a more informative projection. Among all the results, D&C+LBP+SRC on 8×8 division strategy yields the best recognition rate. It demonstrates that only using the divide-and-conquer scheme could improve the performance when more spatial information is included.

Figure 6.4 shows the performance of our proposed approach (D&C+LBP+SRC with the pyramid architecture). In the figure, the improvement of recognition rate is significant. Using only the base level (8×8 sub-blocks) returns the same result as D&C+LBP+SRC in Figure

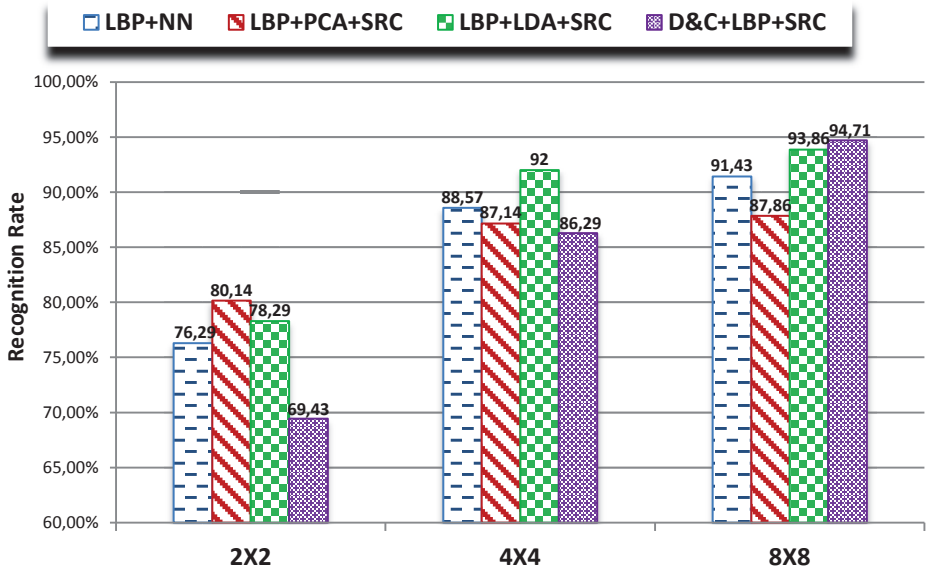


Figure 6.3: Recognition rates of four LBP based methods (LBP+NN, LBP+PCA+SRC, LBP+LDA+SRC and D&C+LBP+SRC) on three different division strategies (2×2, 4×4 and 8×8 sub-blocks).

6.3. But the recognition rate increases progressively when adding more levels in the pyramid. The recognition rate tends to converge as the increasing of pyramid levels. Hence it is inefficient to add more levels in the pyramid since the complexity increases accordingly.

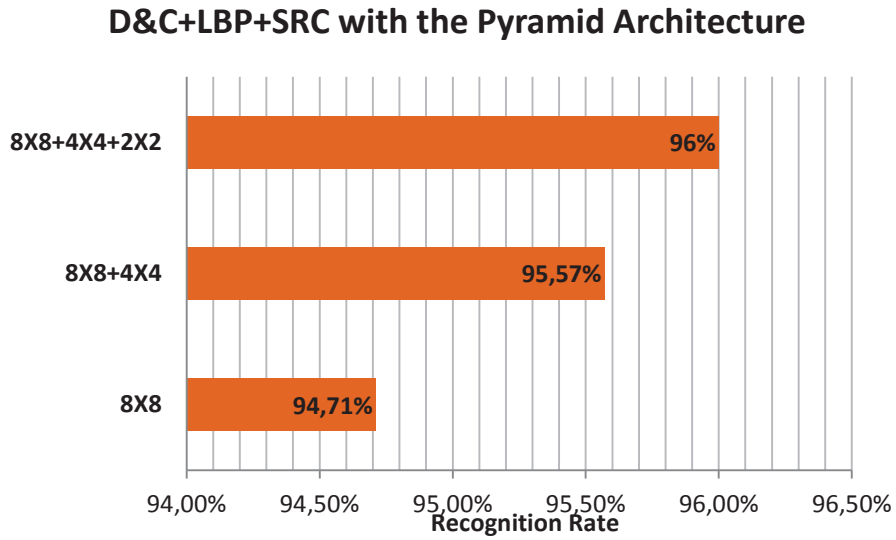


Figure 6.4: Recognition rates of the proposed approach (D&C+LBP+SRC with the pyramid architecture) in different pyramid levels (8×8, $8 \times 8 + 4 \times 4$ and $8 \times 8 + 4 \times 4 + 2 \times 2$).

It should be noticed that the computational burden of the proposed algorithm is higher than the previously discussed algorithms. Suppose that the proposed algorithm has n levels in the pyramid architecture, it requires to compute $(4^n - 1)/3$ times of $l1$-minimization. Therefore we only used a limited number of pyramid levels (i.e. 2×2, 4×4 and 8×8) in our approach. The experimental results reveal that our algorithm with chosen levels yields signifi-

cant recognition rates. In addition, when the parallel computing is considered, the complexity of the proposed approach can be reduced to $\Theta(n^{1/2})$ using the algorithm demonstrated in [136].

On the same dataset, the proposed algorithm reaches the recognition rate at 96%, which is higher than the best result reported by Wright et al. [146] (Fisherface+SRC, up to 94.7%).

6.3 Improving 3D Face Recognition via Multiple Intermediate Registration

In this section, we introduce a new method to improve face recognition in 3D using multiple intermediate references for face registration. In Section 6.3.1, we first give the argument why such a new registration architecture is needed and what is the advantage of the proposed method comparing with works in the literature. Then we validate the proposed registration method by implementing a complete 3D face recognition system which will be described in Section 6.3.2. Finally, experimental analysis on a real world dataset is given in Section 6.3.3.

6.3.1 Overview

One of key issues in 3D face recognition is to align 2 face shapes in a way that they can be better compared, namely face registration. Iterative Closest Point (ICP) methods [24, 40] are the dominating techniques for 3D face registration, since the first work presented by Medioni and Waupotitsch [109]. According to a recent review by Spreeuwers [134], face registration in the literature can be categorized into 3 classes:

1. One-to-all registration (e.g. [61]), where a probe face is registered to all gallery faces.
2. Registration to a face model or atlas, typically to an Average Face Model (AFM) [87]
3. Registration to an Intrinsic Coordinate System (ICS) [134].

The one-to-all approach is known to be accurate but slow in the identification mode since ICP is a time-consuming process. Registration to AFM or ICS has an indirect registration architecture, which requires only 1 time ICP process during an online query. However, the generation of AFM and ICS relies on the good land-marking of gallery faces, which is difficult for the noisy, incomplete and low-resolution faces captured by for example Kinect sensor.

Here, we propose to apply indirect face registration via multiple references (a small number of canonical faces), and we register both the probe face and gallery faces to this set of canonical faces (see Figure 6.5), for the following reason: during an online query, the registration needs only a few ICP processes (e.g. 3-5); there is no effort for land-marking; we demonstrate via experiments that increasing the number of canonical faces can significantly improve the identification results.

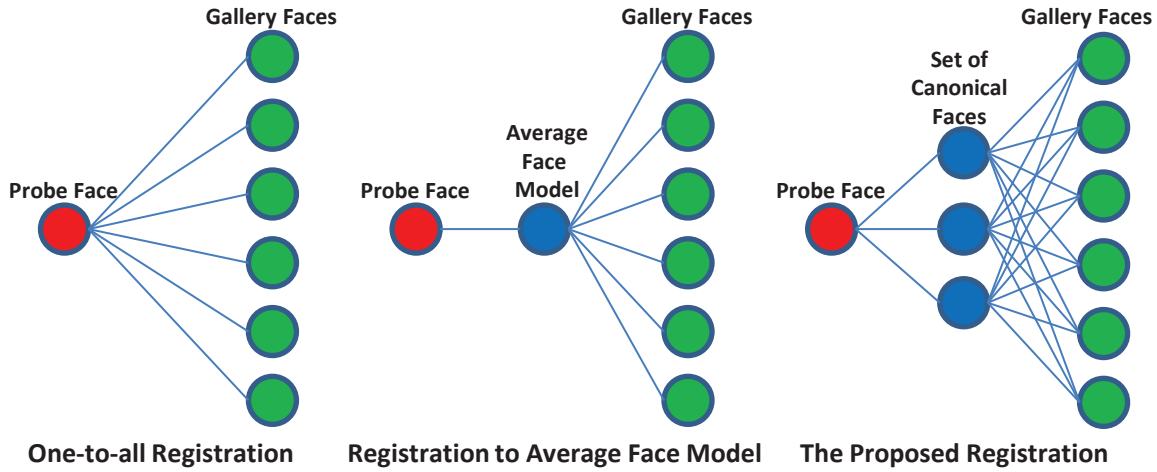


Figure 6.5: Architecture of one-to-all registration (left), architecture of registration to a AFM or ICS (middle), architecture of the proposed registration (right).

To justify the proposed 3D face registration architecture can efficiently improve 3D face recognition, we implement a fast and accurate online 3D face identification system based on ICP and data acquisition from Kinect/PrimeSensor. For the fast computation of ICP, we adopted the EM-ICP [67] algorithm based on a GPU implementation [137]. Our system can then work in real-time, with average speed ranging from 0.04s to 0.38s (depending on the number of canonical faces used). The implemented system is then tested on a real world dataset to demonstrate the superiority of the proposed face registration scheme.

The main contributions of this work can be summarized as follows:

- An efficient 3D face registration architecture via multiple intermediate references is proposed.
- A complete real-time 3D face identification system using a depth camera is implemented to justify the proposed 3D face registration architecture.
- Validation on real-world dataset.

Details of the work and our contributions can be observed in the following sections.

6.3.2 The System

The implemented 3D face recognition system includes a sequence of processing steps including: face detection and segmentation, the core registration module, facial region segmentation and weighting, 2.5D conversion, search space pruning and finally the one-to-many matching. A comprehensive overview is given in Figure 6.6.

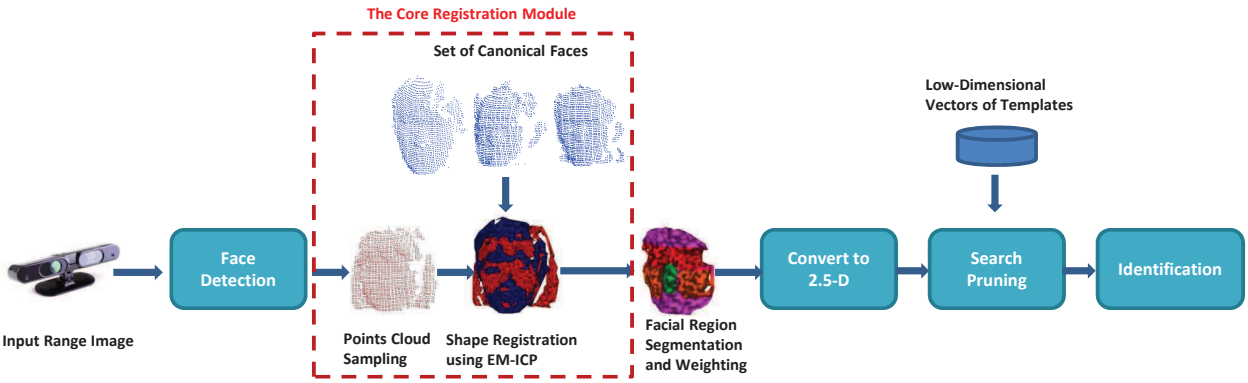


Figure 6.6: System overview.

6.3.2.1 Face Detection and Segmentation

The output from the PrimeSensor includes a RGB image and a depth map at 640×480 resolution. Although the face detection could be achieved by the popular Viola-Jones' method [143] using RGB images, it cannot segment the face/head region exactly from the background/body part. In addition, RGB images are sensitive to illumination variations. Therefore, we focus on the depth information from the range camera (which is illumination-invariant). Because pixels on the head surface have close depth values, given a pre-defined threshold, it is easy to segment the head region according to the depth discontinuity.

Given the segmented visible surface of a head, we first subsample points at a fixed resolution (60×60 in our system) and then compute corresponding real-world 3D coordinates. Any face with a lower resolution is automatically rejected as an invalid face candidate.

6.3.2.2 The Proposed Core Registration Module

The one-to-all registration method is too computationally complex for real-time identification. We therefore propose a simple, efficient and robust face registration strategy, which demands only a few ICP processes during an online query and does not require additional efforts of land-marking as for the registrations to AFM or ICS.

First, we randomly select M faces (selected from the gallery) to form a set of canonical faces (instead of random selection, gallery face clustering can also be adopted to select representative reference faces). During offline enrollment, each gallery face g_i ($i \in [1, N]$, where N is the size of gallery) is aligned with the M canonical faces (using ICP), and thus generates a set of aligned gallery faces $\{g'_{i,k}, k \in [1, M]\}$. During an online query, a probe face p is also aligned with the same M canonical faces, and thus generates a set of aligned probes $\{p'_k, k \in [1, M]\}$. An illustration of the alignment of a face to multiple canonical faces is shown in Figure 6.7. Later in the matching phase, an aligned probe face is matched with the aligned gallery face which is aligned with the same canonical face. Since the recognition capability of ICP based method relies on the fact that registration of different 3D faces from the same identity yields

the same or very similar alignment results, increasing the number of canonical faces for multiple alignments can significantly improve face recognition accuracy.

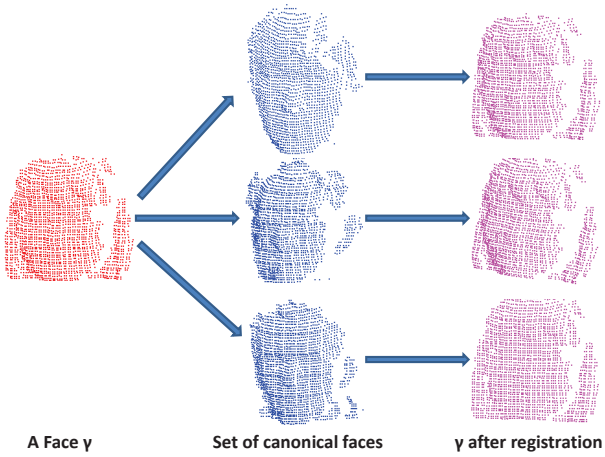


Figure 6.7: 3D face registration to a set of canonical faces.

The major computational burden lies on the registration of all gallery faces (which needs $N \times M$ times alignments), but this is done during offline training. An online query requires only M alignments ($M \ll N$). The proposed registration strategy greatly reduces the computation for online identification.

A generic ICP algorithm is time-consuming (it takes several seconds for a typical face data [134]). Here, in the implemented face recognition system, we adopt an improved version of ICP: the EM-ICP algorithm [67] for face alignment due to its reported accuracy and efficiency. The implementation of the EM-ICP algorithm is provided by Tamaki et al. [137] on CUDA architecture [115] using GPU computing, which is 60 times faster than the OpenMP based implementations on a multi-core CPU.

6.3.2.3 Facial Region Segmentation and Weighting

In order to exclude the unstable features (e.g. hair and boundary aliasing), we segment only the facial region. First, the coordinates of the nose-tip is manually annotated for each canonical face. After registration, we suppose the nose-tip of a probe face is aligned with the nose-tip of a canonical face. Then facial region segmentation is done by masking the registered probe face based on a Euclidean distance from the nose-tip.

We further divide the facial region into different facial areas (nose, eyes region, cheeks region and the rest part) [87]. Each area is associated with a weight (3, 2, 1, 0.5) to indicate its importance in face identification. The choice of these weights is based on experiments on a validation set.

6.3.2.4 Convert from 3D to 2.5D Representation

Matching two 3D points clouds demands looking for "hard" correspondence between sets (each point in one set has a unique mapping to a point in the other set). In the case of one-to-all registration, distances are returned directly by the ICP processes. However, when the probe face and all gallery faces are not directly registered, the distance between a probe and a gallery face requires explicit points indexing. One solution is to construct a K-D tree [22] for each points cloud, which is computational expensive in terms of both construction and query.

To reduce the computation of matching two 3D faces, we convert the registered 3D points cloud into a 2.5D representation via orthographic projection. During matching, the pixel-wise comparison of two 2.5D images does not need any explicit indexing and thus much faster than the matching of two 3D points clouds. These 2.5D images (after the proposed face registration) are good enough to establish identity, without further feature extractions on top of them (as shown in [87, 134]).

6.3.2.5 Vector based Search Pruning

A brute force comparison (even for the simple 2.5D image matching) of the query and the entire gallery may become extremely computational expensive, especially for large databases. The computational complexity is proportional to the number of entries in the gallery. We propose a two-steps non-linear process to first prune the search space using low-dimensional vectors. One way to compute such a vector for a probe face is to compute the $l2$ distances between each facial area on the probe face and the corresponding areas on the canonical faces; those computed distances are formed as the vector. The vectors of template faces are computed during the offline training. In this way, we could roughly narrow the search space more than 60% with 100% confidence in our experiments.

6.3.2.6 Identification

Given the set of a probe face p registered to M canonical faces $p'_k, k \in [1, M]$ (please notice that p'_k, $g'_{i,k}$ in Section 6.3.2.2 are 3D points clouds, whereas here they are range images after the 2.5D conversions described in Section 6.3.2.4) and the pruned search space $g'_{i,k}, i \in [1, N'], k \in [1, M]$ (N' is the number of gallery faces in the pruned search space), we find the probe's identity by the following equation:

$$id(p) = arg \min_{i=1...N'} \sum_{k=1}^{M} dist(p'_k, g'_{i,k}) \tag{6.12}$$

where $dist(p'_k, g'_{i,k})$ represents the Euclidean distance between p'_k and $g'_{i,k}$.

One notable problem in the matching phase is the large number of outliers (e.g. due to self-occlusion by hairs or some sensing errors like holes and spikes). We handle these outliers by imposing a universal threshold ($t = 50$) to all pixels. If the distance between 2 pixels is larger than the threshold, we regard it as an outlier and thus ignore its information.

6.3.3 Results

To assess the performance of the proposed system (in order to show the advantage of the proposed face registration architecture), we built a database using a PrimeSensor and performed a series of experiments on this dataset.

6.3.3.1 Data and Setup

We collected 1054 frontal faces from 20 people (10 to 135 faces for each person) for testing. Each person is asked to sit in front of a PrimeSensor (0.8m-1.2m) for a short period of time in an office environment, potentially with slight head/facial movements. We randomly selected one face from each person as the gallery face of the enrolled identity. All other faces are tested as probe faces for identification. The canonical faces used for face registration are randomly selected from the gallery. The number of canonical faces in the system can be changed according to different configurations.

6.3.3.2 Face Identification Results and Analysis

We first show the Cumulative Match Characteristic (CMC) [95] of the proposed system using various values of M (1 to 5) canonical faces for registration in Figure 6.8 (please note that using 1 canonical face is not equal to using AFM or ICS for the registration since the canonical faces were randomly selected). It is clear that the identification rates converge to 100% faster (in fewer ranks) when more canonical faces are incorporated. However, using more than 3 canonical faces does not produce a significant improvement.

Figure 6.9 shows the rank-1 identification rates of the proposed system when using M=1-5 canonical faces for face registration. In the "Image-to-Image" scenario, each probe face obtained from 1 depth shot is compared with the gallery faces. In the figure, higher identification rates are achieved when more canonical faces are used. When using 5 3D face models as canonical faces, we obtained the highest identification rate in rank-1 (up to 97.91%).

In the "Video-to-Image" scenario, instead of taking one shot of a person as a probe image, we take a short video sequence of the person as the probe video (9 to 134 frames, which are matched with the 20 gallery faces). First, each frame in the probe video is identified individually; then their results are combined by taking the majority decisions to output the identity of the probe video. In the figure, even if only one canonical face is used, we achieve 100% identification rate. This is because the identification result of each frame is already reliable

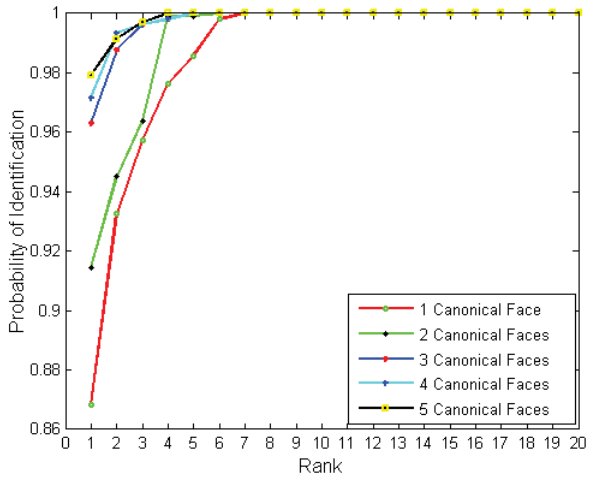


Figure 6.8: The Cumulative Match Characteristic (CMC) using 1-5 canonical faces for registration (M=1:5).

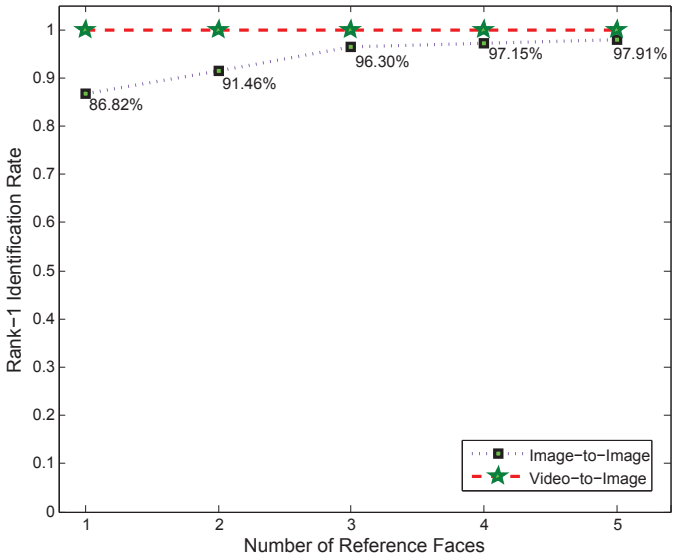


Figure 6.9: Rank-1 identification rates of the "Image-to-Image" approach and the "Video-to-Image" approach based on 1-5 canonical faces (M=1:5) for registration.

(as we have shown in the "Image-to-Image" case), and errors are uniformly distributed for each person (which means that we do not have a lot of errors in one frame sequence but no error for the others). These numbers (100%) are for a 20 persons database. Nevertheless, our results suggest that when extending our system with larger number of subjects in the gallery, the "Video-to-Image" method would ameliorate the identification degenerations due to increased identity variations.

6.3.3.3 Time Complexity Analysis

The proposed system needs to work in real time. Since the system is implemented for a low-cost depth camera (the PrimeSensor), it should perform at interactive rates for daily-use purpose. We conducted our experiments on a consumer-level hardware (Intel(R) Xeon(R) CPU E5520 @ 2.27GHz, NVIDIA(R) Tesla(R) C1060). The average computing time is 0.047s, 0.130s, 0.233s, 0.296s and 0.378s to identify one probe face when using 1-5 canonical faces for face registration.

To validate the proposed face registration architecture for a large scale dataset, we artificially generated 3 gallery sets with different sizes (1000, 2000, and 3000) by data duplication. Figure 6.9 shows the time complexity analysis on the generated large scale dataset. In the figure, we observe that when the number of references is small (e.g. $M = 1$), increasing the number of gallery faces ($N = 1000 - 3000$) maintains similar rates (0.481s-1.316s, still working in real time); however, when M becomes large, the differences among rates become more significant. This result again suggests that one-to-all/many registration is inappropriate for large scale identification. Another observation is when the number of gallery faces (N) increases, the increment of computing time becomes faster (larger slope) along with the increment of M. This is because we generate $M \times N$ registered templates in the gallery; when N becomes large, the total identification time requires more computation time.

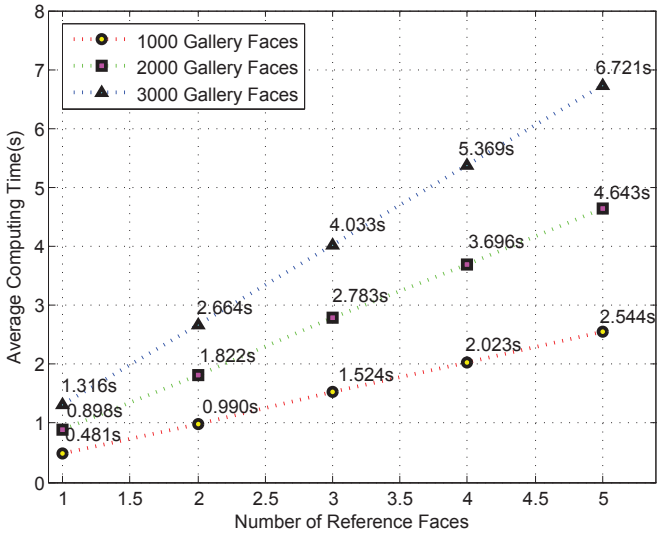


Figure 6.10: Time complexity analysis on large scale dataset (produced by data duplication).

This experiment demonstrates that by selecting an appropriate number of intermediate references (canonical faces M), our system works in real time for identification of thousand of faces. And the proposed system works always faster than one-to-all face registration. However, when the number of faces to identify becomes very huge (where $N \gg 1000$), the proposed face registration architecture can perform less efficient than its performance on small scale dataset.

6.4 Conclusions

In this chapter, we presented two improved versions of the baseline 2D and 3D face recognition algorithms (i.e. LBP, SRC and ICP). The motivation behind our works described in this chapter is the fact that improving the baseline face recognition algorithm can also upgrade the recognition accuracy in occlusion conditions when explicit occlusion analysis and processing (as we suggested in Chapter 3-5) is included.

For the improved 2D face recognition, we proposed a novel method which combines LBP based feature extraction with sparse representation based classification (SRC). Comparing to the state-of-art method, the proposed approach is more appropriated for realistic scenarios (usually there are only few training samples available for each class). The dimensionality problem is resolved by applying the divide-and-conquer strategy and the discriminative power is strengthen by a pyramid architecture which incorporates information from different levels of descriptions. The proposed algorithm is compared with the baseline algorithm [37] as well as the methods combining LBP with SRC based on dimension reduction tools (PCA and LDA); it also yields better recognition results than the best result reported by Wright et al. [146] on the same database. Furthermore, the proposed approach is not only restricted to the basic LBP features but also compatible with other high dimensional histogram descriptors which might achieve improved recognition results.

For the improved 3D face recognition, we presented a complete framework of an online 3-D face identification system based on Kinect/PrimeSensor working in real-time. A fast face registration approach using multiple intermediate references is applied and thus the system accuracy is significantly improved. Our system achieves high identification rates on noisy, incomplete and low-resolution face data and can process up to 20 fps on consumer-level hardware.

As the feature works, other histogram descriptors (such as LGBP, HOG, LPQ etc.) can be exploited for the potential performance increase when combining with SRC. More sophisticated fusion scheme can be used instead of simple summation when fusing the obtained sparse coefficients vectors. For the proposed 3D face registration, how the M canonical faces can be better selected (instead of random selection as we presented), is an interesting topic (which might be achieved by accurate face clustering).

Chapter 7

Conclusions

In this chapter, we elaborate the main achievements of this PhD thesis and discuss the future perspectives.

7.1 Achievements

Face recognition, the least intrusive biometric technique from the acquisition point of view, has been applied to a wide range of commercial and law enforcement applications. With the emphasis on real world scenarios (e.g. face recognition in video surveillance), in the last decade, an enormous amount of research on face recognition under pose/illumination changes and image degradations has been conducted. However, problems caused by occlusions are mostly overlooked, although facial occlusions are quite common in non-cooperative systems such as video surveillance applications. Works in the literature focus on finding corruption-tolerant features or classifiers to reduce the effect of partial occlusions in face representation. However, information from the occluded parts can still hinder the recognition performance and thus they cannot be the optimal solutions. In addition, previous works mainly focus on facial occlusions which are dense and contiguous (e.g. sunglasses, scarf, beards, hat and hand on face), whereas neglecting the other types of facial occlusions.

In this thesis, we focused on handling the facial occlusion problem in face recognition. We approached toward this goal from different aspects, including classical problems addressing, identifying and handling new problems in more advanced conditions, exploiting new sensor for the occlusion problem, as well as improving the baseline algorithms.

Classical Occlusion Handling

We studied the classical problems of occlusion in face recognition (mainly due to facial accessories such as sunglasses and scarf). We figured out that locally emphasized algorithms are insufficient to fully handle the occlusion problem, instead, explicit oc-

clusion analysis can be integrated into the recognition process which greatly improves the performance of face recognition under occlusion conditions. In our work, a novel framework for improving the recognition of occluded faces is proposed; the new techniques to detect and segment facial occlusion are thoroughly described; and extensive experimental analysis is conducted, demonstrating significant performance enhancement using the proposed approaches compared to the state-of-the-art methods under various configurations including robustness against sunglasses, scarves, non-occluded faces, screaming and illumination changes.

Advanced Occlusion Handling

We presented some first solutions to the newly identified sparse occlusion and dynamic occlusion problems in the context of face biometrics in video surveillance. The proposed approaches exploit several advanced image analysis and processing techniques in order to achieve the sophisticated goal. We proposed an approach of first detecting the presence of occlusion in local pixels based on the framework of Robust-PCA, and then inpainting the occluded pixels given the information from the occlusion detection part for face recognition. This approach can significantly improve various face recognition algorithms in complex sparse occlusion scenarios. We also presented a system to detect dynamic occlusion due to cap in the entrance surveillance scenario. The proposed system can be applied to a wide range of applications for security management in nowadays video surveillance. In particular, it can provide prior information of occlusion to face recognition system of identifying suspicious persons.

Exploiting New Sensor

We revealed the capability of Kinect for face recognition, especially for the scenarios when partial occlusion exists. We built the first publicly available face database - the KinectFaceDB based on the emerging Kinect sensor. The database consists of different data modalities (well-aligned and processed 2D, 2.5D, and 3D based face data) and multiple facial variations. Benchmark evaluations and comparative studies are reported in the context of face biometrics. Instead of detecting the presence of occlusion using intensity cue, we proposed to estimate the occlusion probability based on depth cue, so as to improve face recognition using intensity images. We demonstrated via experiments that depth cue is more suitable to uncover the occlusion structure on a face than intensity cue in most cases.

Improving Baseline Algorithms

We proposed two new algorithms which improve the baseline 2D and 3D face recognition algorithms (i.e. LBP, SRC and ICP). We proposed to combine LBP based feature extraction with sparse representation based classification for 2D faces. The dimensionality problem is resolved by applying the divide-and-conquer strategy and the discriminative power is strengthen by a pyramid architecture which incorporates information from different levels of descriptions. We demonstrated via extensive experiments that the proposed method outperform the state-of-the-art works. For the improved 3D face recognition, we proposed a complete framework of an online 3-D face identification system based on Kinect/PrimeSensor working in real-time. A fast face registration ap-

proach using multiple intermediate references is applied and thus the system accuracy is significantly improved. Our system achieves high identification rates on noisy, incomplete and low-resolution face data and can process up to 20 fps on consumer-level hardware.

To summarize, in this thesis, we presented our work for face recognition robust to occlusions. All of the proposed approaches contribute to improve face recognition rates when partial occlusion occurs. The philosophy behind our approaches is indifferent from most of the works in the literature and leads to the state-of-the-art performance in various experiments. Several new approaches and challenges are presented, so that significant improvements for the robustness of face recognition to occlusions are demonstrated.

7.2 Perspectives

We outline the main research directions and possible improvements arisen from this thesis as follows.

New Challenges in Face Recognition under Occlusion Conditions

We illustrated via this thesis that there may exist a large variety of occlusions in real world scenarios. Those facial occlusions may deteriorate face recognition in different applications and may cause severe issues in the surveillance world. Unfortunately, we cannot provide an exhaustive list of all possible facial occlusions in this thesis (our work only focus on the static/dynamic, dense/sparse occlusion problems in face recognition). Here we give one example of those unsolved facial occlusions in practical scenarios. Semi-transparent occlusion, may occur in many practical conditions (e.g. recognition of pilot in car, face recognition behind window/under water). The removal/separation of specular reflection layer has a long history study in computer vision/optics, however it has many intrinsic challenges specific to face recognition context (e.g. to simultaneously find a reference pattern and the face ID). The design of an optimal semi-transparent occlusion removal method is important to improve face recognition in such a scenario. In addition, it is of great interest to find an universal solution for all different types of occlusions in face recognition, with practical time complexity and configurations.

Kinect for Robust Face Recognition

Kinect sensor has depicted a broad prospects of 3-Dimensional data based computer applications. As its success in other computer vision applications, there will be a large room for upcoming researches in facial recognition and analysis (e.g. gender/age/ethnicity recognition, anthropometric measurement and face reconstruction) exploiting Kinect. We provided a preliminary study of using Kinect to improve occluded face recognition. However, a number of Kinect-specific challenges are arising in either general or application-driven contexts, for example, RGB-D co-segmentation of visual objects, per-pixel depth and RGB compensation from each other, 3D/depth

analysis in time (i.e. 3+1D), and multi-modal combination of video, audio and 3D. Processing using Kinect and GPU can also make the system working in real time and potentially be commercialized for real-world products. How to correctly address above issues, implement real-time systems, in addition to fully exploit the capability of Kinect for robust face recognition (particularly robust to occlusion) are very important topics in subsequent years.

Occlusion Problems in presence of Other Facial Variations

So far, almost all occlusion studies focus on the occlusion problem itself (e.g. scarf and sunglasses). However, with emphasis on the real world scenarios, the presence of occlusion can be together with the presence of other facial variations (e.g. illuminations/expressions/poses/image degradations). In order to achieve robust face recognition in non-controlled environments, face recognition which can simultaneously handle multiple facial variations (including occlusion) is a very important topic. For future works, we will investigate on how to handle the occlusion problem in more challenging conditions (e.g. occlusion + pose variations). By providing the face recognition solutions robust to multiple different facial variations can significantly reinforce the security in various public places as well as private residences.

7.3 Conclusions

In this chapter, we conclude the work of this thesis. We summarized our main achievements from different aspects, including different challenges and different approaches. We also discussed some new trends and perspectives for the works in the future.

Bibliography

[1] http://www.scholarpedia.org/article/Local_Binary_Patterns.

[2] http://www.xbox.com/en-US/KINECT.

[3] http://www.konicaminolta.com/.

[4] http://www.face-rec.org/databases/.

[5] http://www.sic.rma.ac.be/∼ beumier/DB/3d_rma.html.

[6] http : //www-users.cs.york.ac.uk/∼ nep/research/3Dface/tomh/3DFaceDatabase.html.

[7] http://www.3dmd.com/.

[8] http://www.cyberware.com/.

[9] http://www.tabularasa-euproject.org/.

[10] http://www.acm.caltech.edu/l1magic/.

[11] The bjut-3d large-scale chinese face database. *Technical Report*.

[12] Di3d. http:/www.di3d.com/.

[13] Inspeck. www.creaform3d.com/.

[14] Low-rank matrix recovery and completion via convex optimization. http://perception.csl.illinois.edu/matrix-rank/.

[15] Uhdb11. http://cbl.uh.edu/URxD/datasets/.

[16] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recogn. Lett.*, 28(14):1885–1906, Oct. 2007.

[17] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037 –2041, dec. 2006.

[18] F. Barak, S. Alexander, M. Meir, and A. Yoel. Depth mapping using projected patterns. *US Patent : 20100118123*, May 2010.

[19] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *ECCV Workshops (1)*, volume 7583 of *Lecture Notes in Computer Science*, pages 433–442. Springer, 2012.

[20] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218 – 233, feb 2003.

[21] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, July 1997.

[22] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept. 1975.

[23] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[24] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, Feb. 1992.

[25] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 232 –237, jun 1998.

[26] P. E. Black. dense graph. in Dictionary of Algorithms and Data Structures, August 2008.

[27] P. E. Black. sparse graph. in Dictionary of Algorithms and Data Structures, August 2008.

[28] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition. *Comput. Vis. Image Underst.*, 101(1):1–15, Jan. 2006.

[29] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition. *Comput. Vis. Image Underst.*, 101(1):1–15, Jan. 2006.

[30] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001.

[31] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[32] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[33] J. C. Brown and P. J. O. Miller. Automatic classification of killer whale vocalizations using dynamic time warping. *The Journal of the Acoustical Society of America*, 122(2):1201–1207, 2007.

[34] G. Burel and D. Carel. Detection and localization of faces on digital images. *Pattern Recognition Letters*, 15(10):963–967, Oct. 1994.

[35] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[36] D. Chai and K. Ngan. Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):551 –564, jun 1999.

[37] C. H. Chan and J. Kittler. Sparse representation of (multiscale) histograms for face recognition robust to registration and illumination problems. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2441 –2444, sept. 2010.

[38] Y.-R. T. Che-Yen Wen, Shih-Hsuan Chiu and C.-P. Lu. The mask detection technology for occluded face analysis in the surveillance system. 50(3), May 2005.

[39] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5):705 –741, may 1995.

[40] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image Vision Comput.*, 10(3):145–155, Apr. 1992.

[41] C. chung Chang and C.-J. Lin. Libsvm: a library for support vector machines, 2001.

[42] A. Colombo, C. Cusano, and R. Schettini. Three-dimensional occlusion detection and restoration of partially occluded faces. *J. Math. Imaging Vis.*, 40(1):105–119, May 2011.

[43] A. Colombo, C. Cusano, and R. Schettini. Three-dimensional occlusion detection and restoration of partially occluded faces. *J. Math. Imaging Vis.*, 40(1):105–119, May 2011.

[44] A. Colombo, C. Cusano, and R. Schettini. Umb-db: A database of partially occluded 3d faces. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2113 –2119, nov. 2011.

[45] A. Colombo, C. Cusano, and R. Schettini. Umb-db: A database of partially occluded 3d faces. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2113 –2119, nov. 2011.

[46] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[47] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1, june 2005.

[48] J. Daugman. Face and gesture recognition: overview. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):675 –676, jul 1997.

[49] J. Daugman. How iris recognition works. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):21 – 30, jan. 2004.

[50] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[51] R. Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer, August 2005.

[52] D. L. Donoho. High-dimensional data analysis: the curses and blessings of dimensionality. In *American Mathematical Society Conf. Math Challenges of the 21st Century*. 2000.

[53] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math*, 59:797–829, 2004.

[54] S. Du, M. Shehata, and W. Badawy. Hard hat detection in video sequences based on face features, motion and color information. In *Computer Research and Development (ICCRD), 2011 3rd International Conference on*, volume 4, pages 25 –29, march 2011.

[55] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In W. Schempp and K. Zeller, editors, *Constructive Theory of Functions of Several Variables*, volume 571 of *Lecture Notes in Mathematics*, pages 85–100. Springer Berlin / Heidelberg, 1977. 10.1007/BFb0086566.

[56] H. K. Ekenel and R. Stiefelhagen. Why is facial occlusion a challenging problem? In *Proceedings of the Third International Conference on Advances in Biometrics*, ICB '09, pages 299–308, Berlin, Heidelberg, 2009. Springer-Verlag.

[57] M. Elendt, editor. *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2011, Vancouver, BC, Canada, August 7-11, 2011, Talks Proceedings*. ACM, 2011.

[58] N. Erdogmus and J.-L. Dugelay. On discriminative properties of TPS warping parameters for 3D face recognition. In *ICIEV 2012, IEEE/IAPR International Conference on Informatics, Electronics & Vision, 18-19 May 2012, Dhaka, Bangladesh*, Dhaka, BANGLADESH, 05 2012.

[59] M. F. Fallon, H. Johannsson, and J. J. Leonard. Efficient scene simulation for robust monte carlo localization using an rgb-d camera. In *ICRA*, pages 1663–1670. IEEE, 2012.

[60] T. Faltemier, K. Bowyer, and P. Flynn. Using a multi-instance enrollment representation to improve 3d face recognition. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1 –6, sept. 2007.

[61] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. Using multi-instance enrollment to improve performance of 3d face recognition. *Computer Vision and Image Understanding*, 112(2):114–125, 2008.

[62] S. Fidler, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):337–350, Mar. 2006.

[63] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, (7):179–188.

[64] A. Fusiello, V. Murino, and R. Cucchiara, editors. *Person Identification Using Full-Body Motion and Anthropometric Biometrics from Kinect Videos.*, volume 7585 of *Lecture Notes in Computer Science*. Springer, 2012.

[65] M. Gabel, E. Renshaw, A. Schuster, and R. Gilad-Bachrach. Full body gait analysis with kinect. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, August 2012.

[66] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721 –741, nov. 1984.

[67] S. Granger and X. Pennec. Multi-scale em-icp: A fast and robust approach for surface registration. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 418–432, London, UK, UK, 2002. Springer-Verlag.

[68] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *Image Processing, IEEE Transactions on*, 19(6):1657 –1663, june 2010.

[69] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik. Texas 3d face recognition database. In *Image Analysis Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*, pages 97 –100, may 2010.

[70] X. He and P. Niyogi. Locality Preserving Projections. Cambridge, MA, 2004. MIT Press.

[71] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):328 –340, march 2005.

[72] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *I. J. Robotic Res.*, 31(5):647–663, 2012.

[73] D. Herrera C., J. Kannala, and J. Heikkilä. Accurate and practical calibration of a depth and color camera pair. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, editors, *Computer Analysis of Images and Patterns*, volume 6855 of *Lecture Notes in Computer Science*, pages 437–445. Springer Berlin / Heidelberg, 2011.

[74] D. Herrera C., J. Kannala, and J. Heikkilä and. Joint depth and color camera calibration with distortion correction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):2058 –2064, oct. 2012.

[75] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):765 –781, nov. 2011.

[76] X. Huang, S. Li, and Y. Wang. Jensen-shannon boosting learning for object recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 144 – 149 vol. 2, june 2005.

[77] Y. Huang, Y. Wang, and T. Tan. Combining statistics of geometrical and correlative features for 3d face recognition. In *Proc. BMVC*, pages 90.1–90.10, 2006. doi:10.5244/C.20.90.

[78] T. Huynh, R. Min, and J.-L. Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *ACCV 2012, Workshop on Computer Vision with Local Binary Pattern Variants, Daejeon, Korea, November 5-9, 2012 / To be published also as LNCS*, Daejeon, KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF, 11 2012.

[79] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10:1–5, 1999.

[80] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, pages 559–568, 2011.

[81] A. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(4):302 –314, apr 1997.

[82] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):4 – 20, jan. 2004.

[83] S. M. Y. Jaywoo Kim, Younghun Sung and B. G. Park. A new video surveillance system employing occluded face detection. 2005.

[84] H. Jia and A. M. Martinez. Support vector machines in face recognition with occlusions. In *in Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 136–141.

[85] R. Johnson, K. O'Hara, A. Sellen, C. Cousins, and A. Criminisi. Exploring the potential for touchless interaction in image-guided interventional radiology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3323–3332, New York, NY, USA, 2011. ACM.

[86] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, Oct. 2002.

[87] I. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):640 –649, april 2007.

[88] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[89] J. Kim, J. Choi, J. Yi, and M. Turk. Effective representation using ica for face recognition robust to local distortion and partial occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1977–1981, 2005.

[90] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:65–81, 2004.

[91] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[92] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):684–698, May 2005.

[93] S. Z. Li. *Encyclopedia of Biometrics*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[94] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. pages 207–212, 2001.

[95] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition, 2nd Edition*. Springer, 2011.

[96] S. Liao and A. K. Jain. Partial face recognition: An alignment free approach. *Biometrics, International Joint Conference on*, 0:1–8, 2011.

[97] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *Proceedings of the 2007 international conference on Advances in Biometrics*, ICB'07, pages 828–837, Berlin, Heidelberg, 2007. Springer-Verlag.

[98] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *Proceedings of the 2007 international conference on Advances in Biometrics*, ICB'07, pages 828–837, Berlin, Heidelberg, 2007. Springer-Verlag.

[99] D.-T. Lin and M.-J. Liu. Face occlusion detection for automated teller machine surveillance. 4319, 2006.

[100] H.-S. Lin. Working set selection using the second order information for training svm. June 2003.

[101] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *ArXiv e-prints*, Sept. 2010.

[102] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150 –1157 vol.2, 1999.

[103] J. Maatta, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using micro-texture analysis. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1 –7, oct. 2011.

[104] J. Maatta, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *Biometrics, IET*, 1(1):3 –10, march 2012.

[105] A. Martinez and R. Benavente. The ar face database. *CVC Technical Report*, (24), 1998.

[106] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998.

[107] A. M. Martinez. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class, 2002.

[108] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, Mar. 2008.

[109] G. Medioni and R. Waupotitsch. Face modeling and recognition in 3-d. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, AMFG '03, pages 232–, Washington, DC, USA, 2003. IEEE Computer Society.

[110] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[111] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay. Real-time 3D face identification from a depth camera. In *ICPR 2012, 21st International Conference on Pattern Recognition, November 11-15, 2012, Tsukuba International Congress Center, Tsukuba Science City, Japan*, Tsukuba, JAPAN, 11 2012.

[112] R. Min, A. D'angelo, and J.-L. Dugelay. Efficient scarf detection prior to face recognition. In *EUSIPCO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark*, Aalborg, DENMARK, 08 2010.

[113] R. Min, A. Hadid, and J.-L. Dugelay. Improving the recognition of faces occluded by facial accessories. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 442 –447, march 2011.

[114] A. B. Moreno and A. Sánchez. GavabDB: a 3D Face Database. In *Workshop on Biometrics on the Internet*, pages 77–85, Vigo, Mar. 2004.

[115] nVidia Corporation. *CUDA CUBLAS Library*, Aug. 2010.

[116] H. J. Oh, K. M. Lee, and S. U. Lee. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image Vision Comput.*, 26(11):1515–1523, Nov. 2008.

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

[118] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *ICISP*, volume 5099 of *Lecture Notes in Computer Science*, pages 236–243. Springer, 2008.

[119] G. P. Otto and T. K. W. Chau. 'region-growing' algorithm for matching of terrain images. *Image Vision Comput.*, 7(2):83–94, May 1989.

[120] B.-G. Park, K.-M. Lee, and S.-U. Lee. Face recognition using face-arg matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1982–1988, Dec. 2005.

[121] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee. Glasses removal from facial image using recursive pca reconstruction. In *Proceedings of the 4th international conference on Audio- and video-based biometric person authentication*, AVBPA'03, pages 369–376, Berlin, Heidelberg, 2003. Springer-Verlag.

[122] P. S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation, 1996.

[123] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 947 – 954 vol. 1, june 2005.

[124] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pages 185–192. IEEE, 2011.

[125] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 137–, Washington, DC, USA, 1997. IEEE Computer Society.

[126] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien. Gait recognition with kinect. In *Proceedings of the First Workshop on Kinect in Pervasive Computing*, 2012.

[127] A. Rama, F. Tarres, L. Goldmann, and T. Sikora. More robust face recognition by considering occlusion information. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1 –6, sept. 2008.

[128] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009.

[129] S. C. K. Sang Min Yoon. Detection of partially occluded face using support vector machines.

[130] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Biometrics and identity management. chapter Bosphorus Database for 3D Face Analysis, pages 47–56. Springer-Verlag, Berlin, Heidelberg, 2008.

[131] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, Washington, DC, USA, 2011. IEEE Computer Society.

[132] S. Singh, A. Gyaourova, G. Bebis, and I. Pavlidis. Infrared and visible image fusion for face recognition. *Proc. SPIE 5404, Biometric Technology for Human Identification*, pages 585–596, 2004.

[133] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948 –1962, nov. 2006.

[134] L. Spreeuwers. Fast and accurate 3d face recognition. *Int. J. Comput. Vision*, 93(3):389–414, July 2011.

[135] E. Stone and M. Skubic. Evaluation of an inexpensive depth camera for in-home gait assessment. *J. Ambient Intell. Smart Environ.*, 3(4):349–361, Dec. 2011.

[136] Q. F. Stout. Supporting divide-and-conquer algorithms for image processing. *J. Parallel Distrib. Comput.*, 4(1):95–115, Feb. 1987.

[137] T. Tamaki, M. Abe, B. Raytchev, and K. Kaneda. Softassign and em-icp on gpu. In *Networking and Computing (ICNC), 2010 First International Conference on*, pages 179 – 183, nov. 2010.

[138] X. Tan, S. Chen, Z. hua Zhou, and F. Zhang. Ieee transactions on neural networks 1 recognizing partially occluded, expression variant faces from single training image per person with som and soft knn ensemble.

[139] R. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *Computer Vision and Pattern Recognition*, 1986.

[140] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, 1991.

[141] C. Velardo and J.-L. Dugelay. Real time extraction of body soft biometric from 3d videos. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 781–782, New York, NY, USA, 2011. ACM.

[142] C. Velardo, J.-L. Dugelay, L. Daniel, A. Dantcheva, N. Erdogmus, N. Kose, R. Min, and X. Zhao. *Introduction to biometry*. Book chapter in "Multimedia Image and Video Processing" (2nd revised edition); CRC Press; 2012; ISBN:978-1439830864, 02 2011.

[143] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[144] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505 – 1518, dec. 2003.

[145] J. Wright and Y. Ma. Dense error correction via l1-minimization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3033 –3036, april 2009.

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

[147] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xu, and Z. Zhang. Automatic eyeglasses removal from face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(3):322–336, Mar. 2004.

[148] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma. Fast *l*1-minimization algorithms and an application in robust face recognition: A review. In *ICIP*, pages 1849–1852, 2010.

[149] A. Y. Yang, J. Wright, Y. Ma, and S. Sastry. Feature selection in face recognition: A sparse representation perspective. Technical Report UCB/EECS-2007-99, 2007.

[150] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *Proceedings of the 11th European conference on Computer vision: Part VI*, ECCV'10, pages 448–461, Berlin, Heidelberg, 2010. Springer-Verlag.

[151] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625 –632, june 2011.

[152] Y.-H. Yang and M. D. Levine. The background primal sketch: an approach for tracking moving objects. *Mach. Vision Appl.*, 5(1):17–34, Jan. 1992.

[153] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–6, 2008.

[154] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:211–216, 2006.

[155] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Transactions on Image Processing*, 16(1):57–68, 2007.

[156] W. Zhang, S. Shan, X. Chen, and W. Gao. Local gabor binary patterns based on kullback leibler divergence for partially occluded face recognition. *Signal Processing Letters, IEEE*, 14(11):875 –878, nov. 2007.

[157] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 786–791, Washington, DC, USA, 2005. IEEE Computer Society.

[158] Z. Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4 –10, feb. 2012.

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

[160] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using markov random fields. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1050 –1057, 29 2009-oct. 2 2009.

[161] X. Zou, J. Kittler, and K. Messer. Illumination invariant face recognition: A survey. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1 –8, sept. 2007.

List of Publications

[1] Rui Min, Kose Neslihan, and Jean-Luc Dugelay. Kinectfacedb: a kinect face database for face recognition. In *submitted to IEEE Transaction on System, Man, and Cybernetics: Systems (T-SMC-S). special issue on "Biometric Systems and Applications"*, 2013.

[2] Rui Min and Jean-Luc Dugelay. Depth assisted 2d face recognition under partial occlusions. In *submitted to ICIP 2013, IEEE International Conference on Image Processing, 15-18 September, 2013, Melbourne, Australia*, Melbourne, Australia, 09 2013.

[3] Tri Huynh, Rui Min, and Jean-Luc Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *ACCV 2012, Workshop on Computer Vision with Local Binary Pattern Variants, Daejeon, Korea, November 5-9, 2012 / To be published also as LNCS*, Daejeon, KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF, 11 2012.

[4] Rui Min, Jongmoo Choi, Gérard Medioni, and Jean-Luc Dugelay. Real-time 3D face identification from a depth camera. In *ICPR 2012, 21st International Conference on Pattern Recognition, November 11-15, 2012, Tsukuba International Congress Center, Tsukuba Science City, Japan*, Tsukuba, JAPAN, 11 2012.

[5] Rui Min and Jean-Luc Dugelay. Inpainting of sparse occlusion in face recognition. In *ICIP 2012, IEEE International Conference on Image Processing, 30 September-3 October, 2012, Orlando, Florida, USA*, Orlando, UNITED STATES, 09 2012.

[6] Rui Min and Jean-Luc Dugelay. Cap detection for moving people in entrance surveillance. In *MM 2011, 19th ACM International Conference on Multimedia, 28 November-1 December, 2011, Scottsdale, USA*, Scottsdale, UNITED STATES, 11 2011.

[7] Rui Min and Jean-Luc Dugelay. Improved combination of LBP and sparse representation based classification (SRC) for face recognition. In *ICME 2011, IEEE International Conference for Multimedia and Expo, July 11-15, 2011, Barcelona, Spain*, Barcelona, SPAIN, 07 2011.

[8] Carmelo Velardo, Jean-Luc Dugelay, Lionel Daniel, Antitza Dantcheva, Nesli Erdogmus, Neslihan Kose, Rui Min, and Xuran Zhao. *Introduction to biometry*. Book chapter in "Multimedia Image and Video Processing" (2nd revised edition); CRC Press; 2012; ISBN:978-1439830864, 02 2011.

[9] Rui Min, Abdenour Hadid, and Jean-Luc Dugelay. Improving the recognition of faces occluded by facial accessories. In *FG 2011, 9th IEEE Conference on Automatic Face and Gesture Recognition, March 21-25, 2011, Santa Barbara, CA, USA*, Santa Barbara, UNITED STATES, 03 2011.

[10] Mourad Ouaret, Antitza Dantcheva, Rui Min, Lionel Daniel, and Jean-Luc Dugelay. BIOFACE, a biometric face demonstrator. In *ACMMM 2010, ACM Multimedia 2010, October 25-29, 2010, Firenze, Italy*, Firenze, ITALY, 10 2010.

[11] Rui Min, Angela D'angelo, and Jean-Luc Dugelay. Efficient scarf detection prior to face recognition. In *EUSIPCO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark*, Aalborg, DENMARK, 08 2010.

List of Figures

1.1 Examples of facial occlusion in different categories: (a) occlusions in daily life; (b) occlusion related to safety issues; (c) occlusion related to security issues. . 3

2.1 Visualization of (a) Eigenface and (b) Fisherface constructed from Yale face database [92] (Image excerpted from [71]). 14
2.2 Visual illustration of LBP thresholding.(Image excerpted from [1]) 14
2.3 Illustration of the Cumulative Matching Characteristics (CMC) in the format of ROC curve. The solid line represents the system that performs better. N is the number of subjects in the gallery. (Image excerpted from [142]) 17
2.4 Illustration of typical examples of biometric system graphs, the two distributions (a) represent the client/impostor scores; by varying the threshold different values of FAR and FRR can be computed. A ROC curve (b) is used to summarize the operating points of a biometric system, for each different application different performances are required to the system. (Image excerpted from [142]) 18
2.5 Example of occlusions in the AR face database. 21
2.6 Illustration of the AMM algorithm: (a) a clean face; (b) an occluded face; (c) dividing the face image into 6 local components, and each component is independently modelled by a mixture of Gaussian model from a set of training samples. (Image excerpted from [107]) 24
2.7 Illustration of the ICA basis (upper) and the LS-ICA basis (lower): a face image can be represented as a linear combination of such basis. It is clear that the LS-ICA basis is more locally distributed. (Image excerpted from [89]) 25
2.8 Illustration (Image excerpted from [89]) 26
2.9 Illustration of the SRC algorithm in case of occlusion occurs: an (well aligned) occluded face can be expressed by a linear combination of faces in the gallery plus the measurement of erroneous on each pixel location. (Image excerpted from [146]) 29
2.10 Illustration of LBMF bases correspond to a non-occluded local patch (Image excerpted from [116]) 31
2.11 Illustration of the occlusion probability estimated using the KLD method (brighter block refers to higher occlusion probability). (Image excerpted from [156]) 32

3.1 Illustration of different types of facial occlusions: (a) ordinary facial occlusions in daily life; (b) facial occlusions related to severe security issues (ATM crimes, football hooligans etc.). . . . 36
3.2 Illustration of different types of facial occlusions: (a) ordinary facial occlusions in daily life; (b) facial occlusions related to severe security issues (ATM crimes, football hooligans etc.). . . . 37
3.3 Overview of occlusion detection in local patches. . . . 38
3.4 Real part of the 40 Gabor wavelets and their magnitudes in five scales. . . . 39
3.5 An example image and the 40 GMPs filtered by Gabor Wavelets. . . . 40
3.6 Illustration of the distributions of 150 occluded (red crosses) and 150 non-occluded faces (blue circles) in the eigenspace. . . . 40
3.7 Examples of our own occlusion dataset with different scarf appearances. . . . 42
3.8 Illustration of our occlusion segmentation: (a) examples of faces occluded by scarf and sunglasses; (b) initial guess of the observation set according to the results from our occlusion detector; (c)(d) are the visualization of $u(i, j)$ in horizontal and vertical directions respectively; (e) the generated occlusion masks ($\omega = 150$). . . . 44
3.9 Flowchart of applying occlusion detection to improve LBP based face recognition under occlusion conditions. . . . 45
3.10 Example of extracting the LBP histogram from the non-occluded facial regions. 46
3.11 The face images are divided into 64 block for LBP based face recognition. . . . 47
3.12 Recognition performance of different methods on three test sets: non-occluded faces, face occluded with scarf and faces occluded with sunglasses. . . . 48
3.13 Illustration of the proposed selective-LGBPHS approach for face recognition in occluded conditions. . . . 50
3.14 Results of PCA, OA-PCA, LBP, OA-LBP, LGBPHS, KLD-LGBPHS, OA-LGBPHS and RSC on three different testing sets (faces are clean and faces occluded by scarf and sunglasses). . . . 52

4.1 Examples of various kinds of facial occlusions: (a) densely occluded faces, (b) sparsely occluded faces. . . . 57
4.2 A high-level work-flow of the proposed method. . . . 60
4.3 A visual interpretation of the robust-PCA (excerpt from [14]). Left: matrix of corrupted observations, in our case, the first column is the occluded faces, all other columns are the canonical faces. Middle: the underling low-rank structure of the clean face subspace. Right: the reconstructed sparse error matrix, in our case, the first column contains large errors caused by occlusion. 61
4.4 Illustration of our sparse occlusion inpainting: (a) faces with different sparse occlusions (stain, text, orthogonal grid, and diagonal grid), PSNR=19.12 dB, 13.92 dB, 13.25 dB, 12.81 dB ; (b) ground truth masks of the sparse occlusions; (c) results of our sparse occlusion detection (τ = 0.004); (d) faces after inpainting using the masks in (b), PSNR=39.12 dB, 34.05 dB, 33.26 dB, 32.51 dB; (e) faces after inpainting using the masks in (c), PSNR=30.43 dB, 28.30 dB, 26.05 dB, 26.50 dB. . . . 63

4.5 Face recognition results based on PCA. 64
4.6 Face recognition results based on SIFT. 65
4.7 Face recognition results based on LBP. 65
4.8 Illustration of Entrance surveillance scenarios. 67
4.9 Illustration of Entrance surveillance scenarios. 68
4.10 Background estimated from 60 frames using LMedS. 69
4.11 The proposed head segmentation method. Left: a silhouette image. Right: its horizontal projection. 70
4.12 The scalable elliptical head tracker: the head size is growing along time according to our monotonic increment assumption. 71
4.13 Perceptual classification of the proposed features (Blue: faces without cap, Red: faces with cap). 72
4.14 Dissimilarity Matrices of 40 video clips (1-20 without cap, 21-40 with cap, cold color indicates low dissimilarity). 73
4.15 Pre-defined ROI 1-3 (from left to right: the whole ellipse, the face region, the upper-eyebrow region). 74

5.1 Demographics of KinectFaceDB validation partition by: (a) gender, (b) age, and (c) ethnicity. 82
5.2 Illustration of different facial variations acquired in our database: (a) neutral face; (b) smiling; (c) mouth open; (d) strong illumination; (e) occlusion by sunglasses; (f) occlusion by hand; (g) occlusion by paper; (h) right face profile and (i) left face profile. (Upper: the RGB images. Lower: the depth maps aligned with above RGB images.) 83
5.3 Acquisition environment for the Kinect face database. 83
5.4 Architecture of a Kinect sensor for RGB-D sensing. 84
5.5 Illustration of the 3D face data: (a) visualization (rescaling to displayable range $[0, 255]$) of a depth map from the pre-cropped region; (b) the 3D points cloud retrieved from (a). 85
5.6 Illustration of the RGB-D alignment: the depth map (left) is aligned with the RGB image (right) captured by Kinect at the same time. 86
5.7 Illustration of color mapping on 3D points cloud of a given face: from left to right views. 86
5.8 The 6 facial anchor points: 1. left eye center, 2. right eye center, 3. nose-tip, 4. left mouth corner, 5. right mouth corner and 6. chin. 87
5.9 Cropped and smoothed 3D face captured by Kinect (upper row) and Minolta (lower row) of the same person, frontal view (left column) and side view (right column). The 3D face from Minolta keeps more details (wrinkles, eyelids etc.). 91
5.10 Overview of the proposed solution for face recognition under occlusion conditions based on Kinect sensor. 93
5.11 The mean depth model for occlusion estimation. 97

5.12 Illustration of the proposed occlusion probability estimation: (a) the intensity image; (b) the depth image; (c) the computed occlusion map from (b); (d) visualization of the occlusion probabilities for different local regions. Row 1 to row 3 corresponds to 3 different types of occlusions (sunglasses, hand and paper). 98
5.13 Recognition rates of LGBP, PRGB-LGBP and PD-LGBP on RGB images. 99
5.14 Recognition rates of LGBP, PRGB-LGBP and PD-LGBP on depth images. . . . 100

6.1 The procedure of block based face representation. 107
6.2 Architecture of the proposed approach. 109
6.3 Recognition rates of four LBP based methods (LBP+NN, LBP+PCA+SRC, LBP+LDA+SRC and D&C+LBP+SRC) on three different division strategies (2×2, 4×4 and 8×8 sub-blocks). 113
6.4 Recognition rates of the proposed approach (D&C+LBP+SRC with the pyramid architecture) in different pyramid levels (8×8, $8 \times 8 + 4 \times 4$ and $8 \times 8 + 4 \times 4 + 2 \times 2$). 113
6.5 Architecture of one-to-all registration (left), architecture of registration to a AFM or ICS (middle), architecture of the proposed registration (right). 115
6.6 System overview. 116
6.7 3D face registration to a set of canonical faces. 117
6.8 The Cumulative Match Characteristic (CMC) using 1-5 canonical faces for registration (M=1:5). 120
6.9 Rank-1 identification rates of the "Image-to-Image" approach and the "Video-to-Image" approach based on 1-5 canonical faces (M=1:5) for registration. . . 120
6.10 Time complexity analysis on large scale dataset (produced by data duplication). 121

List of Tables

2.1 Summary of literature works in occluded face recognition. 23

3.1 Results of occlusion detection 41
3.2 The results on the proposed dataset. 42
3.3 OUR APPROACH VS. S-LNMF [116]. 49
3.4 Robustness to different facial variations. 53

4.1 Results of the proposed system. 74

5.1 Summary of Off-the-Shelf 3D Face Databases. 81
5.2 2D Face Recognition using PCA. 88
5.3 2D Face Recognition using LBP. 88
5.4 2.5D Face Recognition using PCA. 89
5.5 2.5D Face Recognition using LBP. 89
5.6 3D Face Recognition using TPS warping parameters. 89
5.7 Fusion of RGB and Depth for Face Recognition using PCA. 90
5.8 Fusion of RGB and Depth for Face Recognition using LBP. 90
5.9 KinectFaceDB vs. FRGC 91

Une Revue en Français de mes Publications de Thèse

Détection efficace des écharpes avant reconnaissance du visage

Rui Min, Angela D'angelo, and Jean-Luc Dugelay.
In EUSIPCO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark, Aalborg, DENMARK, 08 2010.

L'occultation du visage est un challenge dans la reconnaissance du visage. La performance du système de reconnaissance du visage peut diminuer considérablement en raison de la présence d'occultations partielles sur le visage. Une approche pour surmonter ce problème consiste d'abord à pré-classifier les visages en deux catégories : le visage net et le visage caché; ensuite, visages dans différentes classes sont traités par des systèmes différents de la reconnaissance. Dans ce cas, un algorithme qui est capable de détecter automatiquement la présence d'occultations sur le visage serait un outil utile pour augmenter les performances du système. Dans cet article, nous présentons un algorithme de détection d'écharpe. Dans les résultats expérimentaux , les performances de l'algorithme sont présentés et comparés avec des systèmes de l'état de l'art.

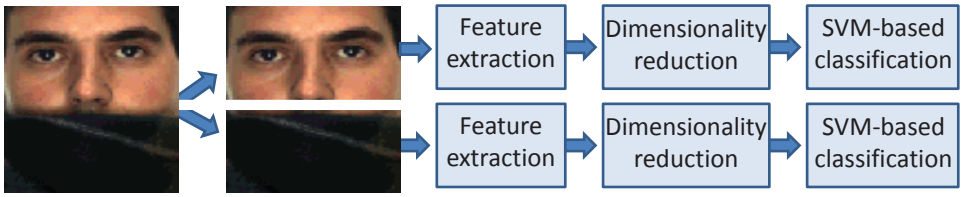


Illustration de la détection des écharpes proposé: l'extraction de caractéristiques, la réduction de la dimension et la classification.

Amélioration de la reconnaissance des visages occluté par des accessoires faciaux

Rui Min, Abdenour Hadid, and Jean-Luc Dugelay.
In FG 2011, 9th IEEE Conference on Automatic Face and Gesture Recognition, March 21-25, 2011, Santa Barbara, CA, USA, Santa Barbara, UNITED STATES, 03 2011.

Les occultations du visage, dues par exemple à des lunettes de soleil, chapeaux, écharpes, barbes, etc, peuvent affecter de manière significative les performances d'un système de reconnaissance faciale. Malheureusement, la présence d'occultations sur le visage est assez fréquent dans les applications du monde réel en particulier lorsque les personnes ne sont pas coopératives avec le système comme dans les scénarios de vidéosurveillance. Alors qu'il y a eu une quantité énorme de recherches sur la reconnaissance de visage avec des changements de pose/d'illumination et des dégradations de l'image, des problèmes causés par les occultations sont souvent négligés. L'objectif de cet article est donc de se concentrer sur les occultations du visage, et en particulier sur la façon d'améliorer la reconnaissance des visages occultés par des lunettes de soleil ou une écharpe. Nous proposons une approche efficace qui consiste à d'abord détecter la présence des écharpes/lunettes de soleil et traitant ensuite des régions faciales non-cachée seuls. Le problème de la détection d'occultation est abordée à l'aide d'ondelettes de Gabor, PCA et support vector machines (SVM), tandis que la reconnaissance de la partie faciale visible est effectuée en utilisant des "local binary patterns" basés sur des blocs. Des expériences sur des données de la base de visages AR ont montré que les méthodes proposées apportent une amélioration significative des performances par rapport aux travaux existants pour reconnaître les visages particialement cachés ou pas. Par ailleurs, la performance de l'approche proposée est également évaluée sous les changements d'illumination et d'expression du visage extrêmes, démontrant des résultats intéressants.

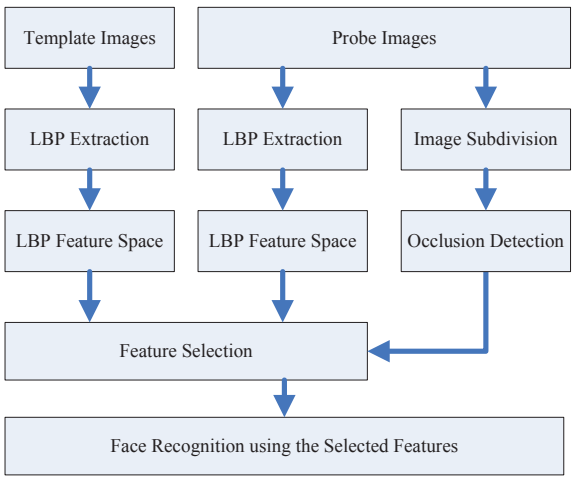


Illustration du système de reconnaissance de visage proposée.

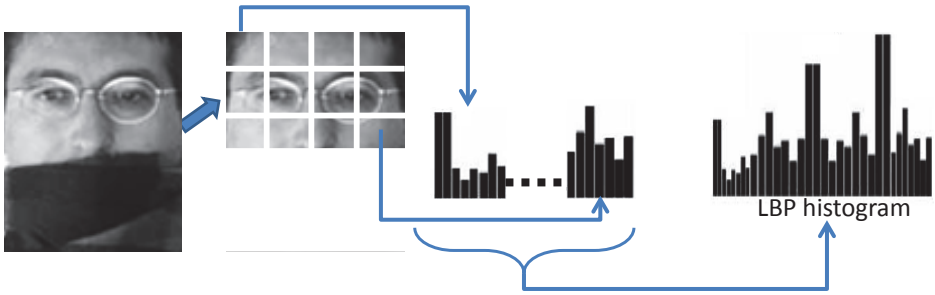


Illustration de l'extraction de l'histogramme LBP des régions faciales non-cachées

Détection efficace de l'occultation avant la reconnaissance robuste du visage

Rui Min, Abdenour Hadid, and Jean-Luc Dugelay.
The Scientific World Journal : Signal Processing (Accepted).

Pour qu'il y a eu une énorme quantité de recherches sur la reconnaissance de visages sous changements des pose/illumination/expression et les dégradations de l'image, des problèmes causés par les occultations ont attirés relativement moins d'attention. Les occultations des visage, dues par exemple à des lunettes de soleil, chapeau/casquette, écharpe, barbe, etc, peuvent détériorer de manière significative les performances des systèmes de reconnaissance faciale dans des environnements non contrôlés comme la surveillance vidéo. Le but de cet article est d'explorer la reconnaissance du visage en présence d'occultations partielles, en insistant sur des scénarios du monde réel (e.g. lunettes de soleil et écharpe). Dans cet article, nous proposons une approche efficace qui consiste d'abord à analyser la présence potentiel d'occultation sur un visage, et puis à effectuer la reconnaissance faciale sur les régions faciales visibles basé sur local Gabor binary patterns sélectifs. Les résultats expérimentaux montrent que la méthode proposée surpasse les travaux de l'état de l'art, y compris KLD-LBPHS, S-LNMF, OA-LBP et RSC. De plus, les performances de l'approche proposée qui sont évalué sous les changements extrêmes d'illumination et d'expression du visage, donnent des résultats aussi significatifs.

Remplissage d'occultations épars en la reconnaissance faciale

Rui Min and Jean-Luc Dugelay.
In ICIP 2012, IEEE International Conference on Image Processing, 30 September-3 October, 2012, Orlando, Florida, USA, Orlando, UNITED STATES, 09 2012.

L'occultation des visages est un problème crucial dans de nombreuses applications de reconnaissance faciale. Les approches existantes de la reconnaissance faciale dans des conditions d'occultation se concentrent principalement sur les accessoires faciaux traditionnels (tels que

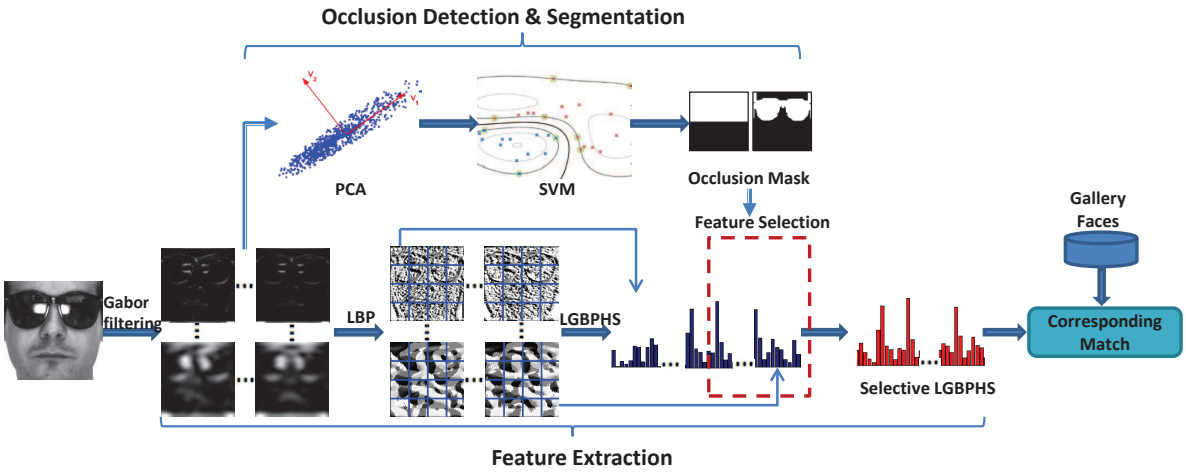


Illustration de la méthode de reconnaissance faciale proposée.



Illustration de la segmentation d'occultation proposé: (a) Exemples de visages cachés par écharpe et lunettes de soleil; (b) supposition initiale de l'observation fixée en fonction des résultats de détection d'occultation; (c) (d) la visualisation horizontale et verticale des structure; (e) les masques d'occultation générés.

lunettes de soleil et écharpe) et donc suppose que la région occluse est dense et contiguë. Pourtant, en raison de la grande variété de sources naturelles qui peut occulter un visage humain dans des environnements non contrôlés, les méthodes basées sur l'hypothèse dense ne sont pas robustes aux occultations minces et distribuées de façon aléatoire. Cet article présente la solution à un problème d'occultation du visage nouvellement identifiée – occultation éparse dans le contexte de la biométrie du visage en vidéo-surveillance. Nous montrons que les pixels manquants peuvent être détectées dans les rangs faibles d'une structure canonique d'un visage selon un cadre Robust-PCA; et la partie cachées peut être retouchée basée uniquement sur la partie visible et un Fields-of-Experts prior via inférences spatiales. Les expériences montrent que l'approche proposée améliorent de manière significative différents algorithmes de reconnaissance de visage en présence d'occultations éparses complexes.

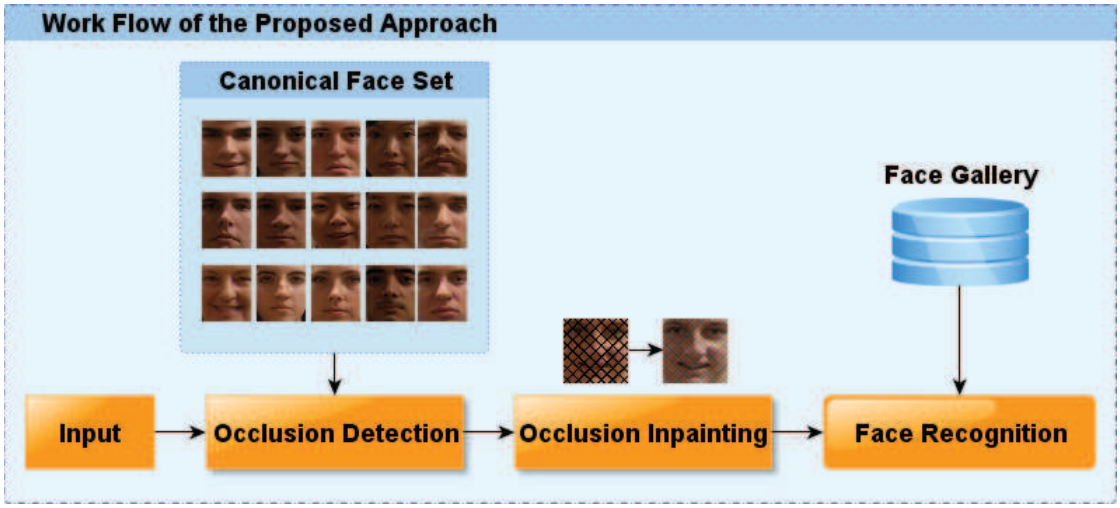


Illustration de la méthode proposée.

Détection de casquettes sur personnes en movement surveillance dans une entrée

Rui Min and Jean-Luc Dugelay.
In MM 2011, 19th ACM International Conference on Multimedia, 28 November-1 December, 2011, Scottsdale, USA, Scottsdale, UNITED STATES, 11 2011.

Tandis qu'il y a eu une énorme quantité de recherches sur la reconnaissance de visages sous changements de pose/illumination et dégradations de l'image, des problèmes causés par les occultations sont souvent négligé. En outre, la plupart des approches existantes de la reconnaissance faciale dans des conditions d'occultation se concentre sur la résolution des problèmes d'occultations du visage dues aux lunettes de soleil et écharpe. Selon nos connaissances, l'occultation due à une casquette n'a jamais été étudiée dans la littérature, mais l'importance de ce problème doit être souligné, car il est connu que les voleurs de banque et des hooligans de football en profitent pour dissimuler leurs visages. Cet article présente une solution à ce problème d'occultation du visage nouvellement identifiée - occultation variable en temps due à une casquette pour de la surveillance de l'entrée, dans le contexte de la biométrie du visage en vidéo-surveillance. L'approche proposée se compose de deux parties: la détection et le suivi des visages cachés dans des sénario complexes du vidéo-surveillance; détection de la présence de casquette par exploitation de l'information temporelle. La partie de détection et de suivi est basé sur la silhouette du corps et traqueur de tête elliptique. Le classement des casquette/ non-casquette des visages utilise "dynamic time warping" (DTW) et regroupement agglomératif hiérarchique. L'algorithme proposé est évalué sur plusieurs vidéo-surveillance et donne de bons taux de détection.

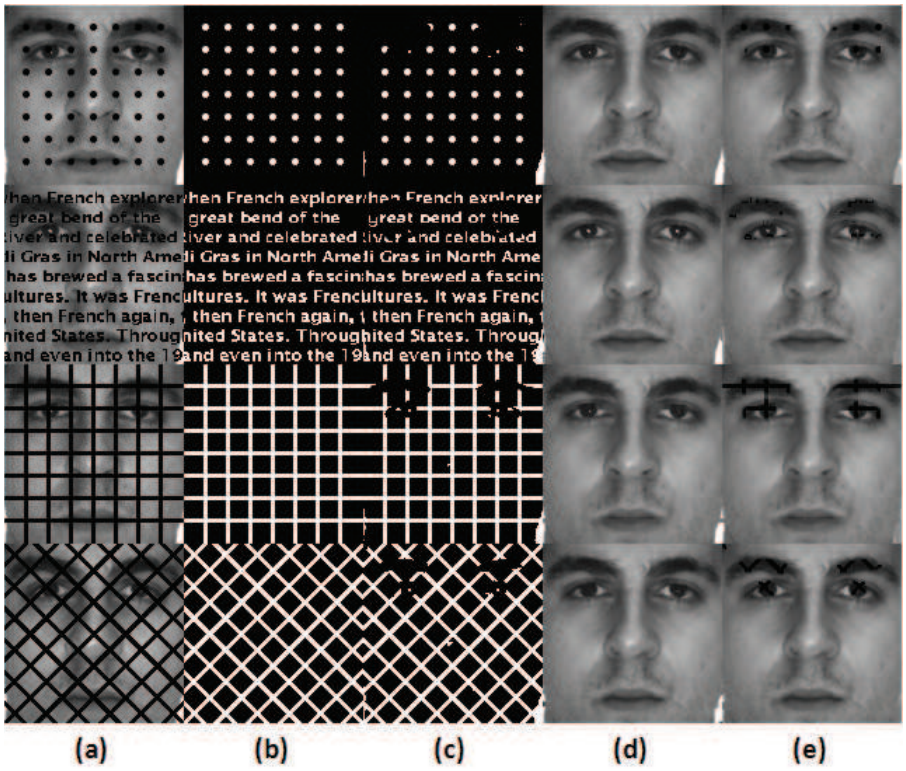


Illustration de notre remplissage d'occultations éparses: (a) visages avec différenes occultations éparses (tache, texte, grille orthogonale, et grille diagonale), PSNR = 19,12 dB, 13,92 dB, dB 13.25, 12.81 dB; (b) masques de vérité de terrain d'occultation éparses; (c) résultats de notre détection d'occultation éparses ; (d) des faces après remplissage à l'aide des masques dans (b), le PSNR = 39,12 dB, 34,05 dB, 33,26 dB, 32,51 dB; (e) visages après remplissage à l'aide des masques dans (c), le PSNR = 30,43 dB, 28,30 dB, 26,05 dB, 26,50 dB.

KinectFaceDB: une base de données de visage Kinect pour la reconnaissance faciale

Rui Min, Kose Neslihan, and Jean-Luc Dugelay.
IEEE Transaction on System (under revision)

Le succès récent des caméras RGB-D comme Kinect ouvre de larges perspectives de données en 3 dimensions informatique. Toutefois, en raison de l'absence de données de base de test standard, il est difficile d'évaluer combien la reconnaissance faciale peut bénéficier de cette technologie. Afin d'établir le lien entre Kinect et la recherche en reconnaissance faciale, dans cet article, nous présentons la première base de données de visage publiquement disponible (KinectFaceDB) basé sur le capteur Kinect pour la reconnaissance faciale. La base de données se compose de différentes modalités des données (bien alignes et traitées en 2D, 2.5D, 3D et vidéos) et de multiples variations du visage. Nous effectuons les évaluations sur la base de données proposée en utilisant des techniques de reconnaissance faciale de base, et démonstrons le gain en performance de Depth en RGB via la fusion au niveau des

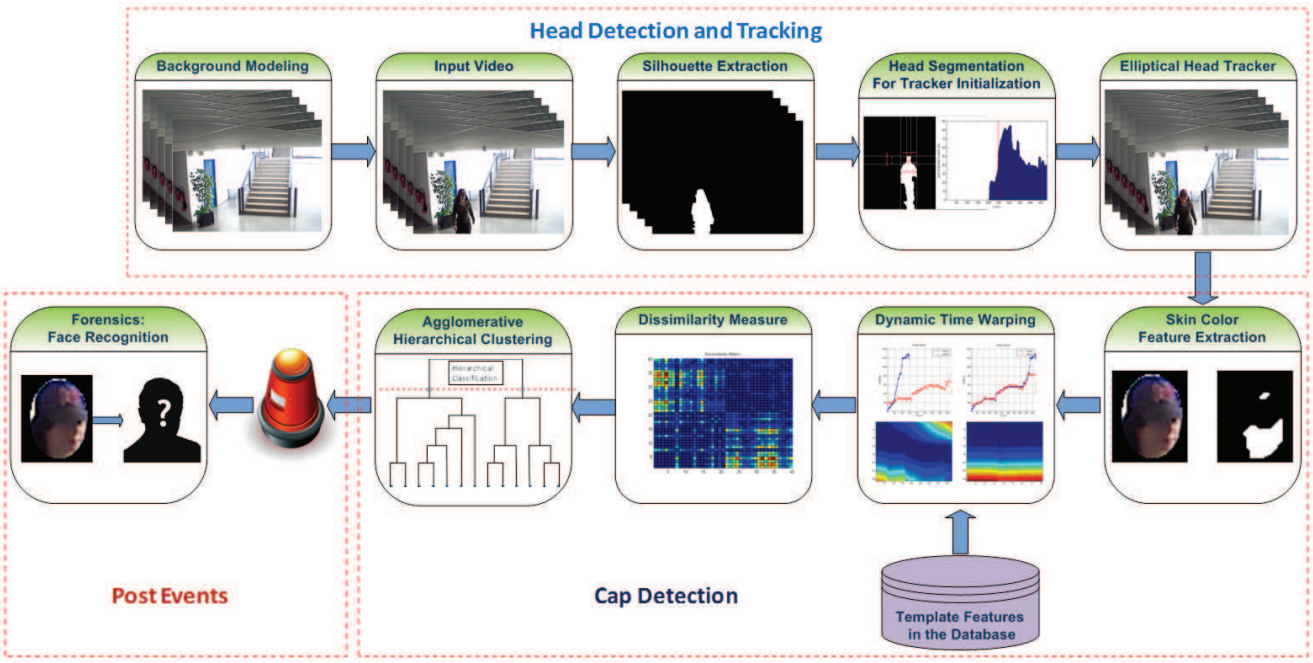


Illustration du système de détection de casquette proposée.

scores. Notre étude rapporte également les comparaisons en performances entre les images de Kinect (KinectFaceDB) et scans 3D traditionnels de meilleure qualité (FRGC) dans le cadre du visage en biométrie, et montre des résultats intéressants.

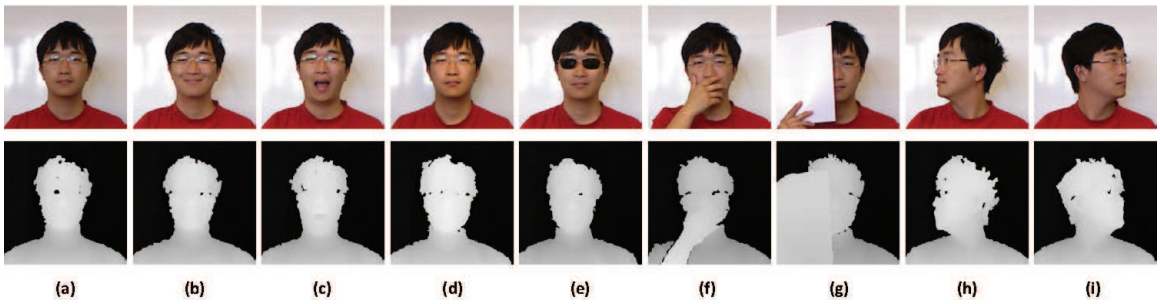


Illustration de la base de données de visage de Kinect proposé.

L'analyse d'occultation du visage pour la reconnaissance du visage robuste basée Kinect

Rui Min, Jean-Luc Dugelay.
Hot3D 2013, 4th IEEE International Workshop on Hot Topics in 3D, July 15, 2013, San Jose, CA, USA.

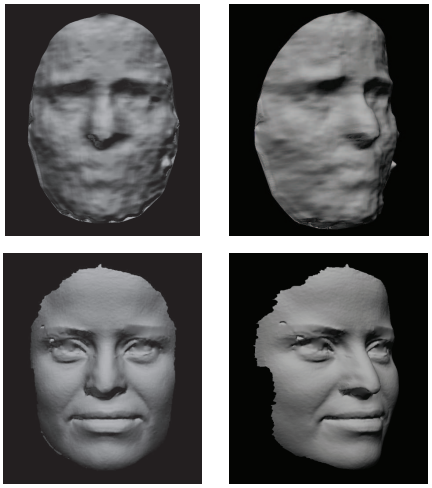


Illustration des données de visage de mauvaise qualité obtenues par Kinect (rangée supérieure) et les données de visage de meuilleur qualité obtenues par Minolta (rangée inférieure) de la même personne.

Nous présentons une approche pour résoudre le problème d'occultations partielles en reconnaissance faciale en utilisant le capteur Kinect. Les méthodes traditionnelles appliquant l'analyse de l'occultation afin d'améliorer la reconnaissance des visages exploitent l'information homogène dans les deux étapes (2D→2D ou 3D→3D). Au lieu de cela, nous utilisons des indices hétérogènes (3D→2D) afin d'améliorer la reconnaissance faciale en présence d'occultations. L'approche proposée mène d'abord l'analyse d'occultation basés sur l'image de profondeur, et puis utiliser cette information pour améliorer la reconnaissance du visage basée LGBP via une stratégie de pondétation. Nous avons construit une base de données de visage Kinect publiquement disponible pour tester la méthode proposée. Les résultats montrent que des améliorations significatives sont atteintes par rapport à LGBP et KLD-LGBP.

Amélioration de la combinaison de LBP et classification basé sur ésparse represenation (SRC) pour la reconnaissance faciale.

Rui Min and Jean-Luc Dugelay. In ICME 2011, IEEE International Conference for Multimedia and Expo, July 11-15, 2011, Barcelona, Spain, Barcelona, SPAIN, 07 2011.

Récemment, des descripteurs basés sur des bariantes du "local binary patterns" (LBP) et une classification basée sur la ésparse represenation (SRC) deviennent à la fois techniques éminents en reconnaissance du visage. Des Techniques préliminaires de la combinaison de LBP et SRC ont été proposées dans la littérature. Cependant, la méthode l'état de l'art souffre de la curse of dimensionality pour les scénarios du monde réel. Dans cet article, un nouvel algorithme de reconnaissance de visage combinant LBP avec SRC est proposé, dans lequel le problème de la dimensionnalité est résolu par "divide-and-conquer" et le pouvoir discriminant est renforcé par son architecture pyramidale. La méthode de reconnaissance



Illustration de l'estimation de la probabilité d'occultation proposée: (a) l'image d'intensité; (b) l'image de profondeur; (c) le plan d'occultation calculée à partir de (b), (d) la visualisation des probabilités d'occultation pour différentes régions locales.

faciale proposée est évaluée sur la base de donnée AR de visages et donne des résultats très impressionnants.

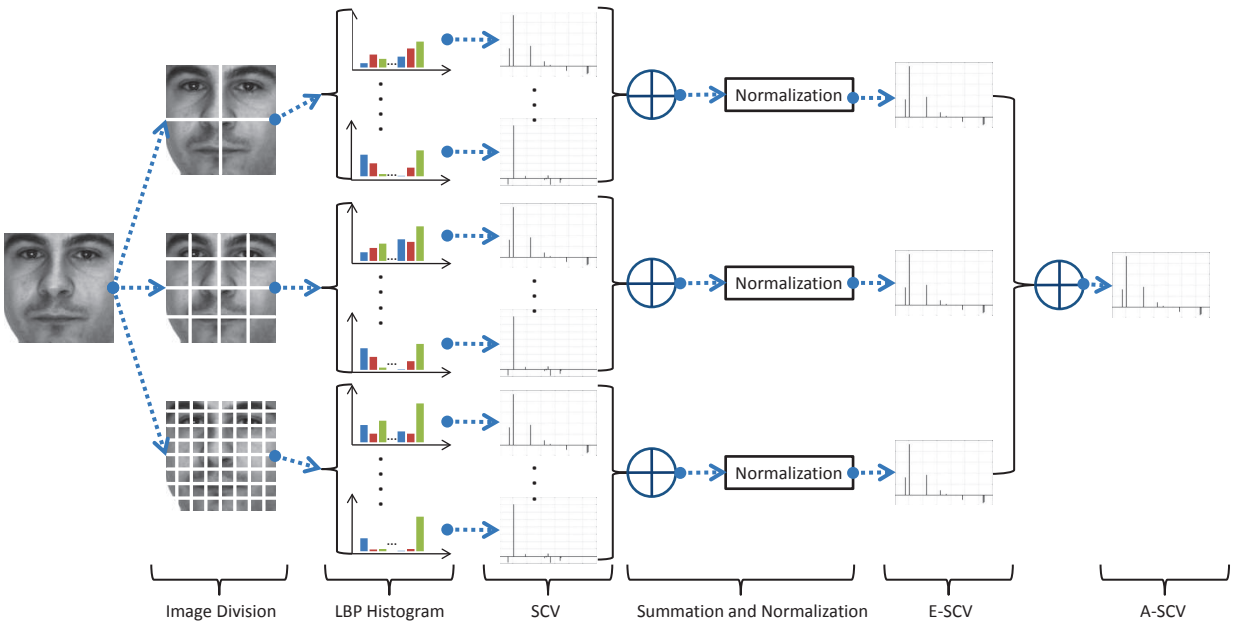


Illustration de l'approche proposée.

Real-Time 3D identification faciale en utilisant d'une caméra Depth

Rui Min, Jongmoo Choi, Gérard Medioni, and Jean-Luc Dugelay.
In ICPR 2012, 21st International Conference on Pattern Recognition, November 11-15, 2012, Tsukuba International Congress Center, Tsukuba Science City, Japan, Tsukuba, JAPAN, 11 2012.

Nous présentons un système 3D temps réel d'identification faciale à l'aide d'une caméra de profondeur grand public (PrimeSensor). Notre système prend une séquence bruitée comme entrée et produit une identification fiable. Au lieu d'enregistrer une example à toutes les instances de la base de données, nous proposons de seulement l'enregistrer avec plusieurs références intermédiaires, ce qui réduit considérablement le traitement, tout en préservant le taux de reconnaissance. Le système présenté réalise régulièrement des taux de 100% d'identification lors de l'appariement d'une séquence vidéo (0.5-4 secondes), et 97,9% pour la reconnaissance d'image isolée. Ces chiffres se réfèrent à un ensemble de données du monde réel de 20 personnes. La méthodologie s'étend directement à de très grandes bases de données. Le processus fonctionne à 20 fps sur un ordinateur portable.

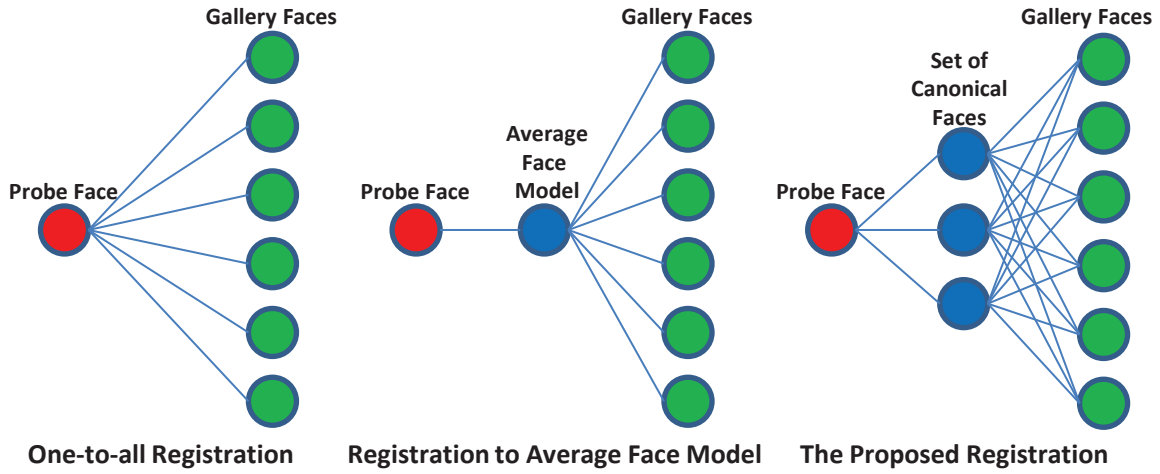


Illustration de registre one-to-all (à gauche), registre à un AFM ou ICS (au milieu), et registre proposé (à droite).

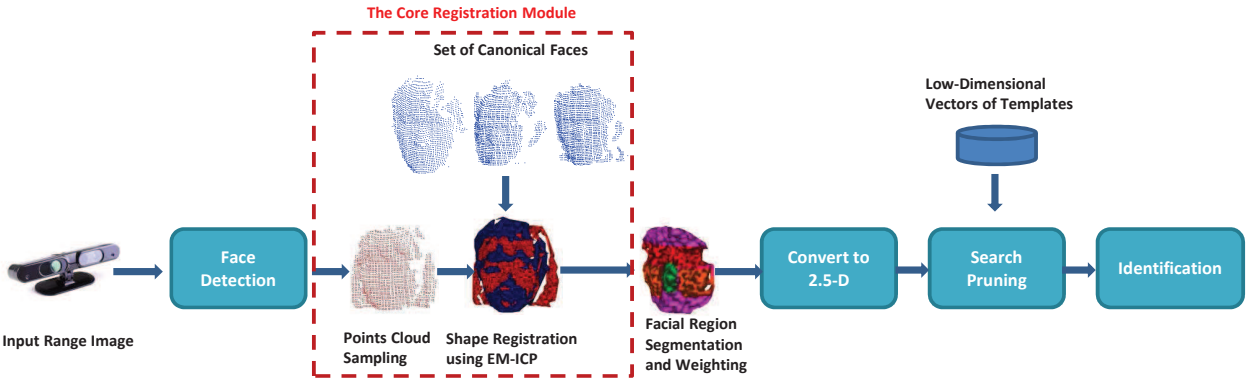


Illustration du système proposé.

Review of my Publications in the Thesis

Efficient Scarf Detection Prior to Face Recognition

Rui Min, Angela D'angelo, and Jean-Luc Dugelay.
In EUSIPCO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark, Aalborg, DENMARK, 08 2010.

Face occlusion is a very challenging problem in face recognition. The performance of face recognition system can decrease drastically due to the presence of partial occlusion on the face. One approach to overcome this problem is to first pre-classify faces into two classes: the clean face and the occluded face; then faces in different classes are treated by different recognition systems. In this case an algorithm which is able to automatically detect the presence of occlusions on the face will be a useful tool to increase the performances of the system. In this paper we present a scarf detection algorithm. In the experimental results the performances of the algorithm are reported and compared with state of the art systems.

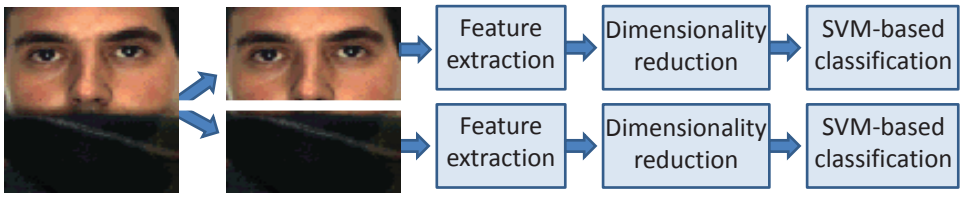


Illustration of the proposed scarf detection: feature extraction, dimensionality reduction and classification.

Improving the Recognition of Faces Occluded by Facial Accessories

Rui Min, Abdenour Hadid, and Jean-Luc Dugelay.
In FG 2011, 9th IEEE Conference on Automatic Face and Gesture Recognition, March 21-25, 2011, Santa Barbara, CA, USA, Santa Barbara, UNITED STATES, 03 2011.

Facial occlusions, due for example to sunglasses, hats, scarf, beards etc., can significantly affect the performance of any face recognition system. Unfortunately, the presence of facial occlusions is quite common in real-world applications especially when the individuals are not cooperative with the system such as in video surveillance scenarios. While there has been an enormous amount of research on face recognition under pose/illumination changes and image degradations, problems caused by occlusions are mostly overlooked. The focus of this paper is thus on facial occlusions, and particularly on how to improve the recognition of faces occluded by sunglasses and scarf. We propose an efficient approach which consists of first detecting the presence of scarf/sunglasses and then processing the non-occluded facial regions only. The occlusion detection problem is approached using Gabor wavelets, PCA and support vector machines (SVM), while the recognition of the non-occluded facial part is performed using block-based local binary patterns. Experiments on AR face database showed that the proposed method yields significant performance improvements compared to existing works for recognizing partially occluded and also non-occluded faces. Furthermore, the performance of the proposed approach is also assessed under illumination and extreme facial expression changes, demonstrating interesting results.

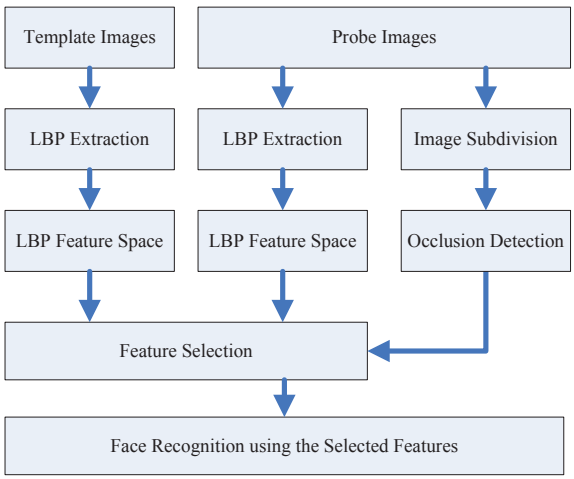


Illustration of the proposed face recognition system.

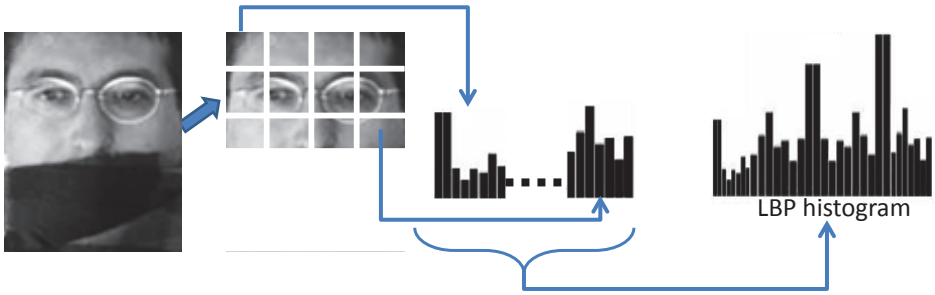


Illustration of extracting the LBP histogram from the non-occluded facial regions.

Efficient detection of occlusion prior to robust face recognition

Rui Min, Abdenour Hadid, and Jean-Luc Dugelay.
The Scientific World Journal : Signal Processing (Accepted).

While there has been an enormous amount of research on face recognition under pose/illumination/expression changes and image degradations, problems caused by occlusions are attracted relatively less attention. Facial occlusions, due for example to sunglasses, hat/cap, scarf, beard etc., can significantly deteriorate performances of face recognition systems in uncontrolled environments such as video surveillance. The goal of this paper is to explore face recognition in presence of partial occlusions, with emphasis on real-world scenarios (e.g. sunglasses and scarf). In this paper, we propose an efficient approach which consists of first analysing the presence of potential occlusion on a face, and then conducting face recognition on the non-occluded facial regions based on selective local Gabor binary patterns. Experiments demonstrate that the proposed method outperforms the state-of-the-art works including KLD-LGBPHS, S-LNMF, OA-LBP and RSC. Furthermore, performances of the proposed approach are evaluated under illumination and extreme facial expression changes, provide also significant results.

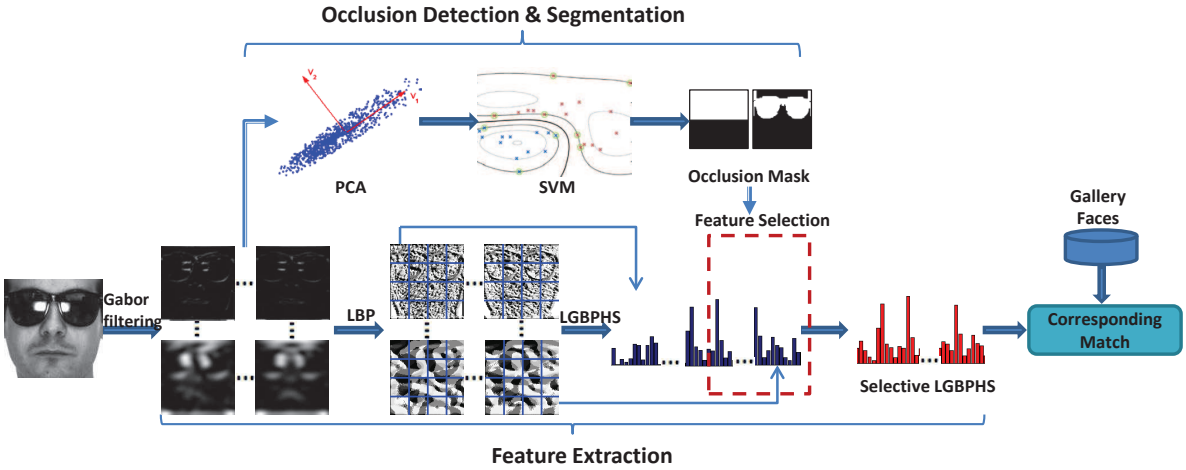


Illustration of the proposed face recognition method.



Illustration of the proposed occlusion segmentation: (a) examples of faces occluded by scarf and sunglasses; (b) initial guess of the observation set according to the results from occlusion detection; (c)(d) are the visualization of horizontal and vertical structural prior; (e) the generated occlusion masks.

Inpainting of Sparse Occlusion in Face Recognition

Rui Min and Jean-Luc Dugelay.
In ICIP 2012, IEEE International Conference on Image Processing, 30 September-3 October, 2012, Orlando, Florida, USA, Orlando, UNITED STATES, 09 2012.

Facial occlusion is a critical issue in many face recognition applications. Existing approaches of face recognition under occlusion conditions mainly focus on the conventional facial accessories (such as sunglasses and scarf) and thus presume that the occluded region is dense and contiguous. Yet due to the wide variety of natural sources which can occlude a human face in uncontrolled environments, methods based on the dense assumption are not robust to thin and randomly distributed occlusions. This paper presents the solution to a newly identified facial occlusion problem – sparse occlusion in the context of face biometrics in video surveillance. We show that the occluded pixels can be detected in the low-rank structure of a canonical face set under the Robust-PCA framework; and the occluded part can be inpainted solely based on the non-occluded part and a Fields-of-Experts prior via spatial inference. Experiments demonstrate that the proposed approach significantly improve various face recognition algorithms in presence of complex sparse occlusions.

Cap Detection for Moving People in Entrance Surveillance

Rui Min and Jean-Luc Dugelay.
In MM 2011, 19th ACM International Conference on Multimedia, 28 November-1 December, 2011, Scottsdale, USA, Scottsdale, UNITED STATES, 11 2011.

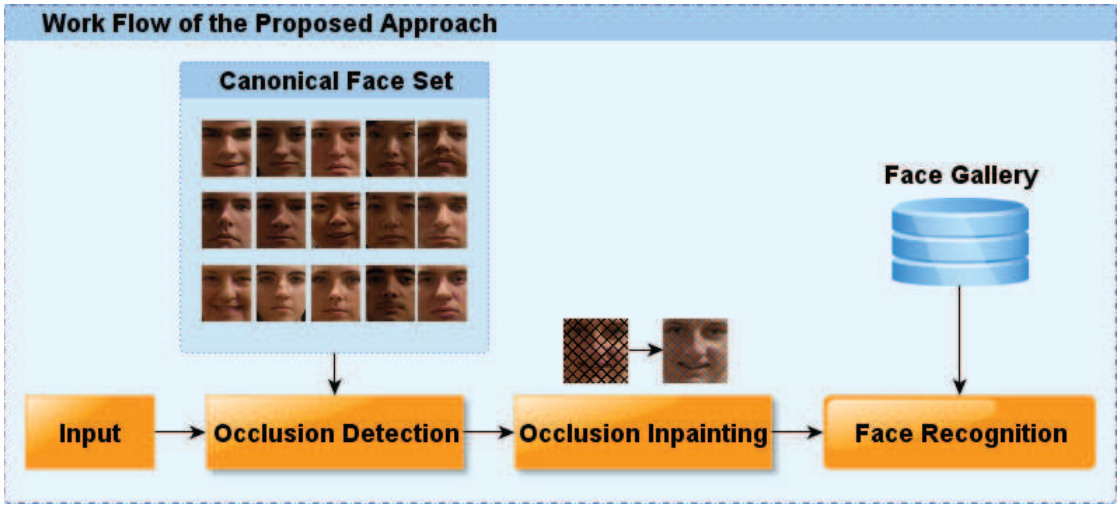


Illustration of the proposed method.

While there has been an enormous amount of research on face recognition under pose/illumination changes and image degradations, problems caused by occlusions are mostly overlooked. Moreover, most of the existing approaches of face recognition under occlusion conditions focus on overcoming facial occlusion problems due to sunglasses and scarf. To the best of our knowledge, occlusion due to cap has never been studied in the literature, but the importance of this problem should be emphasized since it is known that bank robbers and football hooligans take advantage of it for hiding their faces. This paper presents a solution to this newly identified face occlusion problem – the time-variant occlusion due to cap in entrance surveillance, in the context of face biometrics in video surveillance. The proposed approach consists of two parts: detection and tracking of occluded faces in complex surveillance videos; detecting the presence of cap by exploiting temporal information. The detection and tracking part is based upon body silhouette and elliptical head tracker. The classification of cap/non-cap faces utilizes dynamic time warping (DTW) and agglomerative hierarchical clustering. The proposed algorithm is evaluated on several surveillance videos and yields good detection rates.

KinectFaceDB: a Kinect Face Database for Face Recognition

Rui Min, Kose Neslihan, and Jean-Luc Dugelay.
IEEE Transaction on System (under revision)

The recent success of emerging RGB-D cameras such as Kinect depicts a broad prospects of 3-Dimensional data based computer applications. However, due to the lack of standard testing dataset, it is difficult to evaluate how much face recognition may benefit from this technology. In order to establish the connection between Kinect and face recognition researches, in this paper, we present the first publicly available face database (KinectFaceDB1)

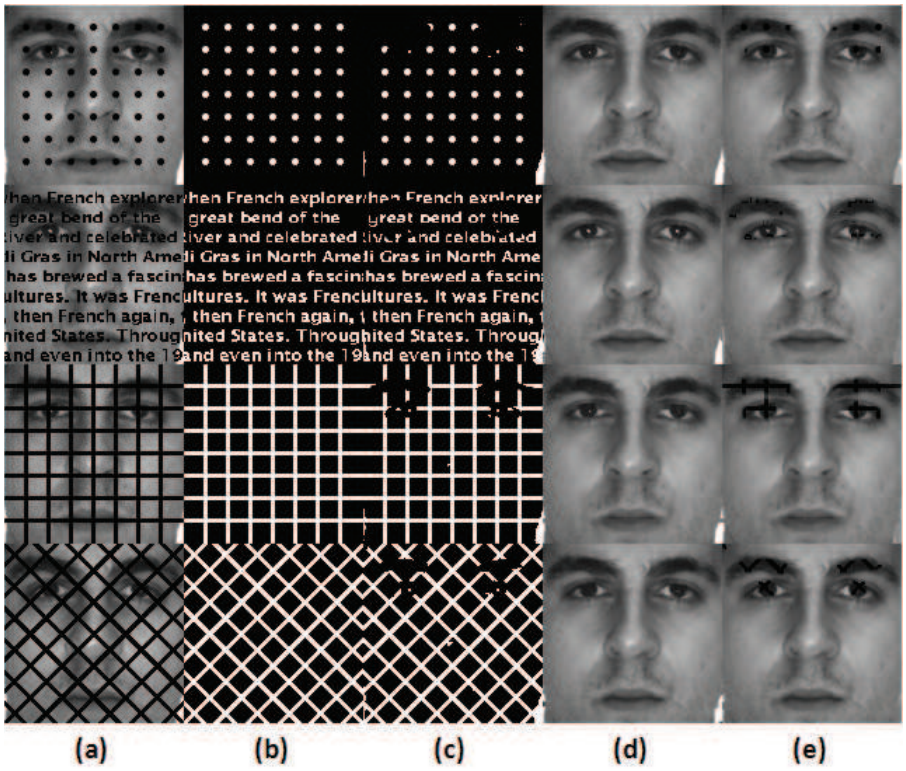


Illustration of our sparse occlusion inpainting: (a) faces with different sparse occlusions (stain, text, orthogonal grid, and diagonal grid), PSNR=19.12 dB, 13.92 dB, 13.25 dB, 12.81 dB ; (b) ground truth masks of the sparse occlusions; (c) results of our sparse occlusion detection; (d) faces after inpainting using the masks in (b), PSNR=39.12 dB, 34.05 dB, 33.26 dB, 32.51 dB; (e) faces after inpainting using the masks in (c), PSNR=30.43 dB, 28.30 dB, 26.05 dB, 26.50 dB.

based on Kinect sensor for face recognition. The database consists of different data modalities (well-aligned and processed 2D, 2.5D, 3D and video based face data) and multiple facial variations. We conduct benchmark evaluations on the proposed database using baseline face recognition techniques, and demonstrate the performance gain from Depth to RGB via score-level fusion. Our study also reports the performance comparisons between Kinect images (KinectFaceDB) and traditional high quality 3D scans (FRGC) in the context of face biometrics, demonstrates interesting results.

Kinect based Facial Occlusion Analysis for Robust Face Recognition

Rui Min, Jean-Luc Dugelay.
Hot3D 2013, 4th IEEE International Workshop on Hot Topics in 3D, July 15, 2013, San Jose, CA, USA.

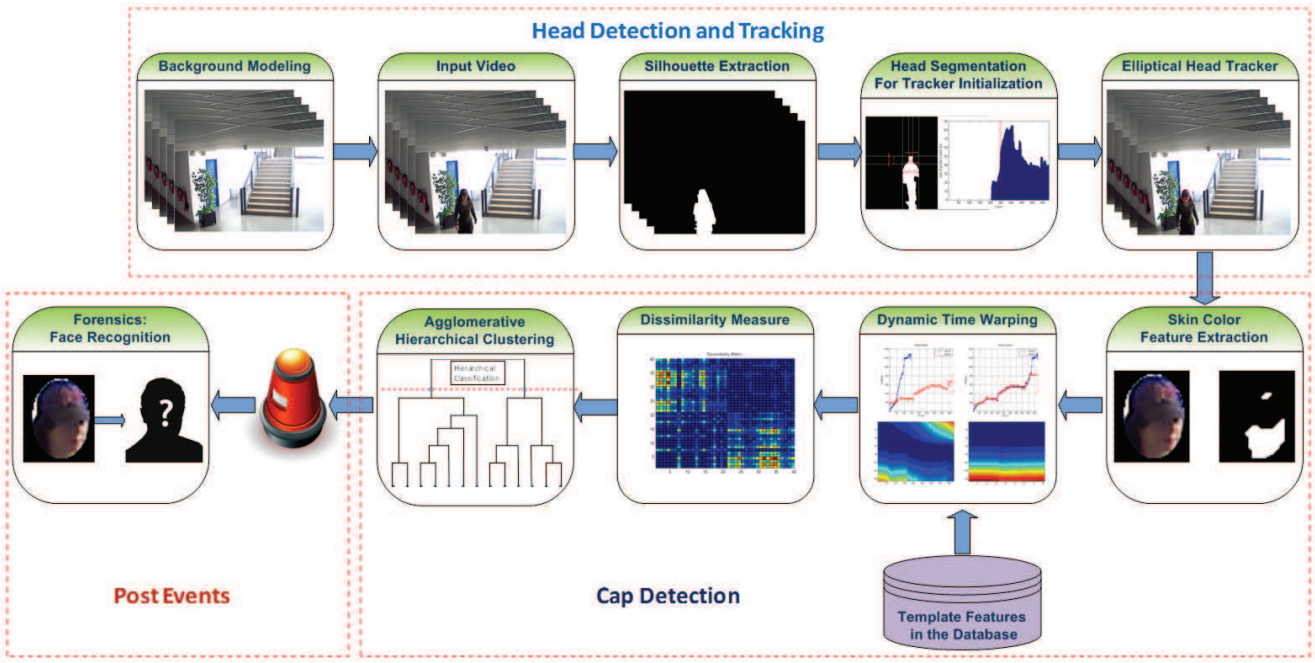


Illustration of the proposed cap detection system.

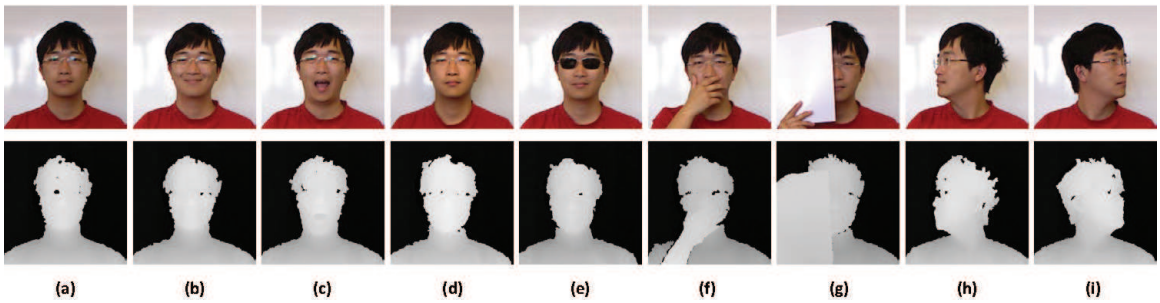


Illustration of the proposed Kinect face database.

We present an approach to address the partial occlusion problem in face recognition using the Kinect sensor. Traditional methods applying occlusion analysis to improve face recognition exploits homogeneous information in both steps (2D→2D or 3D→3D). Instead, we use heterogeneous cues (3D→2D) to improve face recognition in presence of occlusions. The proposed approach first conducts occlusion analysis based on depth image, then use this information to improve LGBP based face recognition via a weighting strategy. We built a publicly available Kinect face database to test the proposed method. Results show that significant improvements are achieved in comparison to LGBP and KLD-LGBP.

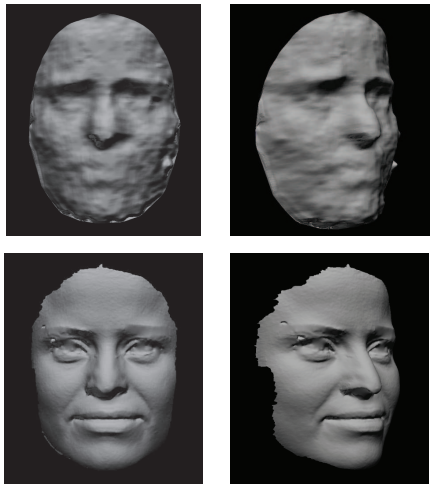


Illustration of the low quality face data obtained by Kinect (upper row) and the high quality face data obtained by Minolta (lower row) of the same person.

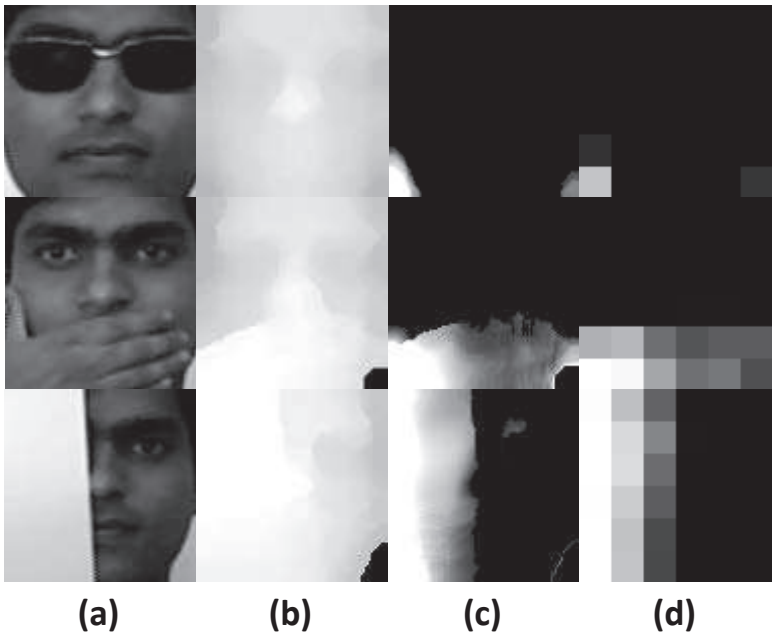


Illustration of the proposed occlusion probability estimation: (a) the intensity image; (b) the depth image; (c) the computed occlusion map from (b); (d) visualization of the occlusion probabilities for different local regions.

Improved Combination of LBP and Sparse Represenation based Classification (SRC) for Face Recognition

Rui Min and Jean-Luc Dugelay. In ICME 2011, IEEE International Conference for Multimedia and Expo, July 11-15, 2011, Barcelona, Spain, Barcelona, SPAIN, 07 2011.

Recently, local binary patterns (LBP) based descriptors and sparse representation based classification (SRC) become both eminent techniques in face recognition. Preliminary techniques

of combining LBP and SRC have been proposed in the literature. However, the state-of-art method suffers from the "curse of dimensionality" for real world scenarios. In this paper, a novel face recognition algorithm of combining LBP with SRC is proposed; in which the dimensionality problem is resolved by divide-and-conquer and the discriminative power is strengthen via its pyramid architecture. The proposed face recognition method is evaluated on AR Face Database and yields very impressive results.

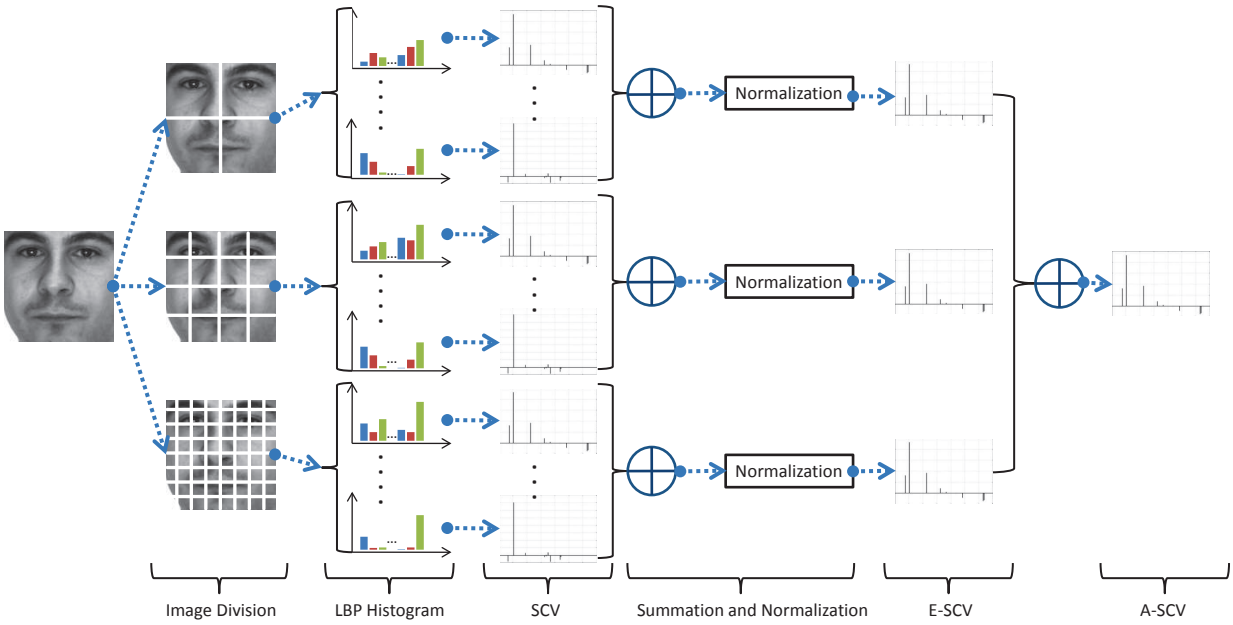


Illustration of the proposed approach.

Real-Time 3D Face Identification from a Depth Camera

Rui Min, Jongmoo Choi, Gérard Medioni, and Jean-Luc Dugelay.
In ICPR 2012, 21st International Conference on Pattern Recognition, November 11-15, 2012, Tsukuba International Congress Center, Tsukuba Science City, Japan, Tsukuba, JAPAN, 11 2012.

We present a real-time 3D face identification system using a consumer level depth camera (PrimeSensor). Our system takes a noisy sequence as input and produces reliable identification. Instead of registering a probe to all instances in the database, we propose to only register it with several intermediate references, which considerably reduces processing, while preserving the recognition rate. The presented system routinely achieves 100% identification rate when matching a (0.5-4 seconds) video sequence, and 97.9% for single frame recognition. These numbers refer to a real-world dataset of 20 people. The methodology extends directly to very large datasets. The process runs at 20fps on an off the shelf laptop.

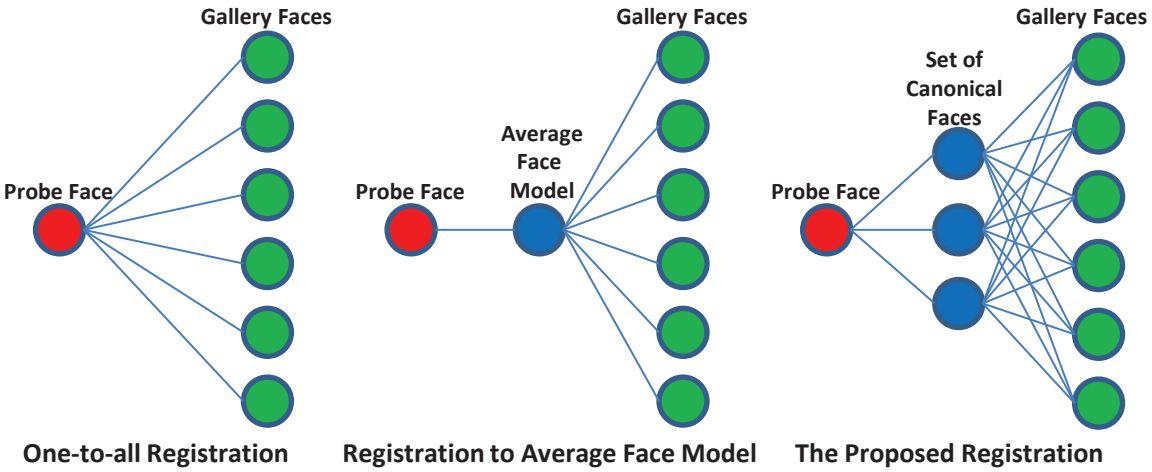


Illustration of one-to-all registration (left), registration to a AFM or ICS (middle), and the proposed registration (right).

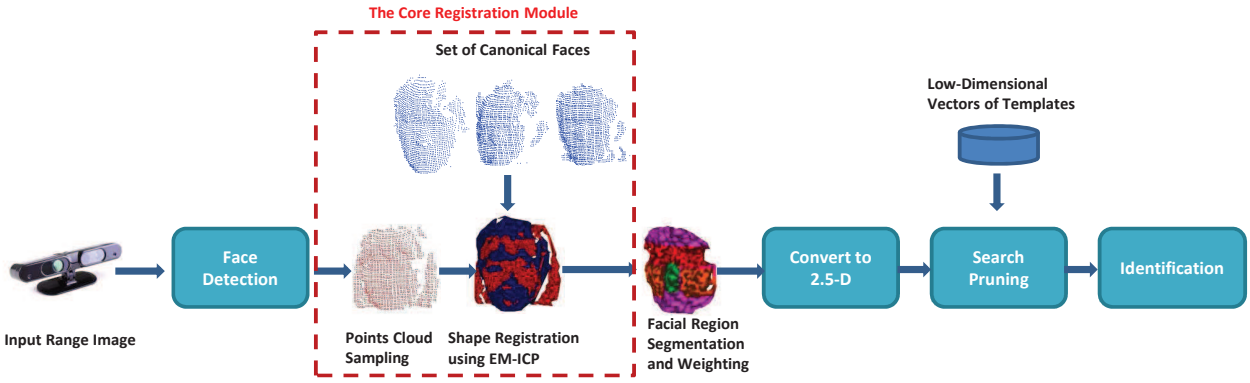


Illustration of the proposed system.