Dengpan Mou

Machine-based Intelligent Face Recognition

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Dengpan Mou

Machine-based Intelligent Face Recognition

Dengpan Mou

Machine-based Intelligent Face Recognition

With 58 figures

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Author
Dengpan Mou
Harman/Becker Automotive Systems GmbH
Becker-Goering-Strasse 16
D-76307, Karlsbad
Germany
E-mail: dengpan.mou@harman.com

ISBN 978-7-04-022355-2
Higher Education Press, Beijing

ISBN 978-3-642-00750-7 e-ISBN 978-3-642-00751-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009922608

Cover design: Frido Steinen-Broo, EStudio Calamar, Spain

Printed on acid-free paper

Springer is a part of Springer Science + Business Media (www.springer.com)

Preface

We can't solve the problems by using the same kind of thinking we used when we created them.
Albert Einstein (1879−1955)

State-of-the-art machine-based face recognition technology, although booming since last decades, is still suffering a lot from critical research challenges, such as the lack of fundamental intelligence, the difficulties of running completely automatically and unsupervisedly without separate training, and the typical failures of dealing with free face pose variations, etc. Those limitations greatly hinder the wide applications it could have had. This book is the first to discuss the general engineering methods of imitating intelligent human brains for video-based face recognition. The advances and evidences from the cognitive science research are introduced in this book, which further strengthen our thoughts and proposals to achieve such a fundamental intelligence in machine vision.

Regarding intelligence, we have defined two directions. The first effort is to simulate the ability of self-learning, self-matching and self-updating. This side of intelligence can be detailed into the following features: the whole recognition procedure is running in an unsupervised, automatic, non-invasive, and self-updated way. It is important to note that, the fully automatic procedure is a generalized face recognition procedure, which includes the task of enrollment (training) and updating as well. However, those steps are typically separate and supervised in machine learning, and therefore missing the essentials of intelligence.

The other main focus of the book is to explore the novel ways on how to implement the high-level analysis in machine-based face recognition, to simu-

late the process in human brains. Through high-level analysis, it is possible to combine multiple available methods, which include conventional machine learning algorithms, image processing approaches, predefined rules, video context, temporal and spatial correlations and even logic deduction. The fusion of multiple approaches contributes significantly to the improved face recognition performance.

Experiments are made through long-term (over years) constructed sequences with more than 30 specific subjects and more than 20 faces from TV news channels. The evaluation results demonstrate the robustness of the proposals in unconstraint video scenarios.

The objective of the author is to provide this book for scientists, researchers and students in the areas of machine-based face recognition. The fundamentals and research backgrounds are provided, aiming to help the beginners to quickly step into the field. Introduction and analysis of the state-of-the-art technology can assist experts to easily keep up to date with the world-wide overview.

The author does hope that the proposals to achieve both intelligence and robustness could be somehow helpful for other researchers, to finally popularize the technology, and to pervasively apply it for designing general machine learning algorithms.

Dengpan Mou
Villingen-Schwenningen, Germany
April 2009

Acknowledgements

Many research results of this book derive from my PhD research in the University of Ulm, Germany, during 2000-2004. Therefore, my first and deepest gratitude is to my PhD advisor, Prof. Dr.-Ing. Albrecht Rothermel, for his invaluable supervision and support. I am deeply grateful to Dr. Rainer Schweer and his group from Deutsche Thomson-Brandt GmbH for stimulating discussions and financially supporting me with such a promising research direction. I would like to express my warmest thanks to Prof. Dr. Heiko Neumann, for his earnest review and comments to shape my PhD thesis.

In writing this book, I have been fortunate to benefit from valuable research discussions with a large number of research groups through my technical reports in many different universities and conferences during the last several years. Although there is no space to mention the hundreds of researchers and scientists for their helpful comments, I would like to take this opportunity to generally thank them all.

My deep acknowledgements go to the following previous colleagues and friends: Dr.-Ing. Ralf Altherr, Dr.-Ing. Markus Buck, Dr.-Ing. Sviatoslav Bulach, Markus Bschorr, Zhen Chang, Richard Geißler, Christian Günter, Roland Hacker, Frank Hagmeyer, Stefan Hirsch, Dr.-Ing. Thomas Kumpf, Martin Lallinger, Dr.-Ing. Roland Lares, Cheng Miao, Dr.-Ing. Oliver Pfänder, Dr.-Ing. Ivan Perisa, Markus Prokein, Xavier Queffelec, Dr.-Ing. Wolfgang Schlecker, Dr.-Ing. Karsten Schmidt, Ralf Schreier, Walter Schweigart, Lei Wang, Lin Wang, Yi Wang, and Chi Zhang for their generous help and contributions to the face databases.

I would like to thank my colleagues at Harman/Becker Automotive Systems GmbH: Michael Maier and Christopher Schmidtmann, who broaden my vision with their extensive industrial research and development experiences.

I am so grateful to my editor, Dr. Liu Ying, from Higher Education Press,

for her great encouragement and generous supports over the years.

In particular, I thank my dearest parents and wife Xiaoying Ge for their endless love and support. Without these emotional and mental supports, the completion of the book could not be possible. This book is dedicated to them.

Contents

List of Figures

List of Tables

1 Introduction

Abstract Face recognition is quite intuitive for most human beings while rather too complex for a machine vision system to be pervasively applied. As the starting point of this book, we are discussing the fundamental pros and cons of state-of-the-art machine vision system versus its human being counterpart. Inspired from the comparison, in the second part of this chapter, a machine-based intelligent face recognition system is briefly introduced. With the high intelligence and high recognition robustness, the system could be widely used for prospective applications such as future smart home, automotive and children mobile phones. In the third part, the discussion of those applications is aiming at broading the visions of the readers. Finally, the outline of the rest part of this book is given.

1.1 Face Recognition—Machine Versus Human

Face recognition has been studied for many years and its attraction grows rapidly due to the wide and promising applications, such as daily identification systems (e.g. automatic banking, access control, computer log-in, etc.), in communication systems (e.g. teleconference and video-phone, etc.), in public security systems (e.g. criminal identification, digital driver license, etc.) and in law enforcement systems.

Face recognition seems instinctive for human beings, but it is really a tough and complex task for a machine-based system. Over the decades, scientists in cognitive and neuroscience are always trying to explore how human

beings recognize faces and why we are in general good at recognizing faces. The recent advances of their research are enlightening [1], but their contributions to the mathematical models and engineering solutions for a machine vision system are still far from enough. Therefore, researchers from the computer science are constructing vast numbers of mathematical models and dedicated algorithms, which may cover the field of artificial intelligence, machine learning, image processing and even video signal processing, etc.

Machine vision systems could have several major advantages over humans. It can have a huge storage medium to deal with much larger amount of people. Furthermore, a machine is not easy to be tired and can work 24 hours a day without any problems. It is always expected to replace the human resources to significantly lower the cost. More importantly, the use of machine can keep the privacy. It is always fair and can 100% follow the predefined rules. For example, suppose that a completely automatic system is required to alarm strangers. A machine will not keep the information of known people and therefore is not violating their privacy. But a human supervisor cannot 100% objectively manage it during the monitoring. And in this case, probably all people under supervision would prefer a machine than a human supervisor.

The most recent face recognition tests [1, 2] claim that, in a particular case, when non-familiar faces (frontal) are tested with significant illumination variations, machine-based face recognition algorithms could be superior to human brains. Nevertheless, in terms of the general recognition precision, the most successful state-of-the-art face recognition technology is still not able to compete with the level of human systems. From the engineering point of view, there is too little attention to the research on face recognition by imitating a human being in a fundamental way, which could combine every means for recognition, not necessarily based on pure psychophysical/neurobiological science or pure mathematical models. For example, it is just a piece of cake for a nine-year old child to recognize people when they turn their heads from frontal to profile views in a video sequence, but can lead to failures for most of current face recognition systems. The reason behind it might be that, in this special case, a child applies his/her multiple approaches including “image processing”, video context, logic deduction, experiences, etc. to recognize the people, while those machine-based systems are merely using the mathematical-based image processing methods to calculate the correlation.

Another crucial drawback of most current machine recognition tech-

niques is their lack of intelligence. Many systems are not able to memorize the faces by themselves without the help of a human supervisor or the cooperation of the users. Any new user is supposed to follow some instructions to be enrolled into the databases. The instructions are normally from a human supervisor who is also required to select and update the databases. They are hence invasive to the users and are difficult to be applied in many areas like consumer electronics. Quite recently, there are some advances [3, 4, 5] in the research of building an automatic face recognition system. But their assumptions restrict the real applications. The most critical one is that they normally assume only one person existing in a video sequence, which greatly decreases the complexity of the automatic procedure. With the same previous example, the child has no difficulty in self-learning and memorizing unknown faces, identifying known faces even if several people existing with free behaviors. But that would make a typical error for those systems.

Robustness is the most concerned question for the researchers. Recent face recognition surveys [6, 7] reveal that lighting changes, indoor/outdoor changes, pose variations and elapsed time databases are the critical parameters which greatly influence the performance of a face recognition system. But the effect from those parameters is significantly database dependent. If a database has already enrolled different mugshots under various environments and can update with recent views, state-of-the-art face recognition techniques can produce robust enough results.

Inspired from the above, we define an ideal machine-based face recognition system, which is unique due to the following features:

- Self-learning: completely automatic and unsupervised;
- Non-invasive;
- Robust: in unconstrained environment, against pose and lighting changes, occlusion and aging problems.

In this book, we are exploring the ways and algorithms to approach such an ideal system, mainly but not limited to the application of consumer electronics.

1.2 Proposed Approach

The rule-based algorithms for computer vision and pattern recognition used to

be popular more than 15 years ago. For example, to detect faces, people used eyes, noses, mouths and other features to define the corresponding rules. When there are several facial features detected, and the location of the features agrees with some predefined rules, a face is expected to exist. Although intuitive, these rules are not robust enough to deal with partial occlusions, pose and other frequent day-to-day face changes when applied in images. Learn-based methods therefore appeal more attractions. Nearly all the best face detection and recognition algorithms from the state-of-the-art are learn-based. However, the machine learning method generally suffers from requiring huge numbers of examples for training which makes it difficult to build an automatic system without any human supervision or user's cooperation. In this book, we explore the ways to combine some general rules as major methods together with learn-based methods to achieve such a system for face recognition in video.

For the convenience of describing the proposed approaches, we list the overview of the functional blocks of the automatic system [8] in Fig. 1.1.

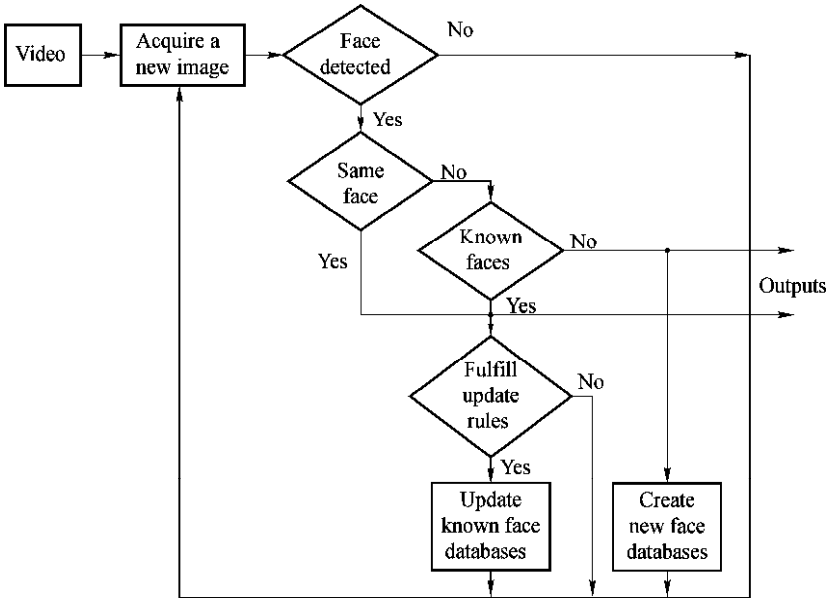


Fig. 1.1 Overview of the System Functional Blocks

Images are continuously acquired from a video source. An image-based face detector and a novel temporal-based face tracker are included to detect whether there are faces in the current image. Any typical image-based face de-

tection techniques (mostly learn-based) can be applied here. The temporal information [9] derives from the limited day to day human moving speed, which is a reasonable processing speed assumption. The human face dimensions are defined according to the anatomical research results. They are therefore rule-based. The combined information from both greatly increases the detection rate although the image-based face detector alone can only detect nearly frontal faces.

The next block is the main block for the same person decision. It is used to determine whether a certain face in the current frame remains the same as in the past several frames. This block is a combination of three components: a face recognition classifier, the temporal-based face tracker and a filter. Any robust face recognition classifier can be included which is mostly learn-based as well. A linear filter with the filter length of n frames is designed to mainly handle a sudden lighting change, shot change and to remove some sporadic face negatives. It is based on predefined rules. A state machine is applied to define a complete list of all possible states produced by the three components. The majority voting method with some assisting rules is used to judge the transition between each state and to deliver the final output of the same face decision.

If it is guaranteed that the current face remains the same, we define it as a known face. Otherwise, the system compares it with all existing databases to find a match by a recognition classifier. After that, the global output notifies whether known faces or new faces are found. For a new face, a corresponding database is created. And for a known face, the update rules help to select qualified mugshots to be enrolled into the databases. A rapid growth is crucial when a database is newly created. The more enrolled mugshots, the higher the recognition quality while the more database redundancy. With more and more mugshots enrolled, we are then concentrating on the variety of the databases. Since faces seem quite different from one day to another, it is also important to keep up with the recent views of them. Those rules guarantee the purity, variety, rapidity, updatability and uniqueness of the databases [9, 10]. The implemented prototype shows robust performance for consumer applications [11, 12].

1.3 Prospective Applications

The proposed non-invasive and unsupervised face recognition system can find prospective applications in consumer electronics. Cumbers [13] describes a passive person identification system which can automatically memorize frequently visiting customers and welcome them. Lin [14] introduces how an adaptive recognition system is used to prevent children from viewing undesirable materials on the internet or TV. In this section, we further foresee three other examples of application.

1.3.1 Recognition in the Future Intelligent Home

The first promising application could be the pervasive recognition for the future intelligent home. The corresponding conceptual intelligent system architecture is given in Fig. 1.2.

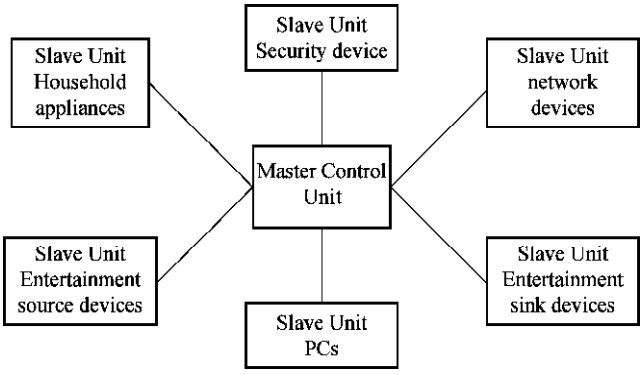


Fig. 1.2 Conceptual Architecture of an Intelligent Home System

The master unit controls all the slave units, which are grouped according to their different application purposes. The household appliances could include the washing machine, kitchen appliances with automatic cooking functions. The entertainment source devices can be audio, video and game sources, such as DVD players, etc.. The entertainment sink devices are displays, speakers, etc. The network devices provide the service for the whole internal intelligent system and external connections to the outside, e.g. through internet, through mobile communications, and cable telephones. The face recognition system can be a part of the security device, which could be connected to the intelligent network through

cable connections (such as Ethernet), interfaces (such as FireWire, USB, etc.), or wireless network connections (such as blue-tooth, wireless LAN, etc.).

The automatic face recognition is functioning as follows: It automatically learns the faces of family members through their daily behavior, even without their awareness. It can be defined to have two modes: a normal mode and a monitor mode. When none of the family members (or only none of the host/hostess) is at home, the system is set to be in a monitor mode, i.e. being alert to strangers. When an unknown face enters the home without accompany of any family members, his/her face shots are autonomously selected and stored in the database. An alarm signal is at the same time sent to the host/hostess who is not at home. Transmission of such an alarm signal as well as a qualified (ideally frontal view) mugshot from a stranger can be handled through the networks devices. For example, when the security system is connected to the home telephone cable, it can automatically dial the mobile phone number of the home owner to send him/her the alarm message as well as the mugshot of the stranger. It is also possible to obtain a higher resolution mugshot through an email via PCs. The owner can then decide whether to call the police for an urgent help. When the owner who is not at home doesn't want to be notified at all, the system is in a normal mode which is still learning new faces and updating known faces. Under this situation, update is therefore quite crucial to keep a low false alarm rate. For convenience, the system should be able to switch from one mode to the other not only by pressing the on-system button but also through a remote control when the owner is away from home.

Another application of the security device could be the keyless main entrance control. Family members that are automatically learned by the system can enter the building without the keys.

Face recognition in the security device might also be applied in entertainment, e.g. as a part of an electronic pet. It is shy to strangers but can learn to get to know him/her with a long enough interaction. It can show some friendly expressions or poses to welcome a known person. Aryananda [15] describes in more detail as such an example, a humanoid robot.

In interactive electronic games, automatic face recognition could be also promising to be applied. Any user would find it cool if the game character resembles him/her, even presenting the real-time facial expressions of the user. It will be just like personally on the scene. When multiple users are playing the game, automatic face recognition is absolutely required.

Face recognition can also be useful for the appliances to be automatically adapted to everyone's specific preferences. For example, each family member might have its own favorite settings of audio, video sink devices, their own preferred entertainment sources, and preferred TV channels, etc.

1.3.2 Automotive

Another interesting application might be an intelligent system inside cars.

A self-learning and adaptive face recognition system can contribute to different functions, such as guard against being stolen, setting of driver's preferences, and detection of driver's attention. The system can automatically learn its owner's face in different views and keep updating the database. The key of the car is not only used to open the door and ignite the engine, but also for keeping the intelligent system in a normal mode. When someone else enters the car without the key, the system is switched into an alarm mode. It can help to send the corresponding warning message to the owner who can decide whether to report to the police. The face shots of the unknown driver are stored in the database. Since GPS and mobile communication systems are more and more popular in cars, the mugshot as well as the car's current position can be sent by using the in-car mobile phone to the car owner if the alarm is verified by the car owner. In this case, the car is not able to drive faster than a certain limit and it can not restart any more when it stops. If the owner thinks it is a false alarm, he/she can use the mobile phone to switch the system back to the normal mode again. In normal mode, the system learns new faces which might be from the spouse, relatives or friends of the owner. It can then automatically memorize the preferences of different drivers. For some high-end cars, different types of keys are normally used to differentiate the preference settings for different drivers (normally a couple). But when the keys are exchanged, it doesn't work at all. With the help of the face recognition system, it is not so annoying any more. During driving, the system can also be used to learn and detect the normal status of eyes for each specific driver when he drives. The information might include the blink times per minute, the opening degrees of the eyes, etc. When a driver is sleepy, the status changes and a warning voice might occur to notify him. Obviously, it is much more precise to detect the usual status of eyes for a specific person than for a generic person model.

1.3.3 Mobile Phone for Children

Mobile phones are more and more widely used. Many parents would like to have their children taking it for security reasons when they are away from home. The mobile phone with the automatic face recognition function can be PIN-free. It is then easier to restrict the children to only call their parents. It can also help to frequently memorize the face shots of unknown persons who are staying with the children. The pictures can be automatically sent to the parents for security purposes.

1.4 Outline

The remainder of the book is arranged as follows.

Chapter 2 introduces the fundamentals and advances in the research of generalized biometric recognition, including both the cognitive-based recognition by human brains and the machine-based algorithms. Through the comparison of multiple biometric methods, we can get to learn the big advantages of using face recognition. Cognitive science research provides us many inspirations of achieving fundamental intelligence. The state-of-the-art research and limitations in machine-based face recognition, especially in video-based face recognition and unsupervised recognition systems, further support the motivations of our research.

Chapter 3 explores the algorithms on how to automatically extract faces of interest from live video input. Face region estimation, temporal-based face detection and tracking and the corresponding detection performance are mainly described in this chapter.

Chapter 4 covers the unsupervised face recognition procedure. Adaptive similarity threshold, temporal filter and the combined same face decision algorithms are explored.

Chapter 5 describes the ways of adaptively constructing the face databases. General supervised and unsupervised learning methods are firstly reviewed and then an improved clustering structure and an adaptive updating threshold are introduced. Finally, the features of an optimum database are proposed.

Chapter 6 introduces a state machine-based method to put all the detection,

recognition, and database construction procedures together as an automatic and self-learning system.

Chapter 7 focuses on implementation issues. Hardware configuration and software implementation are described. A list of additional technology dependent parameter settings is also included.

Chapter 8 demonstrates the performance of all previously proposed algorithms. As two main contributions, the combined same face decision algorithms and the database construction methods are evaluated respectively. The overall performance of the whole recognition system is given in the final part of the chapter, which demonstrates the robustness of our proposals.

Chapter 9 summarizes the contributions of the book and explores the future research directions. We are aiming at enlighting researchers to explore further for achieving the fundamental intelligence, and to apply the proposals dealing with general machine learning and recognition issues.

References

[1] P.Sinha,B.Balas, *et al.*:Face Recognition by Humans:Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of The IEEE,* Vol.94, No.11, 2006

[2] P.J.Phillips,W.T.Scruggs,*et al.*: FRVT 2006 and ICE 2006 Large-Scale Results. http://www.frvt.org, accessed 11 April 2007

[3] J. Weng, C. Evans, and W. Hwang: An Incremental Learning Method for Face Recognition under Continuous Video Stream. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp.251−256

[4] Q. Xiong and C. Jaynes: Mugshot Database Acquisition in Video Surveillance Networks Using Incremental Auto-Clustering Quality Measures. *Proceedings of the 2003 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, 2003, pp.191−198

[5] B. Raytchev and H. Murase: Unsupervised Face Recognition by Associative Chaining. *Pattern Recognition*, Vol. 36, No. 1, 2003, pp.245−257

[6] W. Zhao, R. Chellappa, *et al.*: Face Recognition: A Literature Survey. *Technical Report (CS-TR-4167R), University of Maryland.* ftp://ftp.cfar.umd.edu, accessed 20 August 2005

[7] P. J. Phillips, P. Grother, *et al.*: FRVT 2002 Evaluation Report. *Technical Report*. http://www.frvt.org, accessed 29 March 2007

[8] D. Mou, R. Schweer, and A. Rothermel: Automatisches System zur Erkennung und Verwaltung von Gesichtern bzw. Personen.*21. Jahrestagung der FKTG (Fernseh-und Kinotechnische Gesellschaft e.V.)*, May 2004, pp.24−27

[9] D. Mou, R. Schweer, and A. Rothermel: Automatic Databases for Unsupervised Face Recognition. *IEEE International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, Vol. 5, 2004, pp.90.

[10] D. Mou, R. Schweer, and A. Rothermel (2005): Face Recognition System and Method. *European Patent application,* EP05/001988, Feb 2005

[11] D. Mou, R. Schweer, and A. Rothermel: A Self-learning Video-based Face Recognition System. *Proceedings of IEEE International Conference on Consumer Electronics*, Jan. 2006

[12] D. Mou: Autonomous Face Recognition. Ph.D Dissertation, http://vts.uni-ulm.de/query/longview.meta.asp?document_id=5370, accessed 28 October 2005

[13] B. Cumbers (2003): Passive Biometric Customer Identification and Tracking System. *U.S. Patent*, 6554705, Apr. 2003

[14] Y.T. Lin: Adaptive Facial Recognition System and Method. *U.S. Patent application publication*, US2002/0136433, Sept. 2002

[15] Lijin Aryananda: Recognizing and Remembering Individuals: Online and Unsupervised Face Recognition for Humanoid Robot. *Proceedings of 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*,Vol. 2, 2002, pp.1202-1207

2 Fundamentals and Advances in Biometrics and Face Recognition

Abstract In this chapter, we mainly focus on the fundamentals and advances in the research of biometric recognition. In section 1, generalized biometric procedures and categories are defined and overviewed. This is followed in section 2 by the brief introduction and surveys on current cognitive science research. The essential point here is the fundamental intelligence of human brains. In section 3, machine-based biometric recognition tasks and methods are explored. As the marketing leader, fingerprint recognition is exclusively discussed in more details. Section 4 to 7 explores state-of-the-art research and limitations in machine-based face recognition, in video base-face recognition and in unsupervised recognition systems. This chapter ends by the summary as well as further thoughts inspired from both cognitive and machine recognition research.

2.1 Generalized Biometric Recognition

The term "biometrics" comes from the Greek words "bio" (means life) and "Metric" (means measure). It measures and analyzes the human's unique biological characteristics, either physically or behaviorally, in the purpose of recognition (authentication).

In the following, we define a typical 3-step generalized procedure of biometric recognition: data acquiring, data modeling and decision making.

- Data acquiring: biological characteristics are acquired for further analysis. This procedure can be accomplished via human being sensing (vision,

hearing, smell, etc.), machine sensing or physical samples.

- Data modeling: there are two purposes in data modeling. One is for storage and the other is for matching. The captured data are either to be stored in a database for future recognition or to be compared with the existing database. But both share the same data modeling type. The way of modeling could include: biological models which are quite close to original data, human nerve cell based models which are used by human brains, and mathematical models which are converted from the machine sensed data.
- Decision making: the acquired data for authentication are compared with the database to make the final decision of whether or not being recognized. The complexity of this matching procedure is highly dependent on how the data are modeled.

According to the different methods applied in biometric procedures, biometrics can be categorized in three groups: biology based, cognitive based, and mathematical-model based (machine based).

There are some special biological features, such as dental radiographs, DNA, etc., which have been successfully applied for human recognition in some specific forensic and medical purposes. This group requires physical samples and can be compared through pure biological methods. For instance, dental radiographs are exclusively important for identifying victims of massive disasters since in those cases, most of the traditional biometric features could not be applied. The identification can be examined through the comparison between the so called antemortem (AM) radiographs and postmortem (PM) radiographs. AM radiographs are acquired when a person is alive and the PM ones are from after his death. DNA (deoxyribonucleic acid) is another important biological feature for human recognition, which is substantially applied in law enforcement, court, paternity, and identifying victims. In those applications, only a specific part of the repetitive DNA regions, called DNA profile (DNA fingerprint), can be compared for identification or verification. Identical twins contain identical DNA profile. Except for that, DNA profile is considered to be the highly accurate technique since the probability of two random persons with the same DNA profile is small enough from the perspective of forensic science. Besides its accuracy, DNA also has the advantage in data acquisition. DNA can be found in blood, skin, saliva, hair and bones. However, the biology-based biometric recognition in general still suffers a lot from the following disadvantages: time-consuming (no real-time possibility), quite ex-

pensive, and requirement of human intervention. Therefore, it could be hardly utilized in many prospective applications other than law enforcement and medical applications. In this book, we are hence not going to continue discussing this category any further. If the reader is interested, more detailed information of biology-based biometric recognition can be found in reference [1] for dental radiograph and reference [2] for DNA profile respectively.

Face, voice, gait, and handwriting are another typical biological characteristics studied in cognitive neuroscience for recognition. In this group, all the three recognition procedures are processed purely by humans. Data acquiring is through eyes, ears and other human sensing organs. Data modeling and decision making are all processed in human brains through nerve cells. In Section 2.2, we will focus on how and why human can incredibly achieve the recognition procedure.

The third group is mathematical-based, or in another word, machine-based. This group requires machine sensing during data acquisition. For example, for fingerprint recognition, fingerprint sensors which could be either optical, semiconductor or ultrasound devices for capturing fingerprint images are required. Regarding face recognition, devices such as cameras, video cameras, or even IR (infrared) sensors are needed. Machine sensing can in principle utilize every biological characteristic. The most applicable modalities include physical as well as behavioral features. Physical ones are fingerprints, face, iris, voice, hand geometry, palm prints, palm vein (hand vein blood vessel patterns), retinal, etc. Common behavioral features are gait and signature. And many other characteristics are in various stages of development and research. Data modeling and decision making are processed by machines such as software programs, hardware chips, and even computers through mathematical models. Starting from Section 2.3, all of our discussion belongs to the machine-based recognition, although some fundamental ideas in cognitive science are applied.

The procedure and different categories are summarized in Table 2.1.

Table 2.1 Biometric procedures and categories

	Data acquiring	Data modeling	Decision making
Biology based	Physical sampled data: DNA, dental radiograph, etc.	Modeling according to biology science	Matching result is through biological comparison

Continued

	Data acquiring	Data modeling	Decision making
Cognitive based	Human being sensed data: face, voice, gait, handwriting, etc.	Modeling by human nerve cells	Underlying concepts are still in research
Mathematical based	Machine sensed: fingerprints, face, iris, voice, hand geometry, palm prints, palm vein, retinal, handwriting, gait, etc.	Modeling through mathematical methods	Similarity of the comparison is mathematically calculated and final decision is accordingly made

2.2 Cognitive-based Biometric Recognition

2.2.1 Introduction

For researchers in the field of machine-based biometrics, it is always believed to be beneficial to learn from how the human brain is correspondingly and wonderfully functioning. Neuroscience, physiology, psychology, anthropology, linguistics and artificial intelligence are the disciplines, each of which has its own and specific perspective of understanding minds and intelligence. Philosophical concepts, although crucial, seem too generic to answer our biometric-related questions. Therefore, we can only turn to an interdisciplinary science that embraces all the above-mentioned fields and can synthesize the valuable views and studies of how human brains work.

Cognitive science, which has been arising over several decades, is quite the appropriate discipline to research questions such as why humans could recognize the objects they have seen and furthermore why humans could easily memorize other's faces. The following scenario might occur quite frequently to most people: I must have seen the person before, but unfortunately I have completely forgotten his/her name. The consequent reaction might be the complaint of the poor memory. However, what might be overlooked is the

question of why face is memorized but the name is not. Is it an innate gift of humans to recognize faces or just developed through growing up? Does human have the same ability for face recognition and general pattern recognition such as for furniture, cars, animals, etc.? Those questions have been arising and researched even before the origin of the cognitive science, but even until nowadays most of the questions are still under debates.

In the following discussions, we will briefly introduce the research in this field, especially attempting to provide the inspirations for the study of face recognition from cognitive science.

2.2.2 History of Cognitive Science

The intellectual origins of cognitive science can be traced back to the 1950s. It is commonly considered as a milestone when the conceptual invention of computer machine by the famous British mathematician Alan Turing was published in 1950. Not only is he often considered as the father of modern computer science, but also quite a few cognitive scientists regard him as the founder of cognitive science. Shortly after Turing put forward the historic theory, the first "Universal Turing Machines" were implemented as digital computers. Scientists in different research fields then soon realized that computers in principle could be programmed to behave "intellectual" tasks. Those "intellectual" behaviors such as playing chess etc. were previously believed to be only manageable by humans. Why and how could it happen seemed quite attractive to many researchers in quite different fields, including neuroscience, physiology, psychology, anthropology, linguistics and even philosophy. Through discussions and cooperation, the pioneers were walking more and more close to each other until they decided to found a new research society, Cognitive Science Society in the mid-1970s. Since then, the research directions of this discipline have become clearer. Its major focus is to study the structures and functions of human and animal brains, including: sensation and perception—how the brains acquire knowledge; learning and memory—how the brains represent and further manipulate the acquired knowledge.

As one of the founders of the Cognitive Science Society, Donald A. Norman proposed the twelve issues that should comprise the study in cognitive in 1980 [3], including: belief systems, consciousness, development, emotion, in-

teraction, language, learning, memory, perception, performance, skill, and thought. Hot arguments could be found against his classification method. There exist issues such as performance and skill, which are sometimes not considered to be quite relevant to cognition. Some issues such as consciousness and thought do have overlaps. Some are even subsets of others etc. Anyhow, the famous twelve issues elaborate all the possible research directions and areas of cognitive. What's more, his thought is not only limited to the traditional natural science, but also includes the important philosophical analysis of the discipline. For instance, the "belief systems" could be more easily called "knowledge", but Norman prefers to emphasize all domains of knowledge, including culture knowledge, belief and world knowledge. From his way of depicting the cognitive science, we could be greatly edified for the following conclusion: we as human beings are always using everything we have learned, experienced and lessoned to solve problems. That is to say, besides the biological ability, humans are beautifully able to apply the combination of methods for problem-solving including recognition. If this fundamental way of human thought is also applied in machine-based recognition, we are approaching the essence of machine intelligence.

2.2.3 Human Brain Structure

To briefly explore and understand the common methods applied in cognitive science, we shall at first look through the construction of the human brains.

From the microcosmic point of view, a typical human brain consists of 100 to 200 billion neurons (nerve cells) and the neurons are connected with synapses [4]. If a brain is assumed as a machine, neurons can be analogous to hardware and synaptic connections to software in the brain. All the number of neurons are available with new-born, which can be only decreasing with age. But the synaptic connections are under more significant change. Although there are debates in neuroscience regarding the concrete number of synapses at a certain age, the trend of its change is commonly agreed, as summarized in [4] and [5]. The number of synapses at birth is ca 50 trillion (0.5×10^{14}), and is tremendously increased and peaks to 1000 trillion (10×10^{14}) during infancy. In the early age of childhood, the learning process is almost always activated because nearly everything is new to the baby. On the other side, since the

ability of cognition is in the infancy as well, the learning efficiency is too low, which produces a huge number of extra unnecessary synapses. As the baby is growing up, the redundant synapses are gradually and selectively eliminated. For example, the connections which are repeatedly used tend to be remained and those which are not so frequently used enough are removed. Thanks to this so-called "pruning" process, the typical number of synapses in an adult's brain is decreased to about 500 trillion (5×10^{14}). With such high-efficient and mature "software", the human brain's "hardware" can therefore run in high speed, in a high efficient way, and more importantly, in an unsupervised and automatic manner.

From the "pruning" process, we can obtain one of the fundamental hints on how machine learning, and more concretely, how machine vision should be designed in an automatical and unsupervised way.

Anatomically, a human brain can be divided into two halves, and each half is called cerebral hemisphere. The two cerebral hemispheres have different specific functions. The left hemisphere is typically associated with speech, writing, language and calculation. The right one is typically considered to deal with the spatial perception, visual recognition, and aspects of music perception and production. For the field of machine learning, the researchers normally pay more attention to the cerebral cortex or called neocortex, which is the outer and highly convoluted layer of the cerebral hemisphere. Neocortex is literally meaning "new cortex", because it is evolved later than other brain areas. The neocortex is found only in the brain of mammals and is especially large in higher primates, which is dealing with both conscious and non-conscious processes. Again, debates exist on how many neurons and synapses there are inside the cerebral cortex. G.M. Shepherd gives the number of 10 billion neurons and 60 trillion (0.6×10^{14}) synapses for an adult neocortex [6]. However, C. Koch lists the number of 20 billion neurons and 240 trillion (2.4×10^{14}) synapses [7]. Nevertheless, for us human beings, although neocortex contains only a small percent of the whole neurons, it is the predominant part of the brain regarding learning, which is particularly in charge of perception, emotion, thought and planning.

According to the traditional topological division introduced by Henry Gray in the early 20th century [8], the cerebral cortex has four parts: frontal lobe, parietal lobe, temporal lobe and occipital lobe, as shown in Fig. 2.1. The frontal lobe is associated with higher cognitive functions and initiates voluntary

movements. More specifically, this part is motor in function, organizing responses to events from the outside world. It also demonstrates planning behavior. Occipital cortex is located at the back of the brain, which is primarily responsible for vision-related functions. Temporal lobe is mainly dealing with hearing, some memory, and speech functions (speech production and speech comprehension). Parietal lobe contains somatosensory areas and sensory integration areas. In general, based on the functionality, the last three parts—occipital lobe, temporal lobe and parietal lobe can be generalized as the sensory cortex. All sensory information captured from outside world by corresponding organs (eyes for vision, nose for smell, ears for hearing, finger and other skins for touch, etc.) are mapped into internal representations in the sensory cortex. Those sensory representations are then distributed to frontal lobe, which carries out analyses and arranges actions.

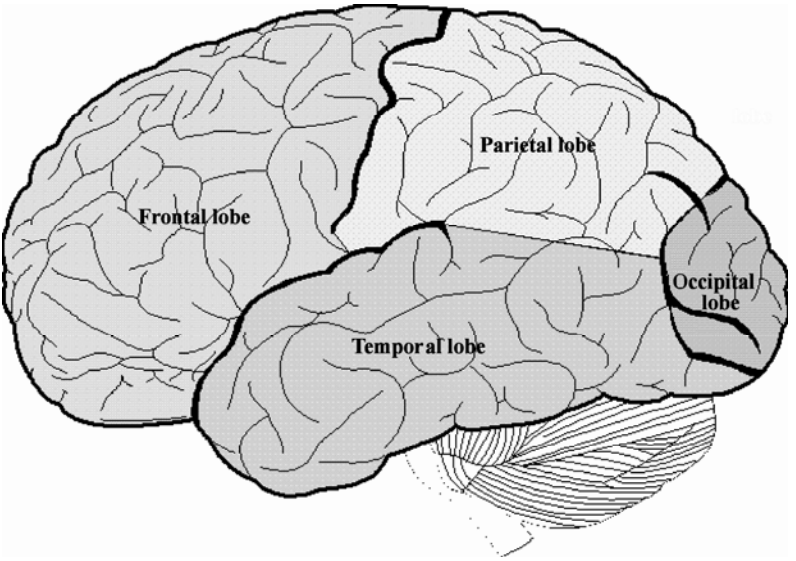


Fig. 2.1 Topological division of neocortex [8], with changes

2.2.4 Generic Methods in Cognitive Science

As mentioned in Section 2.2.1, cognitive science is interdisciplinary, which can benefit from both theoretical and experimental achievements from various fields. It is therefore under hot debates, and is difficult to summarize the generic research methods in cognitive science. Since the brief discussion of the cognitive science here is only aimed to inspire the research in machine learning and recognition, we would prefer a simple and clear description rather than different arguments in details. Therefore, through the exploration of those debates [9, 10, 11], we propose in this book an easy classification of the generic research methods in cognitive science, which includes: analysis, experimental methods, and computational modeling.

Analysis is originated from the psychological, anthropological and more importantly, the philosophical point of view, which deals with the fundamental and underlying issues. For example, examination of internal mental states (such as desire, motivation, belief, emotion, etc.), commence/difference of thoughts in different cultures, the study of rationalism (how brain is acquiring knowledge through logic reasoning) and empiricism (how brain is obtaining knowledge through experiences), etc. The logical analysis can be performed for functional constraints on cognition, providing the knowledge representation of cognitive systems.

The achievements from the analysis can be processed and observed through experiments on human subjects. Many disciplines can contribute their own different observations. In linguistics and physiology, behavioral experiments are made to measure the behavioral responses to stimulate. Likely, in cognitive neuroscience, controlled experiments are also performed to study the nature of the brain. The according experimental method is normally called brain imaging. Thanks to the advances in imaging techniques, there are many brain scan methods available to examine various mental tasks, including computer assisted tomography (CT), positron emission topography (PET), and functional magnetic resonance imaging (FMRI) methods, etc. For example, specific brain regions are identified for mental imagery through such brain scans. Even more microcosmically, the experiments with single-cell recording, neurons and synapses (as mentioned above) are also valuable for the development of cognitive science.

The third part is what we are more familiar with, the computational model-

ing. Although there are quite a few disagreements, many researchers consider that "Thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures" [10]. Through artificial intelligence, corresponding computational and mathematical models, which are mainly statistical models, are created and try to simulate the brain functions. We can certainly draw an analogy between these methods and those applied in machine learning. However, with such pure computational modeling methods, without considering the other two methods, especially the high-level analysis, the performance is far from success. We will discuss this issue with respect to the specific cognitive object recognition and face recognition topic in Section 2.2.6 and 2.2.7.

2.2.5 Visual Function in Human Brain

When people look at objects, millions of photoreceptor neurons and their 46 different brain areas are considered to be activated. Brain scans are normally applied to measure what is going on in a specific area of cortex. Starting from 1990s, through FMRI measurements for human brains, it is demonstrated that much of the posterior human brains contributes to vision. That is to say, the specific vision function is not only from occipital cortex. The human visual cortex is therefore defined, which includes the entire occipital lobe and extends significantly into the temporal and parietal lobes. The human visual cortex is accounting for about 20% of the whole cortex and therefore contains on the order of 5 billion neurons [12].

Most importantly, the visual cortex interprets different aspects of the visual image—such as its spectral composition, motion, orientation, and shape information. The orientation and shape information processing in the visual cortex, if applied for the recognition function, can be considered to an analogy of image-based recognition method in machine vision. A ventral stream of the visual cortex is usually thought of mainly dealing with such a processing procedure. Similarly, the motion information processing in the visual cortex, if applied for recognition, can be an analogy of temporal information and video context which could be used for the video-based or live recognition in machine vision. A dorsal stream is mainly involved in the analysis of motion information.

Readers who are interested in more details in this field are encouraged to refer to [12], where B.A. Wandell et al. reviewed the state-of-the-art research and measurements of human visual cortex, especially the visual field maps.

In conclusion, from the anatomical point of view, we can learn that the visual cortex provides the combined vision information, which is much more than just the still image-based information. Therefore, for the human brain, the function of vision-related processing, such as recognition, is robust. Similarly, for machine vision and recognition, combined methods should always be applied to achieve better performance.

2.2.6 General Cognitive-based Object Recognition

Although in debates, there are recent researches attempting to simulate what exactly happens during human object recognition, and applying the corresponding cognitive-motivated method for machine-based object recognition, as recently described in [13, 14, 15, 16].

Among the recognition approaches motivated by the cortex-like mechanisms, Serre et al. [13] is quite representative. Through a brief overview of this paper, we can get to learn the state-of-the-art research on cognitive inspired recognition systems.

According to the paper, the visual processing in cortex is mostly accepted to be hierarchical. Hence, the recognition tasks can be decomposed into different layers in the visual cortex. For the relatively simple processing stage, the neurons are gradually built to be invariant to easy 2D transformations such as position and scale. More complex processing tasks are implemented by the neurons which are tuned to deal with high-level transformations such as pose, illumination changes, etc. The tuning and the number-increasing of those neurons require long-term learning and visual experience. They therefore describe a hierarchical system that closely follows the organization of visual cortex. Their proposal builds an increasingly complex and invariant feature representation by alternating between a template matching and a maximum pooling operation. In the paper, they claim that the performance of their system is able to compare with many other state-of-the-art computer vision approaches.

However, there exist two major limitations in the typical cognitive-inspired methods:

- Lack of intelligence. As cognitive-inspired, the fundamental feature of the brain—intelligence shall never be overlooked. However, the recognition proposed in [13] is a semisupervised procedure. Moreover, the current research is focusing on the visual cortex and not taking enough consideration of the frontal lobe for recognition. Actually, high-level analysis is synthesized in the frontal lobe [12] which is therefore crucial for intelligence. For example, the important temporal information, the context correlations, and prior knowledge are cognitively integrated inside the frontal lobe to make plans and final decisions, which are finally contributed to recognition. But those discussions are not included in all the current research in [13,14,15,16].
- Computationally complex. For instance, Serre et al. [13] requires tens of seconds for processing, depending on the size of the input image. In contrast, many robust machine-based biometric recognition systems can run in real time.

It is a pity if we fail to see the wood for the trees. For the research in machine vision, in my opinion, it is important to not just only focus on the research details or debates such as how the visual cortex is organized into visual field maps, and how many maps exist for which kind of visual processing. The general inspiration from cognitive science should be paid the most attention to. The fundamental method is therefore proposed in [17] and will be extended and discussed in detail in this book.

2.2.7 Cognitive-based Face Recognition

Following the previous discussion in general cognitive-based object recognition, readers might ask the following questions: Will it be the same for face recognition in visual cortex and therefore for the cognitive-inspired face recognition method? Is face recognition more special than another kind of object recognition?

For the answers, we shall begin with the face recognition processing inside human cortex. Research shows that human newborns do demonstrate the underlying ability for face recognition. In [18], it reported that the newborn infants, even only nine minutes after birth, prefer to look at faces or face-like objects. But the expert discriminatory abilities which are applied for face

recognition are greatly improved through brain development and learning.

There are actually hot arguments on whether face recognition in the human brain is different than generic object recognition. The quite prevailing point of view is that there is a specific area in the brain that works only for face processing. As early as in 1992, Sergent et al. published their study in [19], which reported that region of the fusiform gyrus, which locates in the right hemisphere of the brain, is more active for viewing faces than viewing any other objects. Their claim was tested through the experimental method: positron emission topography (PET). Five years later, Nancy Kanwisher et al. even define the region as "fusiform face area" (FFA) in [20]. They applied the functional magnetic resonance imaging (fMRI) method to conclude that FFA responded much more significantly during face perception than other stimuli. FFA is responsible for color, identification of a face and even recognition of facial expressions. Damage to this area leads to a deficit function of recognizing faces. This argument is supported by different research groups such as the paper by J. V. Haxby et al. [21] and the study by E. H. Aylward et al. [22]. The more recent review in [23] further strengthens this opinion. Through evidences from behavioral, neuropsychological and neurophysiologic investigations, and through reviewing the related research, they conclude with the so called face-specificity hypothesis. That is to say, humans have specialized cognitive and neural mechanisms dedicated to the perception of faces, which is processed in the specific area of FFA.

On the other hand, there are several studies which draw totally different conclusions. One important opinion is that, the FFA is not only contributing to face processing, but with more generalized visual processing mechanism, as pointed out in [24]. The other sort of conclusion is that, besides FFA, there are also other cortex areas that respond to the face processing stimuli, as stated in [25, 26]. For example, it is agreed in [27], and even in [21] and [23] that, there exists a region called fSTS (face-selective region in the posterior part of the superior temporal gyrus). This region is apparently activated when the emotional expression changes, gaze occurs and viewpoint varies. There is another face selective region, named OFA, located in the lateral occipital cortex, which is sensitive to the physical aspects of the face stimuli [28]. Overall, FFA, OFA and fSTS are all primarily involved in distinguishing individual faces. Of course, this party of view is supported by convincing experimental results as well.

Although the two parties who hold opposite opinions provide us much information for the face recognition in cortex, further cognitive research is highly demanding for ending the debates and providing us a clearer answer. However, we, although as researchers in a different field, can now still figure out that, each side has unfortunately one limitation in common: the importance of frontal lobe is not taken into consideration at all. As mentioned earlier, the frontal lobe contributes to the high-level analysis such as reasoning, planning, and problem-solving, etc. Frontal lobe is performing the most complicated task, being expected to be involved in all brain process, and hence demonstrating the fundamental intelligence. This region should be definitely explored for the face recognition procedure. In early 1990s, Gross [29] suggests that the face processing cells are extended to the frontal lobe. In reality, this study focuses on finding the visual ability of the frontal lobe rather than the intelligence of it. More recently, Mechelli et al. [30] and Johnson et al. [31] found out that, the face processing task, although mainly performed in posterior cortical regions such as FFA, OFA and fSTS, is modulated by top-down signals originating in prefrontal cortex. The main purpose in [31] is to point out that, refreshing is a component of more complex modulatory operations such as working memory and mental imagery. And the refresh-related activity may thus be involved in the common activation patterns seen across different cognitive tasks. In summary, most researches are still concentrating on specific and different prospects. However, they convincingly support our fundamental opinion: the high-level intelligence performed in frontal lobe is crucial for face recognition.

It is important to note that, there is a high level research on cognitive-based face recognition, published by P. Sinha et al. [32]. They reported nineteen basic results from the face recognition by humans. Those high-level findings provide us further remarkable insights in designing the corresponding computer vision systems. In the following, we reorganize and analyze the nineteen results for the readers for a better understanding.

- Results for the image-based face recognition algorithms. The findings from Result 3 to Result 12 include the following information that contribute and/or influence the face recognition performance: high-frequency information, facial features, holistic processing, face aspect ratios (width and height dimensions), encoding, shape information, color cues, contrast variations, and illumination changes. Readers who are more interested to

the research in machine face recognition from still images are encouraged to refer the paper [32] for more details.

- Results that imply the intelligence of humans. The paper does not mention the human intelligence at all. However, through our above discussions in cognitive science, we should have already been able to keep this key point in mind to analyzing the findings. They point out that, humans can amazingly recognize familiar faces, even in poor conditions (Result 1, Result 2). The reason behind it is still an open question. However, it strongly implies the intelligence of human brain during face recognition. Since human brains contain much more information of a familiar face than an unfamiliar face, the successful recognition significantly from the ability of utilizing and combining all kinds of related information. Result 13 and Result 14 reveal that, the temporal information and the motion of faces are actually useful rather than harmful for recognition. These findings further strengthen our argument of brain intelligence. Brains can easily combine those temporal information, motion, video context and logic analysis etc. for face recognition.
- Other findings. Result 15 and Result 16 describe the developmental progress inside brains, and Result 17 prefers the argument of face processing in specific areas. Those findings have been so far discussed in this book. Result 18 suggests a feed-forward computation process in cortex. And result 19 claims that face recognition and facial expression recognition might be processed with separate systems. This finding is an important reference for researchers exploring algorithms in face recognition and expression recognition.

There is another side of human brain intelligent that is not well considered in paper [32]: the ability of human brain to learn in an unsupervised, fully automatic, and non-invasive way. To get to learn new faces or recognize known faces, human brains do not need any external supervisors for the training/enrollment. Everything that is required for recognition is obtained automatically inside brains. It will be ridiculous to assume the following scenario, as normally and necessarily occurs in machine-based recognition systems: Watch out! I am getting to learn your face, and I need your cooperation for me to remember your face! Would you please change your head pose with a better front view so that I can correctly enroll your face to my mind?

2.2.8 Inspirations from Cognitive-based Face Recognition

We have introduced the history, development and state-of-the art research of vision and recognition in cognitive science. The exact performance of vision, objects recognition, and face recognition in human brains are still far from clear. However, the advances and research evidences from this field provide us profound inspirations.

Through the human brain structure, either from the microcosmic point of view or from the anatomical point of view, we know that, the brain is still too complicated to be completely explored. Moreover, the human brain is pruning through learning and experiences. Therefore, an intelligent task such as recognition is not an intrinsic process. The brain has the "hardware" ability and is ready to learn foe recognition and improve its ability gradually and automatically.

The generic methods in cognitive science broaden our vision on the study in machine vision. Importantly, besides looking for perfect computational models, we shall always take high-level analysis into consideration, which could significantly contribute to what we are looking for—intelligent vision.

The research on visual functions, object recognition and face recognition further indicate that, although there are dedicated areas that are mainly crucial for vision and recognition, the success of recognition shall greatly benefit from the ability in brains to plan and execute goal-related behaviours.

All the above discussions bring us to the following questions: is there a way for machines to achieve as close as possible the fundamental intelligence of brain for recognition? Is a machine also able to combine temporal information, video context, logic analysis etc. for recognition? Can a machine system also run in an unsupervised, automatic and non-invasive way? We will answer the questions in Section 2.8, after exploring the fundamentals and state-of-the-art research in machine-based biometric recognition and face recognition.

2.3 Machine-based Biometric Recognition

2.3.1 Introduction

From now on, we will be mainly focusing on the mathematical-based (machine-based) biometric recognition. It deserves to be noted that for researchers in computer science and electronic engineering, there is a narrow definition of "biometrics", which is specifically referring to the mathematical model based biometrics. It is representative also due to its pervasive applications involving more and more people. In the following part of the book, we mainly point our discussions of biometrics to this definition.

2.3.2 Biometric Recognition Tasks

The purpose of biometric systems is for recognition, or in another word, authentication. In general, there are two major arguments of differentiating various tasks of the generic term "recognition".

The typical and prevailing opinion is the two-task category: verification and identification, as pointed out in book [34]. Verification is the task to examine whether a person's identity is the same as he or she claims to be. The examination is made through the comparison between the submitted biometric data and the corresponding data of the claimed person in previously enrolled database. It is actually a one-to-one comparison. Identification, however, is a one-to-many comparison. It is the task of looking for the identity of a person by comparing the submitted biometric data with all stored data in the database. There are two cases in identification. One is a close-set identification, in which the person to be examined is known to the database, the task is only to check who it is. In the other case, however, it is not clear whether or not the person to be examined is included in the database. This is called the open-set identification, or named watch list. The first step of the watch list task is the same as that of the close-set identification task, which has to compare the submitted data with the entire database. Then the similar-

ity scores of each comparison are sorted and the top matched one is selected. If the top match is above a predefined threshold, it is declared that the submitted person is inside the database and the identity is therefore provided. Otherwise, it is declared that a new person occurs. Finally, the biometric system determines whether to update the actual database with the new person or just to discard it. Obviously, the open-set identification is a more generalized case.

Another opinion is to divide the biometrics recognition tasks into three parts, derived from the famous FRVT evaluation protocols [79, 80, 81]: verification, identification and watch list, where identification corresponds to the above-mentioned close-set identification. This method arranges the tasks in the order of complexity. The easiest task is the one-to-one comparison, i.e. verification. And the middle one is the close-set one-to-many comparison, i.e. identification. The generalized and the most difficult one is the open-set one-to-many comparison, i.e. watch list. It deserves to point out that, the corresponding database in the watch list task can be called "watch list" as well.

2.3.3 Enrollment—a Special Biometric Procedure

For each biometric recognition task, the same procedure has to be applied: biometric data acquiring, data modeling and decision making, as mentioned earlier in 2.1. But this procedure actually refers to the matching part of recognition. There still exists a prerequisite of carrying out the biometric recognition: enrollment. The enrollment procedure is to apply the data acquiring and data modeling steps to register a new person and to construct the corresponding database. Biometric data of each human subject who is to be registered are captured through machine sensors. The captured data are then digitalized and transformed so that the mathematical features of the whole data can be extracted. The extracted features are labeled with the corresponding name and stored into the database. The constructed database can be denoted as template or template database. Compared to matching, enrollment (registration) features the following characteristics:

- User's cooperation as well as a human supervisor is normally required. The user (subject) has to be informed that it is in the registration procedure so that he/she can be prepared for biometric data acquiring. Moreover, it is

normally necessary to have a human supervisor. The supervisor helps the user to follow some instructions during the enrollment and informs him/her whether or not the registration is successful. For the matching procedure, however, it might be possible to passively recognize a subject and it might be possible that there is no human supervisor.

- The quality of enrollment is critical for recognition. There always exist variations between the biometric data submitted for recognition and its corresponding template (database). An optimized enrollment procedure can construct the database so that the best recognition performance can be achieved. Two major evaluation parameters are introduced: false acceptance rate (FAR) and false rejection rate (FRR). If a biometric system falsely identifies a person as someone else, it is denoted as the false acceptance. If a person could not be identified although his/her biometric data has already been enrolled in the database, it is denoted as the false rejection. Apparently, the template should be in principle so enrolled that the lowest FAR and FRR can be achieved for the matching procedure.

We will talk about the enrollment quality for face recognition in more detail in Section 2.7.

2.3.4 Biometric Methods Overview

Fingerprints, face, iris, voice, hand geometry, palm prints, palm vein, retinal, handwriting (signature), gait, etc., are the typical biometric techniques applied in industrial products. Many other biometric modalities [35,36, 37] such as ear, odor, brain electrical activity, nose, skin spectrum, nail, lips, etc., are in also vigorous research, some of which are even showing promising commercial potentials in the near future.

The revenue of biometric technology has been rapidly growing in the past decades, as stated in the most recent worldwide market research [38] for biometric techniques by International Biometric Group (IBG). The entire sale was only US$ 6.6 million in 1990, exploding to US$ 1.5 billion in 2005, and has been estimated to reach more than US$ 3 billion in 2007.

Fig. 2.2 shows the market share research result of different biometric modalities by IBG [38]. Some biometric techniques such as retina scan, signature recognition etc., are not directly presented in the chart. However, the IBG re-

search result provides a valuable overview of the current marketing share. Obviously, among the techniques, fingerprint (including fingerprint and AFIS/live scan) is the overwhelming leader, which accounts for approximately 70% of the whole market. Face recognition ranks the second and is still expected to play even more and more important roles in the future.

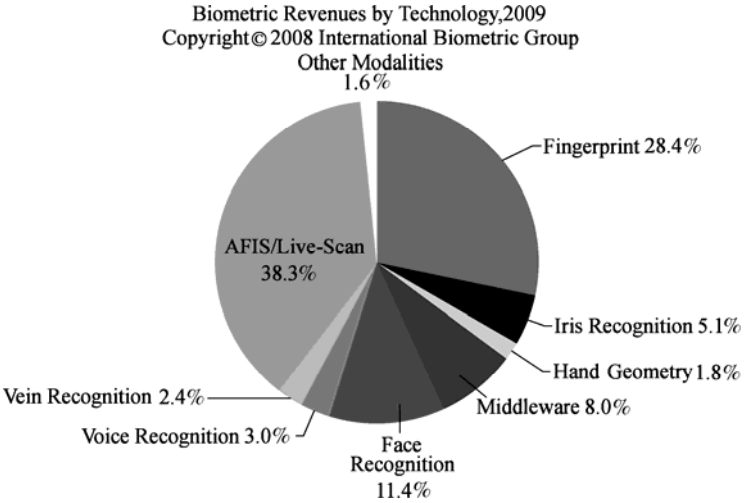


Fig. 2.2 Market share of biometric techniques [38]

In the rest part of this chapter, we are taking a brief overview of commonly applied biometric modalities. Table 2.2 shows such a comparison according to three important factors: robustness, usability and invasiveness.

Table 2.2 Comparison between different biometric recognition techniques

Biometric recognition methods	Robustness	Usability	Invasiveness
Fingerprint	++	++	+
Face	+	++	++
Iris	++	+	–
Voice	–	++	++
Hand geometry	–	++	+
Palm prints	++	–	+
Palm vein	+	–	++
Handwriting (Signature)	–	+	++
Retina	++	– –	– –

Among different biometric modalities, iris, retina, palm print and fingerprint are the most powerful ones. Those techniques are typically used for the systems with fairly huge databases while the false acceptance rate is the most important factor to be considered. All the other issues such as cost, usability, invasiveness, etc., can be relatively negligible. Legal applications, military systems, etc., are the major application areas. For instance, in UK, Project IRIS (Iris Recognition Immigration System) has been launched for checking the identity of at least one million frequent travelers from abroad to the UK. However, the critical challenge of those techniques is that they require substantial constraints on the user to be recognized. We can again take iris recognition for illustration. The human iris typically approximates to 1 cm in diameter while requires 200 pixels or more across the iris to be of "good quality", as proposed in the ISO/IEC 19794-6 standard [39]. Therefore, the acquisition of sufficient quality iris images is both distance and user cooperation dependent. The users have to be extremely close to the capturing system in both enrollment and matching procedures. Furthermore, a users' head has to keep still for quite a certain time until his/her identification is verified. Even a normal and tiny head movement can lead to the failure of the recognition system.

Fingerprint has proven from its success in biometric recognition market shares to be the most prevalent technology in use and the most mature modality. But there are still quite a number of people don't prefer to be fingerprint recorded and identified since it is conventionally applied in law enforcement and especially for recognizing suspects or criminals. Furthermore, due to its invasiveness, the application in consumer electronics is hardly prospective. Anyway, as the marketing leader, fingerprint recognition still deserves to be discussed in a little more detail, which follows in 2.3.5.

Voice is convenient of use and can be passively applied, but the vocal characteristic of a person is changing too much from time to time. Consequently, it is more commonly applied for speech recognition instead of biometric recognition.

2.3.5 Fingerprint Recognition

Fingerprint recognition, as one of the earliest applied biometric recognition method, can be even traced back to thousands of years ago. According to ar-

chaeological and historical records, in about 3000−2000 B.C., fingerprint patterns were used in Pottery in China; in about 2000−1000 B.C., in ancient Babylon, fingerprints were found on clay tablets for securing commercial transactions; since the 3rd century B.C., in Chinese Qin Dynasty, fingerprints were applied as seals for official documents and letters; at the end of the 3rd century B.C., shortly after Qin Dynasty, the law in Chinese Han Dynasty prescribed that criminals must sign their fingerprints for the confession. It was documented in Chinese Song History—Yuan Jiang Biography that, in around 10th century A.D. in Chinese Song Dynasty, Officer Yuan Jiang successfully applied fingerprints for identifying the criminal in a trial.

But it was not until the middle of 19th century A.D., scientific study of fingerprints recognition has started instead of empirical applications. The pioneers were the Scottish Dr. Henry Faulds [40] and the British Sir William Herschel [41]. The first scientific milestone of fingerprints recognition is comprehensively considered to be the publication of the book "Fingerprints" in 1892 by the British scientist Dr. Sir Francis Galton [42]. In the book, he firstly proved that no two fingerprints were exactly the same, and firstly detailed the classification system for fingerprints. Inspired from Galton's study, British Official Sir Edward Richard Henry developed his "Henry Fingerprints Classification System" in 1896, which includes 1024 primary prints types, which is still in wide use today. Since then, fingerprint recognition has started to come into widespread application in law enforcement all around the world.

In 1960s, the development of computerized fingerprint recognition algorithms started. In 1967, US Federal Bureau of Investigation (FBI) and US National Institute of Standards and Technology (NIST) collaborated in kicking off the FBI's Automated Fingerprint Identification System (AFIS). Although the term "automated" was included, AFIS was still a semi-automated system since traditional paper fingerprint cards were manually maintained and fingerprint human experts were highly demanded. AFIS at that time was a manual labor-intensive process and normally took weeks and months for a single check. Since 1999, FBI began to update AFIS to the Integrated Automated Fingerprint Identification System (IAFIS), which was fully automatic. FBI claims that it takes only within two hours for IAFIS to identify a criminal among over 41 million subjects in the criminal fingerprint database [43].

Besides US, similar governmental systems are deployed in other countries as well. Even the same acronym AFIS is used in countries such as Canada,

Germany, UK, etc. AFIS is also used to stand for "Automatic Fingerprint Identification Systems", which is mostly referred to the governmental fingerprint recognition system.

With the maturing of the technology, the cost of such systems has been rapidly going down so that they could be expanded into the civil applications as well. Consequently, many academic and industrial experts use AFIS as a more general appellation which is not restricted to law enforcement purposes. But there are still quite amount of researchers use the term AFIS referring to law enforcement systems and non-AFIS (or fingerprint) referring to other systems [38, 44]. The different meanings of AFIS and categorizing methods of fingerprint recognition systems might be sometimes confusable. In the following, we propose a new way of classifying fingerprint recognition systems mainly according to their applications (see Fig. 2.3).

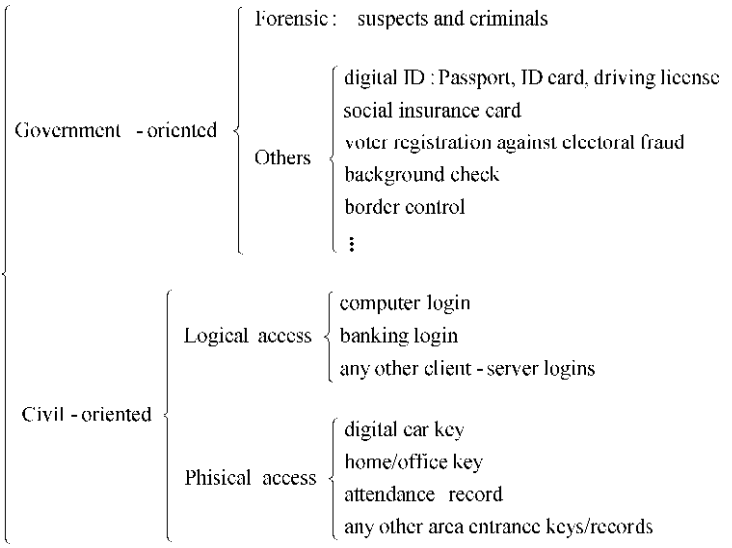


Fig. 2.3 Category of fingerprint recognition applications

As listed in Fig. 2.3, two groups can be briefly divided: government-oriented and civil-oriented. In Government applications, fingerprint database is always huge and the fingerprint template is supposed to be restorable into full fingerprint image. For civil usage, however, considering both personal privacy and fast processing demand, only encoded template extracted from fingerprint image is stored for future recognition. The civil-oriented finger-

print template is normally smaller than 1 KB, while the government-oriented one requires more that 200 KB. In summary, governmental systems can have the following characteristics that set them apart from civil ones:

- Databases are huge, currently in the order of from 100,000 up to million, and are still in growth.
- Fingerprint data in the database are fully fingerprint-image restorable.
- Long processing time, currently in the order of from hours up to days.
- Separate and specific storage and processing devices are required.

2.4 Generalized Face Recognition Procedure

Through our previous discussion on different biometric approaches, we are more interested in finding an identification system that ranks high for all the parameters. Face recognition is appealing because it achieves relatively high accuracy, potentially demands much less user cooperation and human supervision compared to iris, retina and fingerprint recognition. It is therefore less invasive and can be potentially cheap with fast processing speed. Although the advantages of face recognition over other biometric modalities especially over fingerprint are in principle quite apparent, they are unfortunately in practice not. That is primarily due to the technology bottleneck of state-of-the-art face recognition. On the other side, compared to fingerprint recognition, the disadvantage of face recognition is its robustness, which is both in principle and in reality less competitive. From the above-mentioned points, we can mainly explain why fingerprint recognition accounts for 4.5 times more than face recognition in the 2009 biometric market although face recognition is still the number two leader. If face recognition technology can really approach the underlying advantages, its booming growth is prospectively expected.

The success of utilizing those advantages is dependent on the whole generalized face recognition procedure, which contains the face detection, face recognition and face tracking as well. Therefore, it is important to firstly take a brief overview of such generalized face-recognition related research that is in the state-of-the-art, including face detection (see Section 2.5), face tracking (see Section 2.6) and face recognition (see Section 2.7).

2.5 Machine-based Face Detection

2.5.1 Face Detection Categories

As an expanding hot topic, numerous algorithms have been developed for face detection since decades ago. Its task is to search everywhere in an image or in a video stream, detecting whether or not there are faces existing and finding their locations. Face detection is the first and crucial step for any face processing system. According to its applications, face detection can be grouped into image-based and video-based. Image-based face detection methods are the basis. Its development provides different mathematical models for face detection. To detect faces in videos, temporal information is normally included for assistance. In the following, face detection methods are shortly discussed. More detailed surveys can be found in [45, 46, 47]. The categories of face detection methods are summarized in Fig. 2.4.

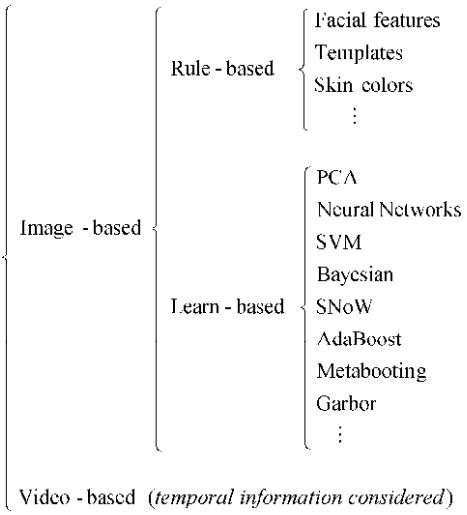


Fig. 2.4 Categories of face detection methods

In [45], face detection methods are divided into two groups: feature-based and image-based. Gong defines the image-based face detection approaches as two groups [48] (see Section 5.1 of the book): local feature-based and holistic-

based. But skin-color approach is not included according to the grouping method. Yang et al. classify the image-based face detection methods into four categories [46]: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. But those categories have large overlaps, especially for knowledge-based and template matching methods. We apply here a new classification rule so that the image-based face detection approaches fall into two groups: rule-based and learn-based, as shown in Fig. 2.4.

In a rule-based method, faces are detected based on some knowledge or predefined rules by experts. Facial features, skin color, and predefined templates all belong to this category. In facial feature models [49], eyes (eyebrows are often considered as secondary features), mouth, and nose (nostrils are often considered as secondary features), etc. are separately searched. The combination of each feature detector decides whether there is a face. This method is quite intuitive but often fails to detect faces with occlusion. Even when a small part of a face is occluded, one of the feature detectors may not work and therefore may result in face detection failure. Skin color, as a rule for face detection is not so obvious to normal people. We define different ethnic people as yellow, white or black, etc. in daily life, but research in [50] shows that the so called different "colors" mainly derives from the brightness rather than the chromatic color. The main problems of this method, however, are its sensitivity to camera parameter changes, sensitivity to varying lighting condition, and difficulties with skin-colored objects in background. For template-based models [51, 52], the rules are general descriptions of what a face looks like. In this method, the image to be examined is compared with the predefined face template. The main limitation of this approach is that it cannot effectively deal with scale, pose, and shape change of faces. Multiple resolutions, subtemplates and deformable templates are subsequently proposed as solutions.

Learn-based methods for face detection are greatly booming during the past several years to deal with changes in facial appearance, lighting, and poses. They generally rely on statistical analysis and machine learning theorem. Examples of face and non face images are normally required. Hence, this approach can also be termed as example-based. Principal component analysis (PCA) is proposed for both detection and recognition in [54]. Neural networks are successfully applied in [55, 56] with exciting results. Osuna et al.[57] proposes the support vector machines (SVM) algorithm to classify face and non-

face patterns. Bayesian approach is explored in [58] with wavelet representations. Sparse network of Winnows (SNoW) is used in [59] to specifically learn very large numbers of features. Recently, Viola et al. [60] applies AdaBoost learning algorithm for real-time frontal face detection and is extended to multi-view detection in [61]. A detector-pyramid architecture is proposed in [62] for fast multiview detection and the training is based on a metaboot learning algorithm.

Video-based face detectors can benefit a lot from the temporal information, which is mostly the motion. Frame difference and background subtraction are two main cues to detect motion. Some video-based systems even apply the motion information alone to detect faces. Raytchev et al. [117] subtracts the current image from the initial background to detect a face. This method, however, requires a stable background with no face. Turk et al. [63] and Palanive et al. [64] make an attempt at segmenting moving faces by intensity changes between neighboring frame images. When stereo cameras are available, stereo disparities are more precisely providing the motion information [53]. But motion detection alone may have difficulties with multiple moving faces, with faces in occlusion, or with relatively static faces. Thus, many robust video-based face processing systems [50, 53, 65, 66] combine temporal information with image-based face detection techniques for face detection.

There are several major terms that have been defined to indicate the performance of a face detection technique. As mentioned in [46], the face detection error can be grouped into two categories: false positives in which an image region is detected to contain a face but in reality does not; and false negatives in which existing faces are not detected. Face detection rate (FDR) is the percentage of correctly detected faces over all testing images. It can be mathematically represented by the following:

$$FDR = 1 - FPR - FNR \tag{2.1}$$

where *FPR* is the false positive rate and *FNR* is the false negative rate.

2.6 Machine-based Face Tracking

To process dynamic faces, the spatial information alone obtained from each frame is not helpful enough. People are almost always in motion and in variation. The frame-based face detection techniques are quite sensitive to the

changes of scales, poses and occlusions. They are not sufficiently robust to locate the faces. Temporal information from video context is therefore important. Tracking techniques are introduced by researchers to use the temporal context. Obviously, tracking is applied to follow a certain face in dynamic scenes and is supposed to compensate for the face motion effect. In general, hands, arms, body blobs and faces are the frequently used human parts to locate and follow a person. Each method has its own advantages and disadvantages and can be selected according to different applications. For example, body blobs are normally used in a security system to analyze the people's behavior. But face tracking attracts the most attention especially in face recognition systems. We briefly survey the popular face tracking algorithms here.

Face tracking is actually the task of prediction and update. It can be normally defined as a statistical process, in which the tracking goal is to estimate the actual state of a face from a sequence of observations in the order of temporal changes. At each video frame, a vector of observations may include scales, colors, and positions, etc. of a certain face in the image. Markov processes are often applied to combine prior information and observations for prediction, which are mathematically represented by propagation of probabilities of detection and the detection parameters over time. Hidden Markov Models (HMM), *Con*ditional *Dens*ity propag*ation* (Condensation) and Kalman filters are the three major filtering techniques. But they are relatively computational expensive for the application in consumer electronics.

Several categories of the observations can be divided for face tracking. Some systems are using the facial features such as eyes, eyebrows, nose, nostrils, and lips, etc. as tracking targets. But they might meet errors when those features are not visible due to pose changes or partial occlusions. Color-based approaches are more advantageous in terms of speed and successful systems, which are claimed in [67,68,69]. But the main problems are their weakness in illumination changes. The combination of different approaches is proposed with improved results. Birchfield [70] fuses the ellipse contour of faces and the color characteristics together to track a face in video. It declares to be robust against out-of-plane head rotation and partial occlusions. But only one face is assumed and a manual initialization is required.

More recently, Verma et al. [71] put forward a system that simultaneously combines the face detection and tracking information. It accumulates probabilities of detection over a sequence which leads to coherent detection over

time with improved detection results. It also predicts the positions, scales, and poses of detection to guarantee the accuracy of accumulation as well as a continuous detection. But the tracking is based on two face detectors: one for frontal face and one for profile. If both detectors fail for several consecutive frames, the tracking eventually fails. Therefore, it cannot improve the detection very well. Another weakness is their difficulties in dealing with crossing-over faces. In their experiments, when one face is gradually occluded by another face, the tracking is lost. When both faces are visible again, however, they are tracked as new faces. It should be noted that most tracking algorithms would have the same problems.

2.7 Machine-based Face Recognition

2.7.1 Overview

Different from face detection, which finds the generic characteristics of human faces that are differentiated from any other objects, face recognition is to find the unique characteristics of every single face to be differentiated from any other face. It is therefore a much more complex task. Therefore, in principle, to achieve better performance, it should be based on machine learning algorithms rather than rule-based methods.

A really booming research has not started until a milestone in machine recognition of faces in [63] in 1991, where an eigenface-based system of detecting and recognizing faces was demonstrated. A large number of algorithms have been developed since then. Chellappa et al. [72] in 1995 and Zhao et al. [73] in 2002 are two important surveys of the general machine-based face recognition approaches. And in [74], the most recent 3D-based face recognition methods are reviewed. Much of our discussion of the image-based face recognition is stimulated by the three papers.

Additionally and more importantly, we are more interested in video-based face recognition and the whole procedure of recognition that are unsupervised and fully automatic running, just as we have learned from the cognitive sci-

ence. The background research of current video-based face recognition and unsupervised face recognition are hence more elaborated in Section 2.7.5 and 2.7.6. The survey on both scientific papers and patent publications is given, which includes both the hot research points and directions in academic as well as industrial groups. The common limitations in current video-based and unsupervised face recognition research are then summarized and discussed, which explains why the limitations greatly hinder the real application of face recognition in many areas, e.g. in consumer electronics.

2.7.2 Benchmark Studies of Face Recognition

As the starting point of our discussion, we introduce in the following the famous benchmark studies of the face recognition technology, from which we can get an overview of the latest progress of the technologies, their most critical challenges, and the demanding research points.

The public available FERET (Face Recognition Technology) database and its protocols [77, 78] provide large databases and standard evaluation methods for assessing different face recognition techniques. There had been three FERET evaluation tests from 1994 to 1997: first Aug1994, second Mar1995, and third Sep1996 (administered in Sep. 1996 and Mar. 1997). The tests were sponsored by U.S. government and appealed many university research groups. The tests did not only provide the possibility to objectively evaluate different face recognition algorithms under almost real-world situations, but also clearly proposed the possible future research directions. There was a gap between 1997 and 2000 and the program has turned its interests to commercial products since 2000. Up to now, there have been three FRVTs (Facial Recognition Vendor Test): FRVT 2000 [79], FRVT 2002 [80] and FRVT 2006 [81], all of which are mainly based on the same FERET evaluation protocols with updated databases. Those tests are useful for customers to decide whether to choose a face as a recognition method and decide which product is better for a certain application. Although not ideal, researchers can still learn from the lessons of the current products. In the following two paragraphs, we list the report results derived from FRVT 2000 and FRVT 2002 to demonstrate the state-of-the-art commercial face recognition techniques.

FRVT 2000 result was published in Feb. 2001 [79]. It identified temporal

and pose variations as two key areas for future research in face recognition. The FRVT 2000 shows that progress has been made in temporal changes, but developing algorithms that can handle temporal variations is still a necessary research area. In addition, developing algorithms that can compensate for pose variations, illumination and distance changes were noted as other areas for future research. The FRVT 2000 experiments on compression confirm the findings of Moon and Phillips that moderate levels of compression do not adversely affect performance. The resolution experiments find that moderately decreasing the resolution can slightly improve performance. In most cases, compression and reducing resolution are lowpass filters. Both results suggest that lowpass filtering probes could increase performance.

FRVT 2002 report [80] is published in March 2003. It finds out that indoor face recognition performance has substantially improved since FRVT 2000 but outdoor recognition performance needs improvement. Other detailed conclusions are:

- Face recognition performance decreases approximately linearly with elapsed time database and new images.
- Top face recognition systems do not appear to be sensitive to normal indoor lighting changes.
- Three-dimensional morphable models [82, 83] substantially improve the ability to recognize non-frontal faces.
- Recognition from video sequences was not better than from still images.
- Males are easier to recognize than females.
- Younger people are harder to recognize than older people.
- For identification and watch list tasks, performance decreases linearly in the logarithm of the database or watch list size.

The latest test is FRVT 2006, which was reported in March 2007 [81]. We summarize below their conclusions:

- Performance of the still-image based face recognition technology has significantly improved, especially with varying illumination across images.
- Performance of the 3D based face recognition technology has significantly improved.
- By comparing human and machine performance, the best-performing machine algorithms are more robust than humans. It was the first time in history that the machine-based faces recognition algorithms outperform humans. However, the comparisons were only made under varying

illuminations.

Regarding the above latest report, we could achieve the following comments and inspirations:

- The evaluation of FRVT 2006 is only limited to the still-image based algorithms, while video-based approaches are not discussed. Therefore, the video-based benefits for recognition such as temporal information, video contexts are not taken into consideration.
- The evaluation of the algorithms dealing with pose variations is not included for the FRVT 2006 test.
- Rather than the claim of better machine algorithms than humans, it is more accurate to only conclude that, compared with humans, machine face recognition algorithms can sometimes deal better with illumination changes. For their human-machine comparison, there are some critical prerequisites. First of all, the tests are only focusing on varying illuminations without including other influences like pose variations, video contexts, etc. As we have mentioned earlier, human has no problem with recognizing faces with online out-of-plane head rotation. But no state-of-the-art algorithms can handle that. Moreover, even under varying conditions, only "non-familiar faces" are used for human subjects. The reason is that, humans have the excellent ability to recognize familiar faces in very poor lighting conditions [32, 81]. It implies in the report that, machine algorithms are still not be able to compete with human for recognizing familiar faces.
- The fundamental intelligence of face recognition by humans is not included in the report because it might be out of their purpose.
- The above four demanding research points derived and inspired from the FRVT reports are just what we would like to mainly discuss in this book.

2.7.3 Some General Terms Used in Face Recognition

There are also some general terms that are often used in face recognition. Face recognition can be further divided into two groups according to their applications: verification and identification. Identification is comparing a new unknown face image with a set of face image groups or a set of templates in the face database to find out who is the person. Verification is comparing a face image, which is claimed to be a certain person, with the corresponding group

or template in the face database and checking whether it is true.

Similar to face detection, two categories of recognition errors are used to evaluate the performance of a method. The false acceptance rate (FAR), sometimes referred to as the false alarm rate, corresponds to false positive rate in face detection. It is the percentage of face shots which have been wrongly identified as somebody else. There are two cases for FAR. One is that an unknown face is wrongly identified as known. The other one is that a certain known face is wrongly identified as another known face.

A false rejection rate (FRR) is corresponding to a false negative rate (FNR) in face detection, that is the percentage of face shots which are detected as unknown but actually they are enrolled (known) faces.

The false acceptance rate and false rejection rate are actually correlated. There is a tradeoff between them. The requirement to receive a lower false acceptance rate can also lead to a higher false rejection rate and vice versa. A plot of numerous false acceptance rate-false rejection rate combinations versus similarity threshold is called FAR/FRR curve. The illustration of the curve can be referred to Fig. 4.3. The curve shows the natural compromise between FAR/FRR of a certain recognition technique. From this curve, we can easily choose a threshold to achieve either a lower FAR or a lower FRR according to different application purposes. For example, a security system to identify a criminal is preferably to keep an extremely low FRR but an automatic banking system would rather take a higher FRR to achieve an extremely low FAR.

Similar to equation (2.1), the recognition rate *FR* can be represented as the following:

$$FR = 1 - FAR - FRR \tag{2.2}$$

2.7.4 Recognition Procedures and Methods

[72] separates the task of face recognition into segmentation (face segmented by detection), feature extraction from face regions, identification and matching. Without considering the face detection, we define here the face recognition as three procedures: enrollment, matching/classification and update. Fig. 2.5 summarizes the face recognition procedures and methods.

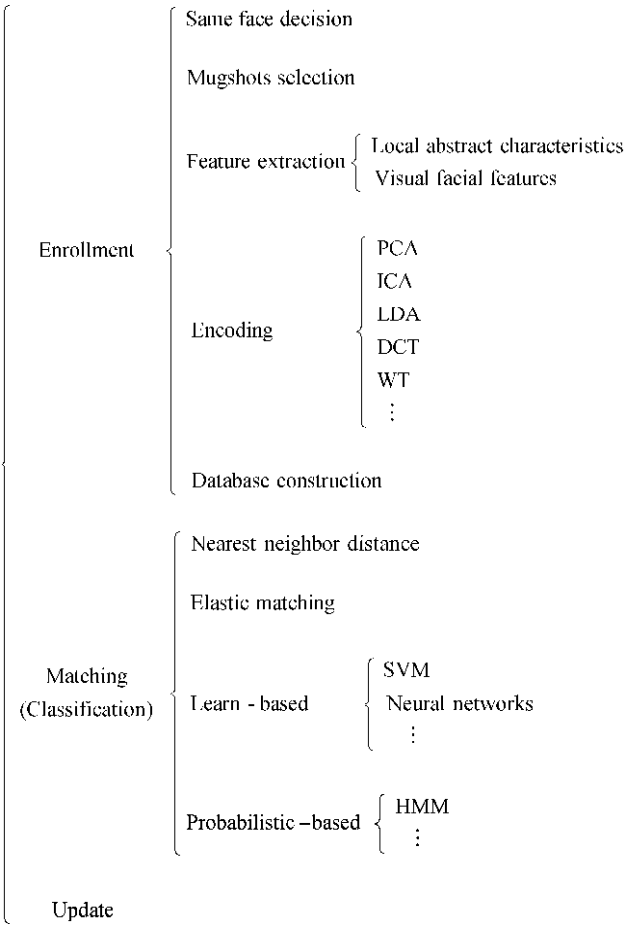


Fig. 2.5 Category of face recognition procedures and methods

The key to successful face recognition is a high quality enrollment. In this step, one or more images of one person's face are grouped and often encoded as one template. Many templates (groups) construct a face database. There are some parameters which represent a high quality enrollment. We divide this procedure into five substeps: same face decision, mugshots selection, feature extraction, encoding and database construction. That is to say, a person's database is expected to be robust enough to recognize him/her afterwards.

Same face decision is the crucial step although most people overlook its importance. For a machine recognition system, it has to make sure that the selected face shots for an expected person are really from the person and not

from anyone else. Since enrollment is traditionally considered as a pre-training phase for any kind of learning method, the task of the same face decision here is normally implemented by a human supervisor. Consequently, it is really a challenging problem for a machine system to manage it by itself.

Once a target face is selected, the second step is to select qualified face shots of it from all available images or data. The selection is more or less dependent on the encoding and classification methods. Some technology might emphasize the variety of head postures; other technology might require the variety of facial expressions. Nevertheless, the purpose of the selection is to ideally make sure that the faces that have been pre-trained should be identified when they show up again. This step is also normally handled with the help of human beings instead of machines themselves. The human assistance in the first two steps appears to be inevitable and has been accepted for many years until quite recently. How can a machine recognition system know that who is who without any pre-trained database? But the amazing thing is human beings have this ability. Therefore, there must be ways for machine-learning systems to achieve the tasks. We will address the related research of these unsupervised face recognition methods in Section 2.7.6.

Feature extraction and encoding are the third step and fourth step respectively in the enrollment, on which most typical research efforts are concentrated. The traditional term “features” indicates facial features such as nose, eyes, mouth, nostrils, chins and forehead. The feature-based method is normally classified as the antonym of the holistic method and is a popular approach for face recognition in earliest research. And the holistic approach plays a more important role for a machine recognition system. It mainly emphasizes the global characteristic and description of faces without considering any facial features. However, biological study [84,85] of human perception systems show that both holistic and feature information are crucial for the perception and recognition of faces [73]. Most successful recognition systems use the hybrid of both. Therefore, we include the feature extraction as the necessary step for enrollment. A more general definition of the term “features” is preferred, which can be either local abstract characteristics in the face region such as lines, curves, edges, fiducial points or areas, or visually facial features such as eyes, mouth, nose, etc..

Encoding is required to decrease the dimension of a data space of face regions. A greyscale image with a resolution of $m \times n$ pixels is normally repre-

sented by a $m \times n$ matrix with each element corresponding to the intensity value of each pixel. The matrix is also called an image space with $m \times n$ dimensions, which needs too many computation efforts to make comparisons and too much storage to save them. Efficient encoding approaches have to be found to greatly reduce the data without losing much information. Eigenfaces [63], which is based on principal component analysis (PCA), independent component analysis (ICA) [86], linear discriminant analysis (LDA, known as Fisher discriminant analysis) [87, 88], discrete cosine transformation (DCT) [89, 90, 91] and wavelet transformation (WT) [95, 96, 97] are the major well-known encoding methods.

The main idea of PCA is trying to project a $m \times n$ dimensional space onto a much smaller subspace, which is a linear combination of orthogonal (uncorrelated) vectors. Those vectors are named as principal components or eigenvectors, which are ranked by the associated eigenvalues of the covariant matrix of the whole image space data. Since each eigenvector has the same dimension of the original face image and is also "face-like" in appearance, it is defined as "eigenface" [63]. PCA is aimed to reduce the redundancy of datasets while remains a minimum loss of information. Thus, the least mean square reconstruction error is always concerned in PCA.

ICA is a generalization of PCA. While the goal of PCA is to achieve the minimum mean-squared reprojection error by minimizing the linear dependencies of input data, the goal of ICA is to minimize both second-order and higher-order dependencies in the input. The normal procedure is to firstly decorrelate the input data by PCA, and then reduce the remaining higher-order dependencies by ICA. Therefore, unlike PCA, the basis vector of ICA is neither orthogonal nor ranked in order. However, there are contradictory arguments on whether PCA or ICA is the better for face recognition, Draper et al. [98] compares both methods in more details.

LDA is trying to represent the input data with the vectors that best discriminate different classes. Those vectors are computed to maximize the inter-class variance and minimize the intra-class variance. From the mathematical point of view, it might be generally believed that the LDA method outperforms PCA when applied into face recognition. The former reconstructs a face by classes with maximum separability while the latter does not pay any particular attention to the underlying class structure. But the experiments in [99] show that it is not always the case. The superiority of PCA over LDA occurs

when the number of samples per class is small.

All traditional PCA, ICA and LDA are linear encoding methods. But the image data are in general nonlinear. Therefore, nonlinear (kernel-based) approaches are introduced quite recently: Kernel-PCA (KPCA), Kernel-ICA (KICA) and kernel-LDA (KLDA/KDA) [100, 101]. There are few papers that make comparison among all PCA/KPCA, ICA/KICA and LDA/KLDA approaches. For example, though independent experiments, Zhang et al. [102] and Draper et al. [98] have drawn contradictory conclusions. Further explorations are required to tell which one in which case is better. But they are beyond the scope of this book and we are just putting forward this question as a research challenge for readers.

DCT is a popular technique in image and video compression especially with the widely used JPEG format. It divides an image into subblocks and each block is decomposed by 2D DCT basis functions. Although it is not as optimum in terms of energy compaction as PCA, ICA and LDA approaches, DCT is still very attractive for the application of face recognition [89, 90, 91] thanks to its computational efficiency and ease of implementation both in hardware and software.

Wavelet transformation is well known for providing a multiresolution analysis of an image. Its computational efficiency also makes it possible for a real-time application. With the extensive application of JPEG2000 (based on wavelet transformation) for image and video compression, wavelet encoding has more and more appeals.

Since pose variation is one of the most critical problems in face recognition, there are accordingly methods suggested to deal with the challenge, including active appearance models [92], multiclassifier fusion approach [93], etc. More recently the face mosaicing model is discussed in [94], where an image fusion scheme is proposed to generate a 2D face mosaic of a person during enrollment. Mosaic applies both the frontal and side-profile face images to generate an extended 2D image for the further matching procedure.

To significantly improve the face recognition performance, especially to be robust against pose variation and illumination changes, 3D-based face models are proposed, as discussed in [74, 75, 76, 82, 83]. 3D models provide both the shape and texture information, which are proven to be significantly superior to the above-mentioned 2D method, as pointed out in the FRVT 2006 test [81]. The 3D model can be either created from original 2D image data or

directly acquired from 3D data. The major limitation of the first 2D-image based enrollment method is the complexity of training process. A large number of training images with various conditions are typically required and the user's cooperation is necessary as well. Compared to the first approach, the real 3D models can be trained in a relatively easier way, while requiring expensive 3D sensors. Although with the high performance, the disadvantages of both 3D encoding methods restrict their pervasive applications.

The final step of enrollment is to arrange the encoded face data into a certain structure, which is known as database construction. An efficient database structure is important to achieve high recognition accuracy.

After enrollment, the second main step is to make classifications, which can be also defined as a matching problem. Matching is to examine the similarity of a subject image with the existing databases. Any kind of classification method from pattern recognition theory can be applied. The popular approaches are nearest neighbor distance, elastic graphic matching, learn-based method and probabilistic approach.

Nearest neighbor distance is the most apparent and simple method. Euclidian distance is applied together with eigenfaces in [63].

Elastic matching approach together with wavelet transformation is proposed in [96]. It is a more complex method which runs heuristically to find the image graph which maximizes the graph similarity function.

Support vector machines (SVM) [103, 104, 105], and neural networks [106, 107], are two major machine learning-based methods, which outperform the Euclidean distance approach but with sacrifice of computation efforts.

As a probabilistic method, Hidden Markov Models (HMM) [89, 90], are successfully applied as face recognition classifiers and [95] claimed that the combination of wavelet encoding together with HMM can outperform other methods. A more complete study of the comparison among the classification approaches for face recognition is still on demand.

Update, as the last procedure for face recognition, is also not a very hot topic of research. Traditionally, there are two approaches for dealing with update. One way is trying to represent a face with complete enough pre-training mugshots and does not require any update. However, as stated in the FVRT 2002 test [80], face recognition performance decreases approximately linearly with elapsed time database and new images. This strongly addresses the importance of update. The other way is to require an additional human supervisor,

who makes the decision on when and which mugshot is to be updated. Apparently, such a tradional way of update could not be non-invasive and fully automatic.

2.7.5 Video-based Recognition

As categorized in Section 2.5.1, there are two main groups of research for face recognition: image-based and video-based. Recognition from video attracts more and more research interests in recent years with the availability of cheap video cameras and fast computers for video processing. It has more application prospects than the image-based research since most practical systems acquire mugshots from video capturing sources. Moreover, in the real world, faces are almost always moving. Recognition through video sequences can find more underlying biological supports. In the previous sections, we have discussed the general procedures and approaches which are all suitable for the image-based methods. But the video-based recognition has a lot of additional features which have not been covered in our preceding discussion. There are a few basic differences between dynamic and static image analysis. In this section, we are mainly talking about the differences between them and are giving an overview of the state-of-the-art of research in video-based approaches.

Compared to static-image analysis, video-based recognition might have the following difficulties:

- **Motion**. Moving faces produce too many changes to make identification. The movement includes local facial movement and global face movement. Faces from different persons with the same facial expression might have a bigger similarity than faces from the same person but with variant facial expressions. Global face movement, especially the pose changes are even more harmful [80].
- **Complex background**. There might be a lot of unpredictable moving objects or people in the video. Those background moving objects/people can occlude the face of interest and hinder the tracking of faces. Another issue is the variation of lighting conditions. A sudden illumination change often occurs with turning on/off the light. Additionally, shadows can mask many valuable facial characteristics.
- **Resolution**. To save storage as well as bandwidth, and to be suitable for

the real-time processing, the available video data are often with low resolution, normally much lower than an ideal face shot in static images. Hence, it obtains less information from a certain video frame.

Those disadvantages slow down the application of face recognition in security systems. A typical example is a criminal-alarm system at the airport. A commercial face recognition system was tested at the Palm Beach International Airport and Boston's Logan Airport in USA, 2002. Although it claimed to be one of the best systems in the market, both airports showed their absolute disappointment with the performance.

However, there are also some benefits available from video:

- **Temporal information**. Video provides temporally correlated image sequences, which make it advantageous over image-based methods. Temporal information is crucial for tracking faces and consequently makes it easier for the same face decision procedure.
- **Abundant data**. More than enough frames for a certain face are usually available for recognition. Recognition errors in part of the frames can still be compensated from those successful frames. It allows the system to discard bad quality frames and select only qualified frames for enrollment and matching.
- **Environment**. Although moving objects/people are a big challenge, videos are often recorded in a constrained area. With a location-fixed video camera, there are also static objects that are not changing frequently. The background subtraction algorithms can be applied as a secondary approach for face detection as a preprocessing step to determine the regions of interests. For that reason, it is for some applications quite useful, e.g. a video surveillance system, to watch whether there is someone in a restricted area and who is showing up.

In earlier research such as [63], [96], although video sequences were used as inputs, the still-image face recognition techniques are applied without considering any above mentioned video advantages.

More real video-based approaches use both spatial information (in each frame) and temporal information (relations between the frames). Those published research before 2002 has been surveyed in [73]. We explore the more recent research here.

Liu and Chen [108] proposes adaptive Hidden Markov Models inspired by speech recognition. It requires a supervised enrollment procedure and an un-

supervised matching procedure. During training, each frame (only containing face portion) is considered as an observation. The statistics and temporal dynamics are learned by HMM. During the matching process, the likelihood scores provided by the HMMs are compared to find the identity of a certain test video sequence. Moreover, each HMM is also adapted with the test sequences to obtain better modelling over time. The performance concludes that it is better than using majority voting of image-base recognition results. But they use cropped images which contain only faces in their experiments, which is a too strict requirement.

Krueger and Zhou [109] presents a general framework which provides a complete statistic description of human identities. Their proposed subspace identity encoding is claimed to be invariant to location, illumination and pose changes. The performances are expected to be degraded with face occlusions.

A comparative study based on experimental analysis between face recognition in static images and videos is demonstrated in [110]. PCA and LDA are chosen as pure image-based methods while HMM [108] is as video-based method. Two major conclusions have been drawn. Firstly, thanks to the combination of spatial and temporal information, it makes the video-based method much less sensitive to bad image quality than the image-based approach. Secondly, the joint spatial-temporal representation is more sensitive to sequence length than the static method. That is to say, short video may lead to worse performance with HMM method, which cannot efficiently extract the dynamic information from too small number of frames.

Tang and Li [111] proposes a multi-classifier approach for video-based face recognition. They first use audio information to align frames in different videos in the order of image similarity. After that, majority voting and a sum rule are applied to combine the unified subspace analysis classifiers from each frame. Their experiments show perfect results, but the subjects are asked to speak predefined sentences to contribute to the audio alignment.

Arandjelovic and Cipolla [112] takes advantages of face motion manifold from video. The Resistor-Average distance (RAD) is computed on nonlinearly mapped data using Kernel PCA. RAD is then used as a dissimilarity measure between distributions of face images. Sources of enrollment errors are modeled and incorporated in the recognition framework to improve the recognition. Videos with random face motion are tested with successful results. But their methods do not handle the well-known illumination problems.

It is important to mention that although the above discussed papers propose different valuable ideas and algorithms for video-based face recognition, all of the performances are tested with highly constrained videos. The most critical one includes: every sequence consists of only one person at one time, moving background objects are normally not considered. Therefore, they are still far from the complex environments in real world.

2.7.6 Unsupervised and Fully Automatic Approaches

The so far discussed face recognition technology requires a separate enrollment (training) step, where face images are collected, selected and labeled manually. Many face recognition approaches, systems, and applications declare to be fully automatic. However, most of them are actually only automatic for matching, not automatic for the whole recognition procedure. A separate training or enrollment procedure is normally necessary, human supervision is therefore always required.

For an online training, where databases are constructed through live video, moreover, the difficulty and importance of update has not been sufficiently addressed by many researchers. There are two conventional ways to construct face databases from videos. One approach is to have a pre-training procedure before recognition, in which the face images from a certain person are carefully selected by a human supervisor. Those face shots are then encoded in a certain format and stored into the corresponding database. The selection criteria are dependent on the recognition methods. Most robust recognition systems collect various face shots of a certain person under varying lighting conditions, multiple views, different head poses and expressions, etc. Although those systems perform well, the training procedure itself demands significant efforts for a human supervisor to select qualified images.

The other way for face database construction requires little effort from human supervision, but the co-operation of subjects is necessary. This approach can be annoying to the subjects. A supervisor normally asks them to change their poses and expressions from time to time during the training phase.

A completely automated system with no help from outside assistance and no requirement for users is hence on demand. The features of such an intelligent system can be described as the following. The system starts with no data-

base at all. At the beginning, every person is a stranger whose face shots are considered to be unknown faces. The system should be able to automatically select qualified face shots of each unknown person, enrolling them into a self-constructed database, which is expected to support recognizing the person when he/she appears again. Moreover, it must not erroneously combine different persons into one database. For known faces which have a corresponding match in existing databases, update is required to keep up with the recent views of a certain person. It is a big challenge to perform the task without any human assistance. Recently, there are advances in automatic and unsupervised face recognition. [113, 114, 115, 116, 117, 119, 121] are the interesting papers from state of the art for further reading.

[63] might be the first article which studies the unsupervised recognition of faces. They use eigenfaces to classify a face image. A new face can be automatically added if its face image has a big enough Euclidean distance when compared with existing databases. However, no temporal information has been used and they have not considered the update procedure.

Wechsel et al. [119] claims an automatic person verification system. It includes video break, face detection and authentication modules. In the video break, motion information is firstly used to set the region of interests (ROI) as face candidates. A simple optical flow method is used when the image SNR level is low. When the SNR level is high, simple pair wise frame differencing is used to detect the moving object. In the face detection module, coarse-to-fine face location methods are used. A decision tree works as a precise approach to determine the face regions. The face images are then used for authentication using a radial basis function (RBF) network. The matching task (actually only the verification task here) is accomplished by comparing with existing face databases. This system is tested on three image sequences: the first is taken indoors with one subject present; the second is taken outdoors with two subjects; the third is taken outdoors with one subject under stormy conditions. Perfect results are reported on all three sequences. It is one of the few automatic systems which are able to deal with multiple people situations. However, the system is in general not a fully automatic system. It achieves only a verification task with a pre-constructed database, which is not expected to automatically enroll new faces. Another limitation is that they have only tested with frontal or nearly frontal faces, which is still a too constrained requirement.

A real autonomous face recognition system was firstly introduced in [113]. They propose an unsupervised learning approximator—SHOSLIF, to incrementally create a tree from training samples. PCA and LDA are used recursively to build a feature space for every internal node (also called a leaf node) of the tree. For any given input, a certain number of top matched leaf nodes are reported. Then a spatial-temporal cluster method is applied to group primitive clusters out of leaf nodes. Eight people are tested which shows perfect performance—100% recognition rate has been achieved. However, their experiments have the following requirements: one person exists in one sequence; each subject always tries to place his/her face in the center of an image, which means that each frame is actually a face image; only frontal views are taken for testing the performance.

[114] is an extension of [113]. An incremental hierarchical discriminating regression (IHDR) decision tree is constructed to deal with the classification problem. The IHDR algorithm applies the discriminant analysis rather than PCA to split data. Virtual labels are formed by clustering in the output space. These virtual labels are used to extract discriminating features in the input space. This procedure is performed recursively. The resulting discriminating subspace is organized in a coarse-to-fine fashion and the corresponding information is stored in a decision tree. The algorithm has been tested with a large number of people—143 different subjects. Each subject is recorded with one video sequence of around 50-60 frames in length. A correct recognition rate of 95.1% has been achieved, which is claimed to be comparable to the nearest neighbor classification method but with much faster speed. Although not clearly stated, the system seems to still inherit the limitations of [113].

Okada et al. [115] proposes a framework for adaptive person recognition. The system consists of two stages: spatial-temporal segmentation; integration of object recognition and knowledge adaptation. For segmentation, motion and convex shape are applied to each frame to set region of interests (ROI). Bunch graph matching is then performed on these ROIs to verify whether they contain faces. Tracking of moving faces are managed by space and time discontinuity cues of the face trajectory. The second stage is further divided into three steps: recognition, knowledge adaptation and forgetting. For recognition, facial similarity between frames is analyzed to obtain a frame-to-frame similarity value. When it cannot provide enough information, torso-color analysis is applied. A simple adaptation rule is applied. When a certain face shot is identi-

fied as a known face, its view representation is always added to the existing databases and a new average of the view representations is computed. For an unknown face, a new entry is created with a new identity label. Forgetting process is able to discard redundant or irrelevant information over time. The system has been successfully tested with sequences recorded in a seminar room. Each sequence contains one person with free motions. Nevertheless, there are several assumptions of the system: the torso-color method requires that one person always dresses in the same clothes; the faces are only showing up with frontal views.

Aryananda [116] presents a humanoid robot which can silently and passively get to know new people interacting with it. The system is based on the face detection method in [61] and the face recognition method in [63]. A simple heuristic clustering method is applied to examine input data, which are collected into batches. For each batch, the system iteratively holds out every image, treating it as a class and calculates the eigenfaces of the remaining images. Threshold values for the maximum allowable distance to the generic face space and an existing known face space are empirically determined. Each face has multiple clusters. There are two major assumptions: one is that most of the input images contain faces and the other is that there are a few images per individual inside each batch. The performance shows that each database for a certain person keeps high purity, i.e. no face shots from any other persons are wrongly enrolled. But the system can learn 4 out of 9 subjects although most of the subjects have long interactions with it. Furthermore, the construction of the databases is wholly dependent on the image-based face detection algorithm without considering the temporal information from video. Another critical limitation is that only a single salient face is allowed to interact with the robot.

An “associative chaining” (AC) method is proposed in [117]. Their system allows all stages of the face recognition system to run automatically, with well recognition rate for both frontal and profile face shots. There are several stages of the AC approach. In the preprocessing stage, a combination of several face segmentation methods is applied. Background subtraction is firstly used to detect moving objects. The binary silhouettes of a certain subject are extracted at each resolution level of a multi-resolution image pyramid. Blob information and the location relation between the shoulder and the head are fused to determine the face region. “Blob” is a frequently used term in computer vision,

which can indicate a human body part or a specific region that contains a body part. A separate face-only image sequence is accordingly produced. In the learning stage, the AC algorithm is run on those face-only image sets to partition them into face clusters without using category-specific information. This is achieved by "chaining" together associations (similar views) by two types of connecting edges depending on local measures of similarity. A minimal spanning tree (MST) is applied for clustering. The learning process is grouped into two modes: a batch mode and an incremental mode according to requirements of different applications. Encouraging results are observed for data of more than 300 face image sequences obtained over several months from 17 subjects. The assumption is that each separate image sequence belongs to the same category (same subject), i.e. different people are not expected to show up at the same time.

Xiong and Jaynes [118] introduces a database quality measure which can automatically select qualified mugshots of both known and unknown faces from a twelve-camera system. A four step procedure is proposed to segment a possibly existing face from each frame. Image subtraction and dynamic threshold between the background image model and every frame are firstly applied. Those pixels which have passed the threshold test are then filtered based on color to detect regions of skin tone. In the next, a morphological erosion operation is applied to the remaining pixels to further decrease the bounding regions which might contain faces. Finally, symmetry of each region is measured to eliminate moving regions that have skin tone color but are actually not the subject's face. To automatically and incrementally construct the database, the Fraunhofer Line Discriminator (FLD) measurement approach is applied to estimate the quality of it. The measure is used to compare the quality difference between the database before and after adding the new mugshot sample. Their experiments show that automatic acquisition of a high-quality database from a twelve-camera network is feasible. Although not clearly stated, their method is subject to fail when there are more than one faces showing up at the same time, and when there is a sudden change that one face disappears and another face is showing up. Their system deals well with only frontal faces due to the limitation of the image-based face recognition method. Additionally, each of the twelve cameras needs to be connected to a PC, which is quite a big hardware requirement.

Patents and published patent applications are another important reference

resource. They normally demonstrate the industry's interests on a specific research area and are especially referential for engineers. There are some published patents and patent applications dealing with the problem of unsupervised and automatic face recognition. In the following, we survey four important ones.

An earlier patent in [120] describes an autonomous face recognition machine. It is claimed to be insensitive to variations in brightness, scale, focus and can operate without any human intervention or inputs. Image differencing method is applied to determine a region of moving object from a frame image. A facial feature based face detector is performed on the motion region to detect whether it contains a face. Then an elliptical mask with the facial characteristics (is defined as a gestalt-face) is extracted and compared with the stored data. The system assumes the following: faces are always moving (stable faces are not able to be detected); there is only one face existing at one time; a constant illumination for the scene; faces are showing up with frontal or at least nearly frontal views. The system is applied to identify known persons, which does not deal with learning new faces and update known faces.

Cumbers [53] describes a passive system for biometric customer identification and tracking, which is a more recent patent. It defines a more systematic idea on how such an automatic system will behave. It is expected to passively memorize new customers, update known ones and remove the people of non-interests (those who come once or occasionally in a certain period of time). The automatic system can frequently welcome visiting clients and can be alert to those with bad credits. The patent is useful as a meaningful guidance for research directions. It has not discussed any technical details, such as which biometric data are useful and which algorithms are suitable for such definitions.

[121] is a published patent application, in which an adaptive facial recognition system and method is demonstrated. It is a more detailed system than [53], applying the facial characteristics for recognition. Faces are firstly detected by a generic template of eigenfaces. A tracking method (no details) is also included to track a detected face (a multi-face situation is not discussed). The detected face is then compared with the database by any existing distance criterion (actually a classifier by our definition). If the face shot is determined to belong to any known classes, it is to be updated based on selection rules (not mentioned in any detail). Otherwise, a new class is created.

Center JR [122] patents a real-time facial recognition and verification system. Motion, color and blob information are combined to localize a face region in an image. A series of templates (eigen templates) are used to process the detected face region. Matching is also based on PCA method. Although somehow similar to [120], the proposed ideas are to be performed with a much higher speed.

In conclusion, we list in the following the common limitations of most current automatic systems:

- Firstly and most importantly, they assume a single face case within a certain sequence. The occurrence of multiple people with occlusions or a sudden change from one person to another may lead to failure. Although not subject to happen for video sequences from live cameras, the sudden shot change often occurs in films or TV programs.
- The one-face assumption greatly decreases the complexity of the automatic procedure. However, it is a very strict requirement for the environment and does not have enough practical application prospects.
- Moreover, the general ways to further compensate for face detection and face recognition limitations by using the temporal information from video sequences are not sufficiently explored.
- Another crucial point is that research on automatically building and updating high efficient and high quality databases is far from adequate. As we can learn from the 3D-model based approaches, the success of database construction is one of the key to the success of recognition. Furthermore, how to online update the already constructed database is another issue for achieving the robust recognition.

2.8 Summary and Discussions

In this Chapter, we start our fundamental discussion with the definition of generic biometric recognition, its procedures and categories.

From the introduction of cognitive science, we have learned that, the complex human brains and its self-development makes it possible for performing successful vision, object recognition and face recognition. Although how exactly occur during face recognition in cortex is still not clear, the research advances in cognitive science provide us profound inspirations on designing a

machine face recognition system.

The exploration of machine-based biometric procedures, comparison of different biometric methods, and their market shares imply the possible and promising future of applying the face recognition technology. Through the face recognition procedure, we have put much emphasize on the enrollment and update procedure, which are typically not taken into much consideration. The survey and corresponding comments on the research in face detection, face tracking, and face recognition, help us get to learn the state-of-the-art in this area, and more importantly the fundamental research demands in face recognition.

By combining the inspirations from cognitive research and demands in machine-based face recognition technologies, we are ready to define an ideal system which is able to mimic the fundamental intelligence of the human brains, which should include: being able to combine all available information during recognition, and to run in an unsupervised, fully automatic and non-invasive way. The whole recognition procedure containing the enrollment/training step, and the update step should be completely automatic.

In the following chapters, we are mainly focusing on the methods and proposals on how to achieve this goal. In another words, we are introducing here the face recognition system that features the following: firstly and most importantly, the system is free of any human supervisor and has no requirement for user's cooperation—that means fully automatic and completely unsupervised, both for the recognition procedure and the training procedure. Stable running is accordingly important for an autonomous system. We have to guarantee that no crash and no exertion can happen. Self-adaptive is another feature to keep updating the most recent face shots. At the same time, the system is expected to keep robust performance.

References

[1] H. Chen, and A. K. Jain: Dental Biometrics: Alignment and Matching of Dental Radiographs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, 2005, pp.1319−1326

[2] D. Lazer, Ed.: *DNA and the Criminal Justice System: The Technology of Justice*, MIT Press, Cambridge, MA, 2004

[3] D. A. Norman: Twelve Issues for Cognitive Science. *Cognitive Science*, Vol. 4, Issue 1, 1980, pp.1−32

[4] Halfon N, Shulman E, and Hochstein M, eds.: *Brain Development in Early Childhood*. Building Community Systems for Young Children, UCLA Center for Healthier Children, Families and Communities, 2001

[5] J.P. de Magalhaes, and A. Sandberg: Cognitive aging as an extension of brain development: A model linking learning, brain plasticity, and neurodegeneration. *Mechanisms of Ageing and Development*, Vol. 126, 2005, pp.1026−1033

[6] G.M. Shepherd: The Synaptic Organization of the Brain 5th Edition, Oxford, Oxford Univ. Press, 2004, p.6

[7] C. Koch: Biophysics of Computation. *Information Processing in Single Neurons*, New York, Oxford Univ. Press, 1999, p.87

[8] Henry Gray: *Anatomy of the Human Body*, 1918

[9] R. S. Michalski., G. Carbonell, and T. M. Mitchell: *Machine Learning: An Artificial Intelligence Approach*, Berlin, Springer-Verlag, 1984

[10] P. Thagard: Mind: Introduction to Cognitive Science, 2nd Edition. Cambridge, The MIT Press, 2005

[11] B.A. Wandell: What's in your mind? *Nature Neuroscience*, Vol. 11, No. 4, 2008

[12] B.A. Wandell, S.O. Dumoulin, and A. A. Brewer: Visual Field Maps in Human Cortex. *Neuron*, Vol. 56, No. 2, 2007

[13] T. Serre, L. Wolf, *et al*.: Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 3, 2007, pp.411−426

[14] L. Wiskott: How does our visual system achieve shift and size invariance? In: J.L. van Hemmen and T.J. Sejnowski, *23 Problems in Systems Neuroscience*. Oxford, Oxford University Press, 2006

[15] J. Mutch and D. Lowe: Multiclass object recognition using sparse, localized features. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006

[16] M. Ranzato, F. Huang, *et al*: Unsupervised learning of invariant feature hierarchies, with application to object recognition. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007

[17] D. Mou: Autonomous Face Recognition. Ph.D Dissertation, http://vts.uni-ulm.de/query/longview.meta.asp?document_id=5370, accessed 28 October 2005

[18] M. Johnson, S. Dziurawiec, *et al*.: Newborns preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, Vol. 40, 1991, pp.1−19

[19] J. Sergent, S. Ohta, and B. MacDonald: Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain*, Vol. 15, No. 1, 1992, pp.15−36

[20] N. Kanwisher, J. McDermott, and M. M. Chun: The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, Vol. 17, No. 11, 1997, pp.4302−4311

[21] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini: The distributed human neural system for face perception. *Trends in Cognitive Sciences*, Vol. 4, Issue 6, 2000, pp.223−231

[22] E. H. Aylward, J. E. Park, *et al*.: Brain Activation during Face Perception: Evidence of a Developmental Change. *Journal of Cognitive Neuroscience*, Vol. 17, Issue 2, 2005

[23] N. Kanwisher, and G. Yovel: The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, Vol. 361, 2006, pp.2109-2128

[24] M. J. Tarr, and I. Gauthier: FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, Vol. 3, No. 8, 2000

[25] I. Gauthier, and N. K. Logothetis: Is face recognition not so unique, after all? *Cognitive Neuropsychology*, Vol. 17, 2000, pp.125−142

[26] M. Riesenhuber and T. Poggio: Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, Vol. 12, 2002, pp.162−168

[27] T. J. Andrews and D. Schluppeck: Neural responses to Mooney images reveal a modular representation of faces in human visual cortex. *Neuroimage*, Vol. 21, Issue 1, 2004

[28] P. Rotshtein, R. N. Henson, *et al*.: Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, Vol. 8, No. 1, 2005

[29] C. G. Gross: Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London B.*, Vol. 335, 1992, pp.3−10

[30] A. Mechelli, C.J. Price, *et al*.: Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral Cortex*, Vol. 14, No. 11, 2004

[31] M.R. Johnson, K.J. Mitchell, *et al*.: A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *Neuroimage*, Vol. 37, Issue 1, 2007

[32] P. Sinha, B. Balas, *et al*.: Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of The IEEE*, Vol. 94, No. 11, 2006

[33] M. Dawson: *Understanding Cognitive Science*. Malden, Blackwell, 1998

[34] A A. Ross, K. Nandakumar and A.K. Jain: Handbook of Multibiometrics. Boston, Springer, 2006

[35] H. Chen, and B. Bhanu: Human Ear Recognition in 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 4, 2007, pp.718−737

[36] Z. Korotkaya: Biometrics Person Authentication: Odor. http://www.it.lut.fi/kurssit/03-04/010970000/ seminars/Korotkaya.pdf, accessed 08 December 2005

[37] R. Palaniappan, and D. P. Mandic: Biometrics from Brain Electrical Activity: A Machine Learning Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 4, 2007, pp.738−742

[38] http://www.biometricgroup.com/reports/public/market_report.html, accessed 15 December 2007

[39] Information Technology. Biometric Data Interchange Formats. Iris Image Data. ISO/IEC 19794-6:2005

[40] Faulds, Henry: On the Skin-furrows of the Hand. *Nature*, Macmillan and Co., London, October 28, 1880, p. 605

[41] Herschel, W. J.: Skin Furrows of the Hand. *Nature*, Macmillan and Co., London, Nov. 25, 1880, p. 76

[42] Galton, Sir Francis: *Finger Prints*. Macmillan and Co., London, 1892.

[43] http://www.fbi.gov/hq/cjisd/iafis.htm, accessed 15 December 2007

[44] D Maltoni, D Maio, *et al.*: Handbook of Fingerprint Recognition, New York, Springer, 2003

[45] E. Hjelmas and B.K. Low: Face Detection: A Survey. *Computer Vision and Image Understanding*, 2001, Vol. 83, No. 3, 2001, pp.236−274

[46] M. Yang, D.J. Kriegman, and N. Ahuja: Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, 2002, pp.34−58

[47] M. Yang: Recent Advances in Face Detection. *IEEE ICIP 2004 Tutorial*, Cambridge, UK, http://vision.ai.uiuc.edu/mhyang/face-detection-survey.html, accessed 13 October 2005

[48] S. Gong, S. McKenna, and A. Psarrou: *Dynamic Vision: From Images to Face Recognition*, London, Imperial College Press, 2000

[49] K. C. Yow and R. Cipolla: Feature-Based Human Face Detection. *Image and Vision Computing*, Vol. 15, No. 9, 1997, pp.713−735

[50] J. Yang and A. Waibel, "A Real-Time Face Tracker", *Proceedings of the 3rd Workshop on Applications of Computer Vision (WACV'96)*, 1996, pp.142−147

[51] G. Yang and T. Huang: Human Face Dtection in Complex Background. *Pattern Recognition*, Vol. 27, No. 1, 1994, pp.53−63

[52] C. Kotropoulos and I. Pitas: Rule-Based Face Detection in Frontal Views. *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP'97)*, Vol. 4, 1997, pp.2537−2540

[53] B. Cumbers (2003): Passive Biometric Customer Identification and Tracking System. *U.S. Patent*, 6554705, April 2003

[54] K. Sung and T. Poggio: Example-Based learning for view-Based Human Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, 1998, pp.39−51

[55] H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, 1998, pp.23−38

[56] R, Féraud, O. Bernier, *et al.*: A Fast and Accurate Face Detector Based on Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, 2001, pp.42−53

[57] E. Osuna, R. Freund, and F. Girosi: Training Support Vector Machines: An Application to Face Detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp.130−136

[58] H. Schneiderman and T. Kanade: A Statistical Method for 3D Object Detection Applied to Faces and Cars. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2000, pp.746−751

[59] M. Yang, D. Roth, and N. Ahuja: A SnoW-Based Linear Subspaces for Face Detection. In: S. Solla, T. Leen, and K. Müller, eds. *Advances in Neural Information Processing System 12*, MIT Press, 2000, pp.855−861

[60] P. Viola and M. Jones: Robust Real-time Object Detection. *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, July 13, 2001

[61] M. Jones, P. Viola: Fast Multi-view Face Detection. *Mitsubishi Electric Research Laboratories Technical Reports*, TR2003-96, 2003, http://www. merl.com/ reports/ docs/ TR2003-96.pdf, accessed 12 October 2005

[62] Z. Zhang, L. Zhu, *et al.*: Real-Time Multi-View Face Detection. Proceedings of *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, May 2002, pp.149-154

[63] M. Turk and A. Pentland: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp.72−86

[64] S. Palanive, B.S. Venkatesh, and B. Yegnanarayana: Real time face recognition system using autoassociative neural network models. *IEEE Conference Proceedings on Acoustics*, *Speech, and Signal Processing (ICASSP '03)*, Vol. 2, 2003, pp.833−836

[65] T. kim, S. Lee, *et al.*: Integrated approach of multiple face detection for video surveillance. *Proceedings of IEEE 16th Conference on Pattern Recognition*, Vol. 2, 2002, pp.394−397

[66] D. Butler, C. McCool, *et al.*: Robust Face Localisation Using Motion. Colour & Fusion *Proceedings of Digital Image Computing: Techniques and Applications (DICTA 2003)*, 2003, pp.899−908

[67] C. Wren, A. Azerbayejani, *et al.*: Pfinder: A Real-Time Tracking of Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp.780-785

[68] Y. Raja, S.J. McKenna, and S. Gong: Tracking and Segmenting People in Varying Lighting Conditions Using Color. *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, 1998, pp.228-233

[69] K. Schwerdt and J. Crowley: Robust Face Tracking Using Colour. *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, 2000, pp.90-95

[70] S. Birchfield: Elliptical Head Tracking Using Intensity Gradients and Color Histograms. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp.232-237

[71] R.C. Verma, C. Schmid, and K. Mikolajczyk: Face detection and tracking in a video by propagating detection probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, Issue 10, 2003, pp.1215−1228

[72] R. Chellappa, C.L. Wilson and S. Sirohey: Human and Machine Recognition of Faces: A Survey. *Proceedings of IEEE*, Vol. 83, No. 5, 1995, pp.705−740

[73] W. Zhao, R. Chellappa, *et al.*: Face Recognition: A Literature Survey. *Technical Report (CS-TR-4167R)*, *University of Maryland*. ftp://ftp.cfar.umd.edu, accessed 20 August 2005

[74] K. Bowyer, K. Chang, and P. Flynn: A survey of Approaches and Challenges in 3D and Multi-Modal 3D 2D Face Recognition. *IEEE Transactions on Computer Vision and Image Understanding*, Vol. 101, No. 1, 2006, pp.1−15

[75] X. Lu, and A. Jain: Deformation Modeling for Robust 3D Face Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 8, 2008, pp.1346−1357

[76] A. Franco, D. Maio and D. Maltoni: 2D Face Recognition based on Supervised Subspace Learning from 3D Models. *Pattern Recognition*, Vol. 41, No. 12, 2008, pp.3822−3833

[77] P. J. Phillips, P. Rauss and S. Der: FERET (Face Recognition Technology) Recognition Algorithm Development and Test Report. *Technical Report ARL-TR 995*, U.S. Army Research Laboratory, 1996

[78] P. J. Phillips, H. Moon, *et al.*: The FERET Evaluation Method for Face Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp.1090−1104

[79] D. M. Blackburn, M. Bone and P.J. Phillips: FRVT 2000 Evaluation Report. *Technical Report*, Feb. 16th, 2001, http://www.frvt.org, accessed 29 March 2007

[80] P. J. Phillips, P. Grother, *et al.*: FRVT 2002 Evaluation Report, *Technical Report*, March, 2003, http://www.frvt.org, accessed 29 March 2007

[81] P. J. Phillips, W. T. Scruggs, *et al.*: FRVT 2006 and ICE 2006 Large-Scale Results. *Technical Report*, March 2007, http://www.frvt.org, accessed 29 March 2007

[82] V. Blanz, S. Romdhami, and T. Vetter: Face identification across different poses and illuminations with a 3D morphable model. *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 2002, pp.202−207

[83] V. Blanz and T. Vetter: Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 25, 2003, pp.106−1074

[84] V. Bruce: *Recognizing Faces*. London, Lawrence Erlbaum Associates, 1988

[85] V. Bruce, P.J.B. Hancock, and A.M. Burton: Human Face Perception and Identification. In: *Face Recognition: From Theory to Applications*, Berlin, Springer-Verlag, 1998, pp.51−72

[86] M.S, Bartlett, J.R. Movellan and T.J. Sejnowski: Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks*, Vol. 13, No. 6, 2002, pp.1450−1464

[87] D.L. Swets and J.J. Weng: Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, 1996, pp.831−836

[88] P. Belhumeur, J.P. Hespanha, and D.J. Kriegman: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp.711−720

[89] A.V. Nefian and H.H. Hayes III: Hidden Markov Models for Face Recognition. *IEEE International Conference on Acoustic, Speech and Signal Processing*, Vol. 5, 1998, pp.2721−2724

[90] V. V. Kohir and U. B. Desai: Face recognition using DCT-HMM approach. *Workshop on Advances in Facial Image Analysis and Recognition Technology (AFIART)*, June 1998

[91] R. Tjahyadi, W. Liu, and S. Venkatesh: Application of the DCT Energy Histogram for Face Recognition. *Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)*, 2004, pp.305−310

[92] H. Kang, T. F. Cootes, and C. J. Taylor: A comparison of face verification algorithms using appearance models. *Proceedings of The British Machine Vision Conference*, Vol. 2, 2002, pp.477−486

[93] X. Lu, Y. Wang, and A. K. Jain: Combining classifiers for face recognition. *Proceedings of the IEEE International Conference on Multimedia & Expo*, Vol. 3, July 2003, pp.13−16

[94] R. Singh, M. Vatsa, *et al.*: A Mosaicing Scheme for Pose Invariant Face Recognition. *IEEE Transactions on Systems, Mans and Cybernetics - B, Special Issue on Biometrics*, Vol. 37, Issue 5, 2007, pp.1212−1225

[95] M. Bicego, U. Castellani, V. Murino: Using Hidden Markov Models and Wavelets for face recognition. *Proceedings of IEEE International Conference on Image Analysis and Processing (ICIAP03)*, 2003, pp.52−56

[96] L. Wiskott, J. Fellous, *et al.*: Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp.775−779

[97] C. Liu and H. Wechsler: A Gabor Feature Classifier for Face Recognition. *Proceedings of Eighth IEEE International Conference on Computer Vision*, Vol. 2, 2001, pp.270−275

[98] B.A. Draper, K. Baek, *et al.*: Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, Vol. 91, No. 1, 2003, pp.115−137

[99] A.M. Martinez and A.C. Kak: PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, 2001, pp.228−233

[100] M.H. Yang: Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods. *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'02)*, May 2002, pp.215−220

[101] J. Lu, K.N.Plataniotis, and A.N. Venetsanopoulos: Face Recognition Using Kernel Direct Discriminant Analysis Algorithms. *IEEE Transactions on Neural Networks*, Vol. 14, No. 1, 2003, pp.117−126

[102] Y. Zhang, L. Lang and O. Hamsici: Subspace Analysis for Facial Image Recognition: A Comparative Study. http://www.stat.ohio-state.edu/~goel/ STATLEARN/, accessed 12 October 2006.

[103] G. Guo, S. Z. Li, and C. Kapluk: Face recognition by support vector machines. *Image and Vision Computing*, *Special Issue on Artificial Neural Networks for Image Analysis and Computer Vision*, Vol. 19, No. 9-10, 2001, pp.631−638

[104] B. Heisele, P. Ho and T. Poggio: Face Recognition with Support Vector Machines: Global versus Component-based Approach. *Proceedings of IEEE International Conference on Computer Vision*, 2001, pp.688−694

[105] J. Huang, V. Blanz, and B. Heisele: Face Recognition Using Component-Based SVM Classification and Morphable Models. *SVM 2002*, 2002, pp.334−341

[106] S. Lawrence, C.L. Giles, *et al.*: Face Recognition: A Convolutional Neural Network Approach. *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, 1997, pp.98−113

[107] T. Kurita, M. Pic, and T. Takahashi: Recognition and Detection of Occluded Faces by A Neural Network Classifier with Recursive Data Reconstruction. *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, 2003, pp.53−58

[108] Xiaoming Liu, and Tsuhan Chen: Video-Based Face Recognition Using Adaptive Hidden Markov Models. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, 2003, pp.340−345

[109] V. Krueger and S. Zhou: Exemplar-based Face Recognition from Video. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, May 21−22, 2002, pp.175−180

[110] A: Hadid and M. Pietikäinen: From Still Image to Video-Based Face Recognition: An Experimental Analysis. *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)*, 2004, pp.813−818

[111] X. Tang and Z. Li: Video Based Face Recognition Using Multiple Classifiers. *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)*, 2004, pp.345−349

[112] O. Arandjelovic and R. Cipolla: Face Recognition from Face Motion Manifolds using Robust Kernel Register-Average Distance. *IEEE International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, Vol. 5, 2004, p.70

[113] J. Weng and W. Hwang: Toward Automation of Learning: The State Self-Organization Problem for a Face Recognizer. *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, 1998, pp.384−389

[114] J. Weng, C. Evans, and W. Hwang: An Incremental Learning Method for Face Recognition under Continuous Video Stream. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp.251−256

[115] K. Okada, L. Kite, and C. von der Malsburg: An Adaptive Person Recognition System. *Proceedings of the IEEE International Workshop on Robot-Human Interactive Communication*, 2001, pp.436−441

[116] Lijin Aryananda: Recognizing and Remembering Individuals: Online and Unsupervised Face Recognition for Humanoid Robot. *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, Vol. 2, 2002, pp.1202−1207

[117] B. Raytchev, H. Murase: Unsupervised Face Recognition by Associative Chaining. *Pattern Recognition*, Vol. 36, No. 1, 2003, pp.245−257

[118] Q. Xiong and C. Jaynes: Mugshot Database Acquisition in Video Surveillance Networks Using Incremental Auto-Clustering Quality Measures. *Proceedings of the 2003 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, 2003, pp.191−198

[119] H. Wechsel, V. Kakkad, *et al.*: Automatic Video-Based Person Authentication Using the RBF Network. *Proceedings of 1st International Conference on Audio And Video-based Biometric Person Authentication*, 1997, pp.85−92

[120] C. Lambert (1991): Autonomous Face Recognition Machine. *U.S. Patent*, 5012522, April, 1991

[121] Y.T. Lin (2002): Adaptive Facial Recognition System and Method. *U.S. Patent application publication*, US2002/0136433, Sep. 26, 2002

[122] J.L. Center JR (2003).: Real-time Facial Recognition and Verification System. *U.S. Patent application publication*, US2003/0059124, Mar. 27, 2003

3 Combined Face Detection and Tracking Methods

Abstract This chapter explores the algorithms of automatically detecting and extracting faces of interest from live video input. It is the first but crucial step for an intelligent, automatic and unsupervised face recognition system. In section 1, the proposals of face detection are briefly overviewed. Imaged-based face detection issues are discussed in section 2, including the choice of detection algorithms, the definition of face region based on the eye distance, and the critical cases of evaluating detection performance. In section 3, temporal-based face detection algorithms are proposed. Through the definition of "search region" and the analysis of temporal change, the motion-based face detector can significantly benefit from the temporal context from video sequences. Summary and future research directions are the final parts of this chapter.

3.1 Introduction

As mentioned in Section 2.1, the combination of motion detection and image-based face detection techniques is commonly applied in many video-based face processing systems. Although motion information alone can be used to detect a face in video [1], it is normally used in the preprocessing step [2, 3, 4] to segment moving objects from background, e.g. to determine the regions of interests. An image-based (frame-based) face detector is then applied to verify the faces in those regions.

The image-based detector can reduce the number of false positives from the temporal detector, but it has difficulties in detecting the false negatives. That is to say, the faces that are not detected in the temporal detector are not likely be detected in the image-based detector as well. And motion is subject to produce many false negatives in real cases without any human supervision. Background subtraction and frame differencing are the two main methods to detect motion. The first one requires a static background. However, there are unavoidable failures when subtracting the initialized background from the current frame with a varying background. Complicated adaptive background models are introduced as a compensation in [5], but with significant additional computation efforts. For the frame differencing method, when a face is not moving apparently, or other moving objects existing in the background, it might not work well.

To achieve better results, we include an image-based face detector (IFD) in the first step and then apply the motion-based face detector (MFD) as a post-processing procedure rather than in the preprocessing step. Apparently, as the first step, IFDs are generally the most robust and can achieve much higher face detection rate than MFDs. But no technology is perfect. When the first step fails, the corresponding motion information can still help to complete the face detection procedure.

Inside IFD, besides the detector, we further introduce a face region estimation method based on anatomy. The region is used not only for extracting faces but also for supporting the following MFD. Temporal information and the logic deduction of human movement analysis are combined for the MFD. The hybrid methods, especially the non-mathematical fundamentals make the detection procedure intelligent. From the technical point of view, the detection rate is significantly improved while the mathematical complex still keeps low. The achievement of higher face detection rate is due to the improvement of false negatives. The low computational effort is achieved based on the daily people moving speed and adjacent frame differencing. Fig. 3.1 illustrates the procedure for face detection.

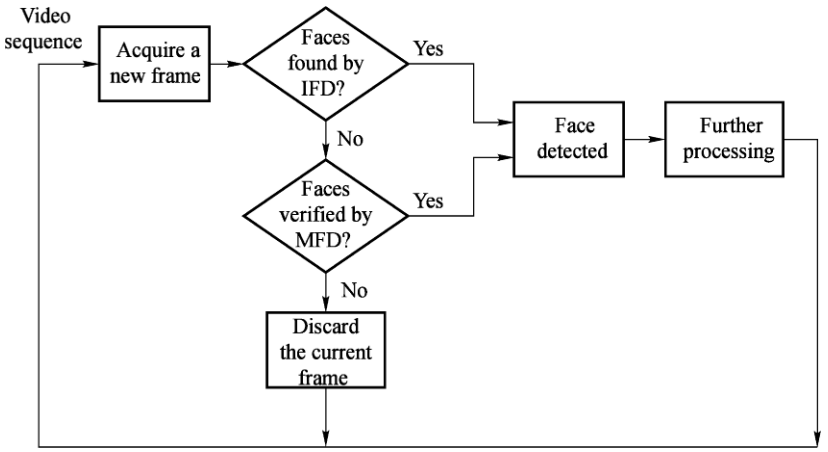


Fig. 3.1 Function blocks of face detection

3.2 Image-based Face Detection

The IFD includes two parts. One is the face detector which indicates where the face is. Moreover, it should also be able to detect the eye positions. The other one is the face region estimator which can extract the face region according to the detected eye positions. The face region estimator as well as the detected eye positions is required for a following MFD.

3.2.1 Choice of the Detection Algorithm

From the performance point of view, it is always an optimized choice if we can include the best available technology for our system. The two main third-party components we need for the whole automatic system are an image-based face detector and a recognition classifier. The face recognition technology FaceVACS® (from Cognitec Systems GmbH) ranked the first in the FRVT 2002 test [6]. Hence, we take its SDK (Software Development Kit) with a special academic price offer. The SDK also includes an image-based face detector. Although the detector is not as outstanding as the recognition classifier, it still meets our requirements, eye detection algorithms are included and a fast proc-

essing speed can be achieved. For the above reasons, we have taken the FaceVACS® as our face detector rather than other popular detection algorithms such as [7] and [8]. Nevertheless, any other image-based face detection method is encouraged for the readers to experiment with our proposals. The only requirement is that the eye positions are expected to be available. Additionally, fast processing speed is preferred.

3.2.2 Overview of the Detection Algorithm

The detector applied here is a rule-based face detection method by using a predefined face template, according to the face detection categories described in Section 2.1. The following two paragraphs briefly introduce the detection algorithm from the publicly provided information in [9].

To locate one face, a so-called image pyramid is formed from the original image. An image pyramid is a set of copies of the original image at different scales, thus representing a set of different resolutions. A mask is moved pixelwise over each image in the pyramid, and at each position, the image section under the mask is passed to a function that assesses the similarity of the image section to a predefined face template. If the similarity value is high enough, the presence of a face at that position and the corresponding resolution is assumed. From that position and resolution, the position and size of the face in the original image can therefore be calculated.

From the position of the detected face, an estimation of the eye positions can be derived. In a neighborhood around those estimated positions, a search for the exact eye positions is started. This search is very similar to the search for the face position, the main difference being that the resolution of the images in the pyramid is higher than the resolution at which the face was found before. The positions yielding the highest similarity values are taken as final estimates of the eye positions.

3.2.3 Face Region Estimation

The study of anatomy [14] reveals that most human faces follows a certain

range of width-height ratio, which can be also represented by the eye distance. In this book, we empirically apply the following equation to extract the face region from detected eye positions.

$$\begin{aligned} F_{\mathrm{W}} &= 2.5E_{\mathrm{D}} \\ F_{\mathrm{H}} &= 3.5E_{\mathrm{D}} \end{aligned} \tag{3.1}$$

where E_{D} is the eye distance, F_{W} and F_{H} are face width and face height respectively.

Fig. 3.2 uses grids to proportionally illustrate the relationship between the estimated face height, face width and the detected eye distance.

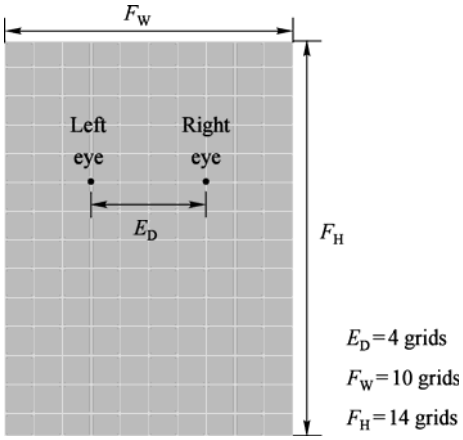


Fig. 3.2 Proportional illustration of face region depending on the eye distance

To evaluate the face extraction method, we have collected images containing a face from internet, TVs and video cameras with different resolutions. Each image is processed through the above mentioned image-based face detection method to obtain the eye distance of a certain face. Face regions are then extracted by equation (3.1).

Fig. 3.3 shows examples of the extracted faces and the thumbnails of the corresponding original images. For the sake of displaying, the whole images are re-sampled to the same resolution while the face regions are shown in their original size. Thumbnails re-sampled from TV programs and video cameras are listed in (a). The corresponding face regions in their original resolutions are extracted from each image in (a), which are shown in (b). It can be seen that although there is a large resolution difference of the images, each face is

well extracted without losing any useful facial information.

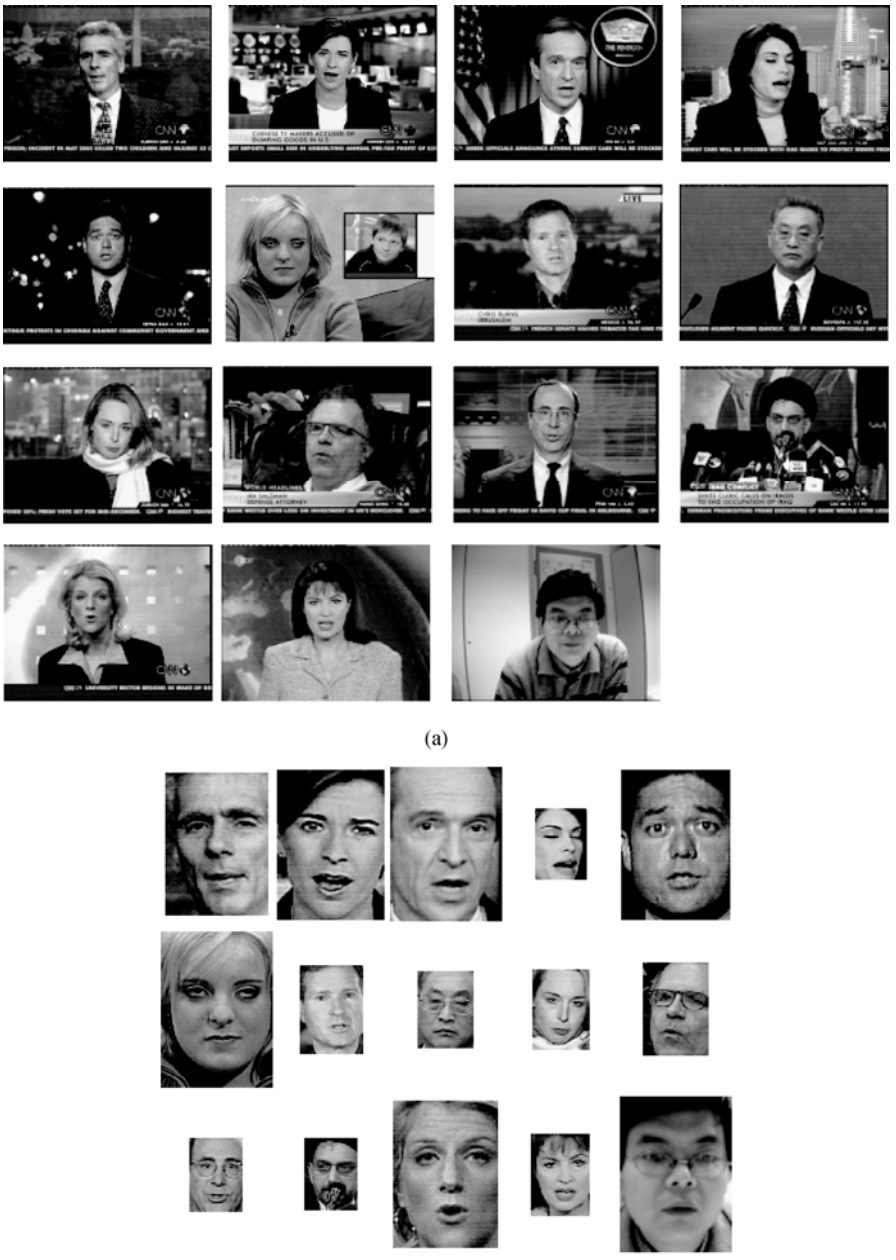


Fig. 3.3 Example of face extraction

3.2.4 Face Detection Quality

To evaluate the performance of a face detection algorithm, the following factors shall be considered:

- Head pose
- Scale
- Facial expression
- Lighting
- Motion blur (detection from video)

Variance of head poses can greatly influence the success of detection. Three dimensional motion of a face includes pitch, yaw, roll or combination of them. Fig. 3.4 shows examples of those pose changes.

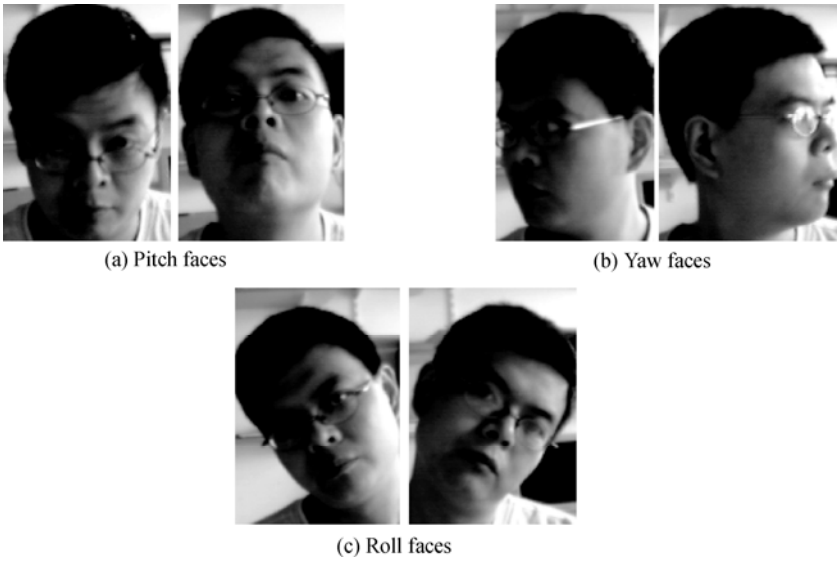


Fig. 3.4 Variance of 3D head poses

Scale and facial expressions are important factors for the image-based detection algorithms since they significantly change the appearance of faces. Varying lighting conditions can lead to failures of many face detection algorithms especially for color models. Motion blur typically occurs with moving faces taken from standard videos.

We have made experiments to qualitatively evaluate the FaceVACS detection method, which is shown in detail in Section 8.2. Although relatively ro-

bust for pitch, scale and facial expression changes, it is sensitive to roll faces and may fail with profiles.

To keep the generality, we study here all common failures of face detection algorithms and explore the ways to benefit from temporal information from video. False positives from video can be further divided into two major groups. One type of failure occurs when some static face-like objects are in the background. The other kind occurs for some objects which are falsely detected as faces due to motion artifacts. We will discuss the false positive problem and the solution for that in more details in Section 4.3.3. In the following section, we mainly focus on how to handle the false negatives.

3.3 Temporal-based Face Detection

3.3.1 Overview

Widely used algorithms for predicting motions are Kalman filters, Hidden Markov Models and Conditional Density Propagation (Condensation). In [11], those methods are described in detail especially for face tracking. We introduce here a simple frame differencing method for motion detection, which is based on the eye distance from IFD. No motion estimation vector is required. It is therefore computationally more efficient for a real-time system if compared to existing techniques. The proposed temporal-based face detector can be divided into three parts by its functionality. A face region is defined for each detected face so that it contains one and only one face. An expanded region centered on the face region is then marked according to an average day-to-day maximum expected moving speed. The region is named the search region, inside which a face is expected to stay for at least two consecutive frames of video. Motion detection is applied in each search region to decide whether there is much temporal difference. Small motion meets the expectation that a certain face is in the search region while big motion indicates that the corresponding face does not exist in the search region any more. The temporal-based face detector is equivalent to the functional block “Face found by MFD ?” in Fig. 3.1

and can be further decomposed into four sub-blocks as shown in Fig. 3.5.

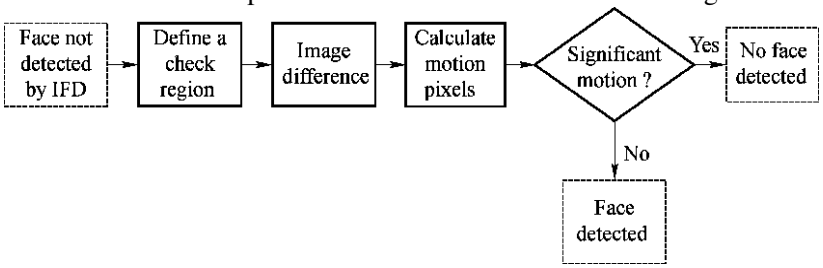


Fig. 3.5 Function blocks of temporal-based face detector

3.3.2 Search Region Estimation

In daily life, most people don't move extremely fast. The general walking speed of a person is around 4-5 km/hour, approximated as 110-140cm/s, which can be estimated as the motion speed of a face. We accordingly suppose that a human face within a video sequence follows that speed limit, providing valuable temporal information for detecting faces. Stereoscopic literature [15] finds out that the vast majority of adults have interpupillary distance in the range 5.5-7cm, which is about 1/20 of the assumed face motion speed by only considering the absolute value. The motion speed of one face in videos can be thus represented by the following equation:

$$D_F \le 20E_D \tag{3.2}$$

where D_F denotes the maximum speed that one face can move, and E_D equals to the absolute value of the detected eye distance although it is physical representing speed.

For the convenience of processing frame images, Equation (3.2) can be converted to:

$$M_F \le 20E_D\ \Delta t \tag{3.3}$$

where M_F denotes the maximum distance that a face can move between two successive frames and Δt is the time interval between the two frames.

M_F actually defines a region inside which a certain face is expected to remain for at least two consecutive frames. This region is named the search region here.

We have made an experiment to assess the relationship between the eye distance and the search region. Two video sequences are recorded with one

person moving as fast as possible. The resolution is 384×288 pixels and the recording speed is based on PAL standard with 25 fps. The thumbnails shown in Fig. 3.6 are the main face regions cut from the original images. Fig. 3.6(a) lists the consecutive frame images with a person moving the head up and down as fast as possible. Fig. 3.6 (b) shows the consecutive frame images with a person moves the head left and right as fast as possible.

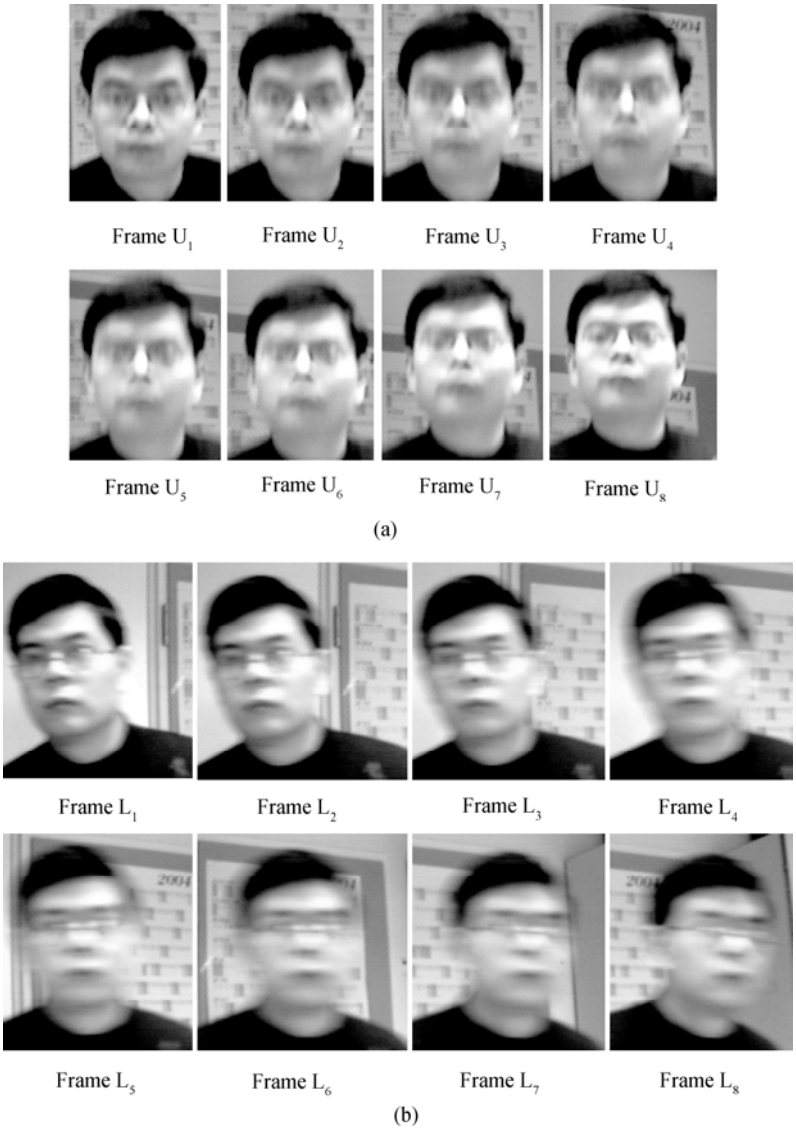


Fig. 3.6 Two examples of face region fast movement

Table 3.1 Comparison of M_F and actual speed shown in Fig. 3.6

	Image name	Left eye position	Right eye position	Actual Speed (in pixel)	Eye Distance (in pixel)	M_F
Example (a)	Frame U_1	(76, 208)	(109,207)		31	20×31×1/25 =24.8
	Frame U_2	(77,193)	(109,193)	15		
	Frame U_3	(79,173)	(109,175)	20		
	Frame U_4	(79,153)	(110,153)	20		
	Frame U_5	(81,132)	(112,133)	21		
	Frame U_6	(82,111)	(111,111)	22		
	Frame U_7	(84,93)	(113,92)	22		
	Frame U_8	(84,75)	(114,75)	17		
Example (b)	Frame L_1	(52, 65)	(84,61)		33	20×33×1/25 =26.4
	Frame L_2	(67,66)	(99,61)	15		
	Frame L_3	(91,66)	(121,62)	24		
	Frame L_4	(116,63)	(149,61)	25		
	FrameL_5	(149,62)	(182,60)	33		
	Frame L_6	(187,62)	(220,61)	38		
	Frame L_7	(228,61)	(260,61)	41		
	Frame L_8	(266,63)	(298,65)	38		

We manually measure the eye positions and eye distance of each frame image in pixels and accordingly obtain the actual speed (in pixels) of the face region in every frame. M_F is calculated by Equation (3.3). The results are listed in Table 3.1, which point to the following conclusions:

- All image frames except L_5, L_6, L_7 and L_8 fulfill Equation (3.3).
- Due to extremely fast moving speed, U_5, U_6, U_7 and U_8 contain significant motion blur and are not suitable for further processing any more. With a commonly used 25fps capturing system, such moving speeds can be omitted without loosing the generality. For applications in high-speed motion with high speed capturing systems, Equation (3.3) can be easily revised with a bigger coefficient.

According to Equation (3.3), however, there is a typical case that the search region estimation does not have enough contribution to face detection. In the above example, the sequences are processed offline, which indicates the assumption of the processing speed with 25fps. If the detection system runs slower

or the sequences are captured with a less fps, e.g. with Δt equals to 1s, the calculated search region is even larger than the original image. By such a big search region, motion detection is not helpful any more since moving objects in the background or a changing background may lead to errors. Moreover, the actual processing speed has to be online detected according to Equation (3.3), which brings more calculation efforts. For further simplification, we suppose that the detection system is working with a constant speed of 20fps and Δt can be accordingly removed from Equation (3.3). This assumption is logic enough if a face processing system is built on a chip. Even for a system implemented by software, it is becoming less critical to achieve a real-time processing speed. Recent advances in real-time robust face detection [7, 8, 13] strongly support the above assumption. Equation (3.3) is converted as the following:

$$M_{\mathrm{F}} \le E_{\mathrm{D}} \tag{3.4}$$

Fig. 3.7 proportionally illustrates the search region by this simplification. Under this condition, the search region dimension varies with the eye distance by the following equation:

$$\begin{aligned} C_{\mathrm{W}} &= 4.5E_{\mathrm{D}} \\ C_{\mathrm{H}} &= 5.5E_{\mathrm{D}} \end{aligned} \tag{3.5}$$

where E_{D} is the eye distance, C_{W} and C_{H} are width and height of the search region respectively.

In example (a) of Table 3.1, A_{F} is about 8049 pixel^2 and A_{C} about 23785 pixel^2, which respectively accounts for 7.9% and 23% of the whole image area.

The area of the search region A_{C} and the area of the face region A_{F} can be therefore represented by the following equation:

$$\begin{aligned} A_{\mathrm{C}} &= 4.5E_{\mathrm{D}} \times 5.5E_{\mathrm{D}} = 24.75E_{\mathrm{D}}^2 \\ A_{\mathrm{F}} &= 2.5E_{\mathrm{D}} \times 3.5E_{\mathrm{D}} = 8.75E_{\mathrm{D}}^2 \\ A_{\mathrm{C}} &\approx 2.8A_{\mathrm{F}} \end{aligned} \tag{3.6}$$

Empirically, when a person's face is not purposely moving extremely fast, the search region is not only suitable for a 20 fps or a higher speed, but also working well with 10-20 fps speed.

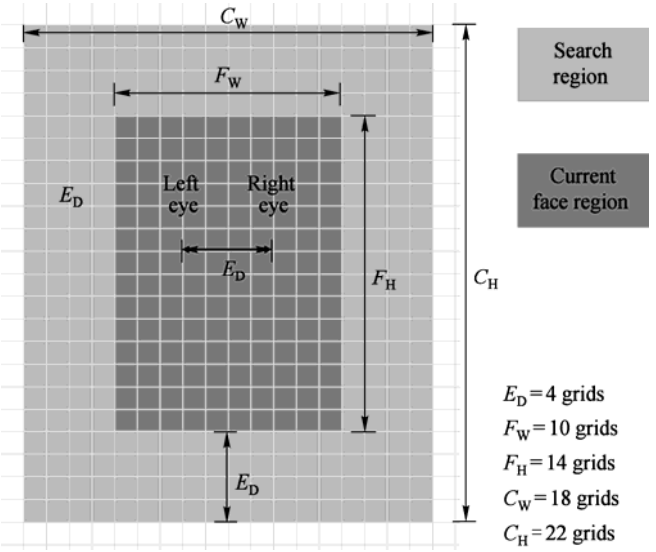


Fig. 3.7 Proportional illustration of face region and search region

3.3.3 Analysis of Temporal Changes

Once a certain search region is determined, we can apply the temporal information inside this region to further examine whether an expected face is still there. The design of the temporal-based detection should be computationally simple to be fit for the real-time processing purpose. As shown in Fig. 3.5, motion detection is implemented by three functional blocks: image difference, motion pixels calculation and significant motion detection.

Suppose that I_N and I_{N-1} denote the search regions of successive frame images respectively. Image difference is implemented by:

$$I_D(x,y) = I_N(x,y) - I_{N-1}(x,y) \tag{3.7}$$

where $I_D(x, y)$ is the intensity change at each image coordinate (x, y), and $I_N(x, y)$ is the N th frame of a certain sequence.

The following equation is used to calculate the motion pixels:

$$
\begin{aligned}
I_{\mathrm{M}}(x,y) &= \begin{cases} 1, & if\ I_{\mathrm{D}}(x,y) \geq T_{\mathrm{M}} \\ 0, & if\ I_{\mathrm{D}}(x,y) < T_{\mathrm{M}} \end{cases} \\
N_{\mathrm{D}} &= \sum I_{\mathrm{M}}(x,y) \\
M_{\mathrm{P}} &= \frac{N_{\mathrm{D}}}{N_{\mathrm{S}}} \times 100\%
\end{aligned}
\tag{3.8}
$$

$I_{\mathrm{M}}(x, y)$ is the binary moving pixels in the search region recalculated by a predefined motion threshold T_{M}, N_{D} is the number of motion pixels, N_{S} is the total number of pixels of the search region, and M_{P} is the motion parameter representing the percentage of motion pixels in the search region. A pixel is defined as a motion pixel when intensity change of a pixel is bigger than a predefined threshold.

There is a dilemma for the selection of T_{M}. On one hand, it should not be too small to be influenced by noise. Intensity changes due to noise can vary from one pixel to even more than 10 pixels. On the other hand, it makes no sense if too big value of T_{M} is chosen. With an extremely big T_{M}, motion cannot be detected at all.

To roughly estimate the range of T_{M}, we have recorded sequences to a PC with a TV card. The video sources are from TV news channels and from live video captured by a CMOS camera for comparison.

Fig. 3.8 shows the example images of the experiment. There are three image pairs with (a1)-(a2) and (b1)-(b2) from camera and (c1)-(c2) from TV. (a1)-(a2) pair is with relatively fast motion, (b1)-(b2) pair and (c1)-(c2) pair are with slow motion. The search regions and face regions are set by the first frame images, as shown in (a1) and (b1).

(a1) (a2)

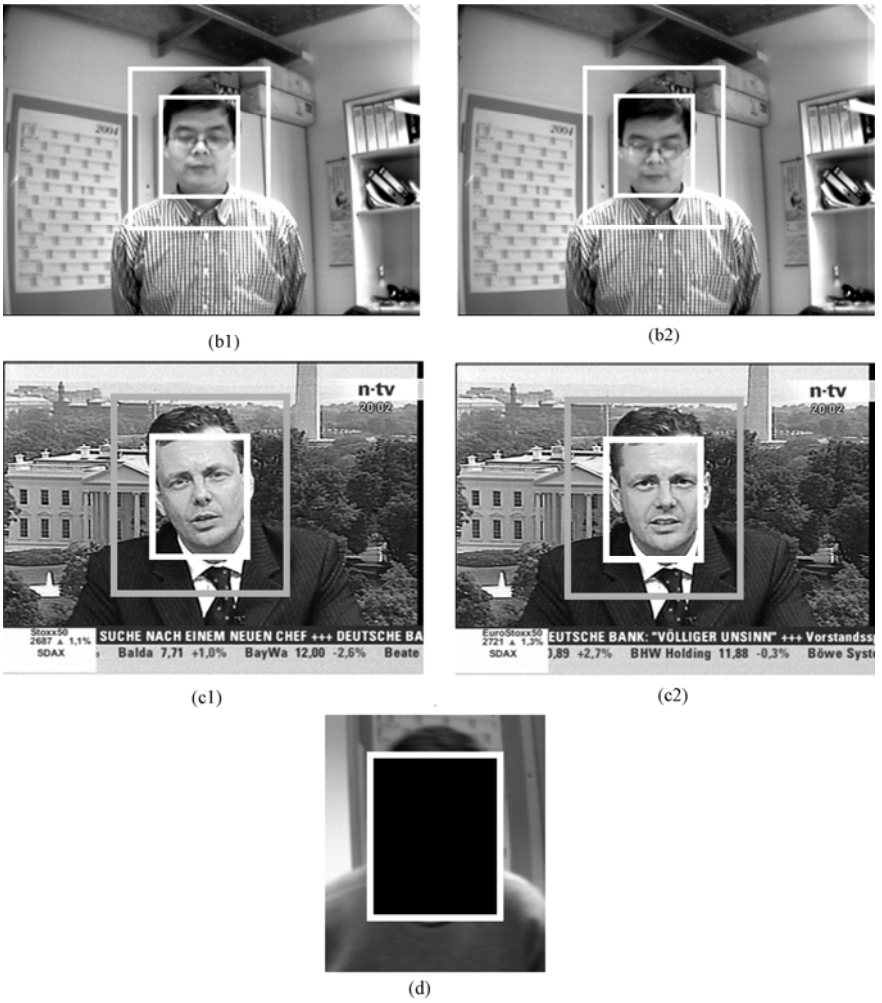


Fig. 3.8 Example images to find the range of T_M

To examine the lower range of T_M, we are interested in the motion pixels in the part of the search region excluding the face region, which is the area between the white frame and the grey frame in Fig. 3.8. That is because any facial expression change and face movement inside the face region can affect the analysis of the noise influence. We marked every face region with a black block as shown in Fig. 3.8(d) for the convenience of calculation. The respective percentage of motion pixels varied with T_M for each three pair is illustrated in Fig. 3.9. We can estimate the motion pixels from Fig. 3.8. Around 15% of the search region in (a1)-(a2) pair is in motion while the (b1)-(b2) and (c1)-(c2) pairs have

hardly moving pixels (less than 5%). We accordingly take the range of 15~35 pixels for T_M for further testing.

Regarding the decision of significant motion, the critical parameter is the threshold of the motion pixels M_P, which we define as M_{th}. If M_P is below M_{th}, the same face is expected to be remained in the search region, independent from the image-based face detector result.

The area of the face region is about 1/3 of the search region according to Equation (3.6). When there is a sudden change in the face region, there are roughly 30%~40% pixels of the search region in motion.

Adapted from (3.8) together with the above parameter discussions, we summarize in (3.9) the conditions under which the detection of a certain face is verified by the Motion-based Face Detector (MFD).

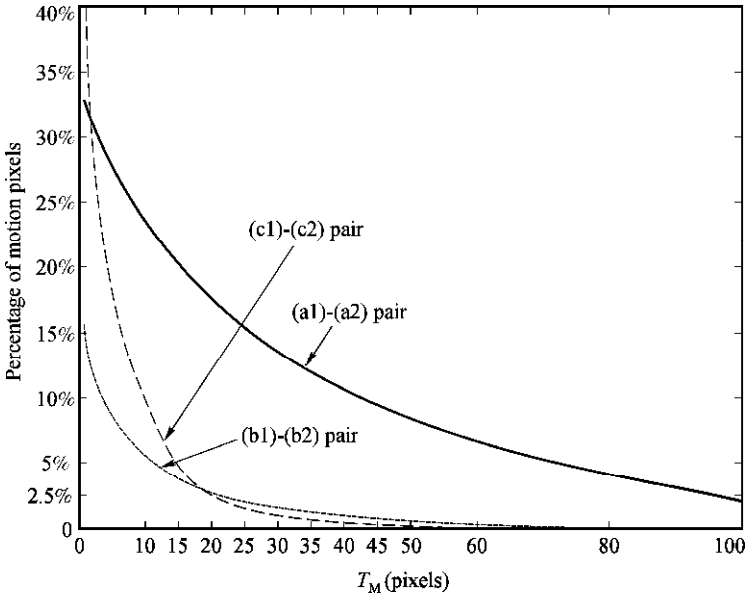


Fig. 3.9 Percent of the motion pixels in the search region (face region excluded)

$$I_M(x,y) = \begin{cases} 1, & \text{if } I_D(x,y) \geq T_M \\ 0, & \text{if } I_D(x,y) < T_M \end{cases}, \text{with } T_M \in (15,35)$$

$$M_P = \frac{N_D}{N_S} \times 100\%, \text{ with } N_D = \sum I_M(x,y) \tag{3.9}$$

$$\text{Face verified by MFD} = \begin{cases} \text{True}, & \text{if } M_P \leq M_{th} \\ \text{False}, & \text{if } M_P > M_{th} \end{cases}, \text{ with } M_{th} \in (30\%, 40\%)$$

3.4 Summary

The fusion of IFD and MFD guarantees that the detection procedure benefits from both the image-processing advantage in each single frame and the temporal context in video sequence. The automatic procedure is accordingly achieved without any assistance outside the system. This procedure is the first step to assure the whole autonomous recognition system.

Moreover, the combination of the multiple face detection methods significantly improves the detection rate especially for the critical cases like head pose variations and facial expression changes. In Section 8.2, we will elaborate the detection performance analysis. The two main benefits of improving the face detection false negative rate are deserved to be emphasized here:

- The failure of the face detection step leads to the failure of the whole system. No recognition classifier can work without face(s) detected. It is therefore always beneficial to have one system which features as low as possible the false negative face detection rate.
- The detector can also work as a face tracker which is contributed to the same face decision algorithms for face recognition. We will explore the tracking advantages in Section 4.4.

The face detection false positive rate is not so harmful. If a non-face object is detected as a face, there are still possibilities to remove it. Section 4.3.3 provides such a solution.

3.5 Further Discussions

The IFD can only output the most salient face for each image and therefore is not well in detecting multiple faces simultaneously. Since each search region of our proposed MFD is independently calculated from each corresponding face, the proposed temporal-based method has no trouble in handling multiple faces when a multiple IFD is available. Typical examples of multi-face sequences are shown in Fig. 3.10.

To detect multiple faces with the same detection quality as the single face detection with the specific IFD that we are using, the following idea is proposed, as illustrated in Fig. 3.11. This method is a workaround for the specific

IFD. Therefore, we do not intend to discuss it in detail here.

For future research, it is proposed to apply other IFDs including multi-IFDs to test the generality of our approaches.

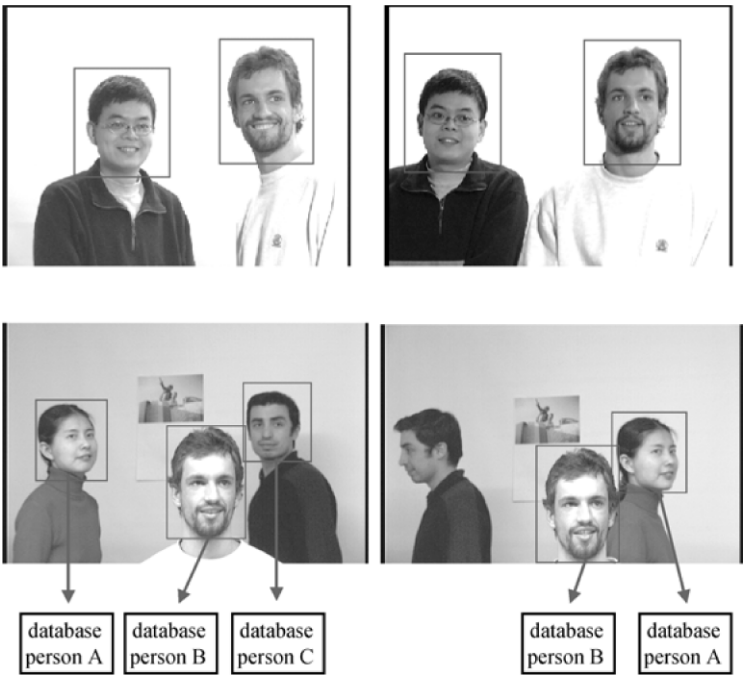


Fig. 3.10 Sequences of multiple people

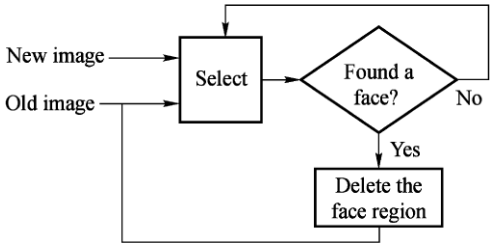


Fig. 3.11 Proposal for the multiple IFD

References

[1] B. Raytchev, H. Murase: Unsupervised Face Recognition by Associative Chaining. *Pattern Recognition*, Vol. 36, No. 1, 2003, pp.245−257

[2] J. Steffens, E. Elagin, and H. Neven: PersonSpotter-Fast and Robust System for Human Detection, Tracking, and Recognition. *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp.516−521

[3] T. kim, S. Lee, *et al.*: Integrated approach of multiple face detection for video surveillance. *Proceedings of IEEE 16th Conference on Pattern Recognition*, Vol. 2, 2002, pp.394−397

[4] D. Butler, C. McCool, *et al.*: Robust Face Localisation Using Motion. Colour & Fusion. *Proceedings of Digital Image Computing: Techniques and Applications (DICTA 2003)*, 2003, pp.899−908

[5] B. Stenger, V. Ramesh, *et al.*: Topology Free Hidden Markov Models: Application to Background Modeling. *IEEE 8th Conference on Computer Vision (ICCV 2001)*, Vol.1, 2001, pp.294−301

[6] P. J. Phillips, P. Grother, *et al.*: FRVT 2002 Evaluation Report, *Technical Report*, Mar. 2003, http://www.frvt.org, accessed 29 Mar 2007

[7] M. Jones, P. Viola: Fast Multi-view Face Detection. *Mitsubishi Electric Research Laboratories Technical Reports*, TR2003-96, 2003, http://www. merl.com/ reports/ docs/ TR2003-96.pdf, accessed 12 October 2005

[8] Z. Zhang, L. Zhu, S. Li, H. Zhang : Real-Time Multi-View Face Detection. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, May 2002, pp.149−154

[9] http://www.cognitec-systems.de/brochures/FaceVACSalgorithms.pdf, accessed 12 October 2005

[10] http://www.portrait-artist.org/face/structure4.html, accessed 12 December 2006

[11] S. Gong, S. McKenna, and A. Psarrou: *Dynamic Vision: From Images to Face Recognition*, London, Imperial College Press, 2000

[12] N. A. Dodgson: Variation and extrema of human interpupillary distance. *Stereoscopic Displays and Applications XI*, May 21, 2004, pp.36-46

[13] P. Viola and M. Jones: Robust Real-time Object Detection. *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, July 13, 2001

[14] *Human Engineering Design Data Digest*, Department of Defense Human Factors Engineering Technical Advisory Group (DOD HFE TAG), April 2000, http://hfetag.com/hfs_docs.html, accessed 15 December 2006

[15] N. A. Dodgson: Variation and extrema of human interpupillary distance. *Stereoscopic Displays and Applications XI*, May 21, 2004, pp.36−46

4 Automatic Face Recognition

Abstract In this chapter, we mainly focus on the discussion of unsupervised face recognition, which corresponds to the matching procedure and the classification-related sub-steps of the enrollment procedure, including same face decision, feature extraction and encoding. If each single step is running passively, the autonomous procedure of recognition is guaranteed. After the introduction section, feature extraction and encoding methods are briefly covered in section 2. The matching algorithms are described in detail in section 3, which includes an image-based classification method, an adaptive way of selecting the thresholds for the classifier, and a video-based temporal filtering method. In section 4, the combined same face decision algorithms are discussed. Finally is the summary of the automatic and intelligent face recognition procedure.

4.1 Overview

When faces are detected in a certain video sequence, the next two roles for an automatic face recognition procedure are: to determine the identification of the person and to determine whether to put it into the database. Identification, or termed as recognition, means to identify whether a certain live face is a known face or an unknown face. As mentioned in Section 2.7.4, we classify the face recognition task in a broad sense into three procedures: enrollment, matching (classification) and update. In this chapter, we mainly focus on recognition, which corresponds to the matching procedure and the classification-related

sub-steps of the enrollment procedure, including same face decision, feature extraction and encoding. The database-related sub-steps such as mugshot selection, database construction as well as the update procedure will be discussed in the next chapter.

4.2 Feature Extraction and Encoding

To determine the identification of a certain live face, feature extraction and encoding have to be applied before comparison. Any robust method could be suitable for this purpose. Here we still use the techniques from FaceVACS, the same commercial product as we have used for image-based face detection. In the following, we will briefly introduce the techniques according to [1].

A face image is firstly divided into an $M \times N$ block, as shown in Fig. 4.1(a). Feature extraction is then performed in each of these blocks. The extraction method is based on local abstract characteristics according to our category method discussed in Section 2.7.4. Local information in each block is extracted and transformed into a vector. As a commercial product, the specific applied encoding (transformation) method is not open to the public. Any of the mature techniques could be used, such as PCA, ICA, LDA, DCT, WT, etc. Every local area is transformed and the amplitude values in the frequency domain can be taken as local feature information. For a certain face image, an $M \times N$ number of vectors are accordingly determined to make a vector set, as illustrated in Fig.4.1(b). A global transformation is further applied to the vector set to construct a new feature vector, which is expected to represent the face image efficiently. In this way, each enrolled face shot of a certain person is encoded as one feature vector. Fig. 4.1(b), (c) show the global transformation procedure from a vector set to a feature vector. The parameter of the global transformation should be selected to achieve the maximum ratio of the inter-person variation to the intra-person variation.

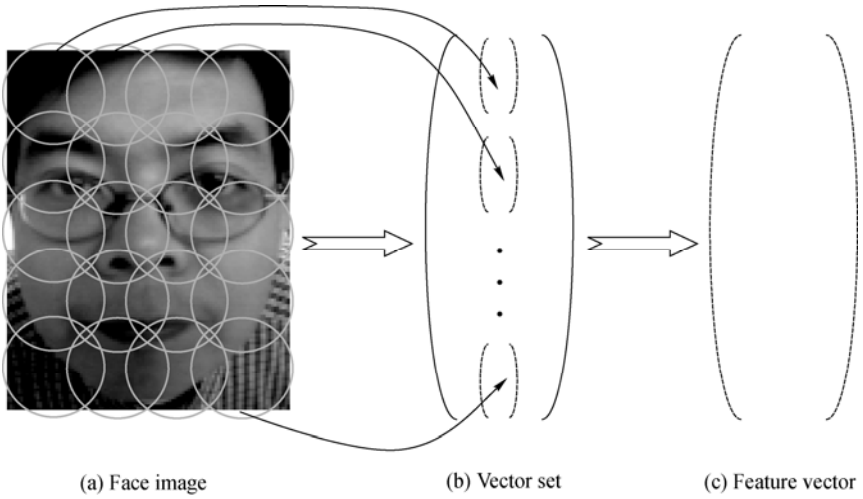


Fig. 4.1 Feature extraction of a face region

4.3 Matching/Classification

4.3.1 Image-based Classifier

When the features are encoded, they are to be arranged in a certain structure as the database construction, which is normally based on the clustering methods if no human teacher is available. How to build an optimum face database for high quality recognition will be discussed in more detail in Chapter 5.

Suppose that face databases already available, the next step after feature extraction is matching. Any classification method as discussed in Section 2.7.4 can be applied. How to select a well classifier is quite a critical point. The most important characteristic of a classifier might be the distance measure method. The intra-class and inter-class similarity are represented by distance measure. As mentioned earlier, the FaceVACS classifier seems to be a good candidate since it performs the best during the FRVT 2002 test [2]. However,

the FRVT test only takes each classifier as a black box, and having only made high-level performance analysis, such as the influence of the recognition rate by lighting change, scale variation, aging, etc. It does not indicate which distance measure algorithm is used in FaceVACS since the algorithm is not open to the public. Therefore, it seems to be really necessary to analyze the distance measure performance through more experiments.

Among many distance measure methods, minimum distance and Bayesian decision are more frequently applied in face recognition. Fig. 4.2 sketches the differences between them. As shown in Fig. 4.2(a), while minimum distance method only measures the Euclidean distances, Bayesian distance additionally considers the probability distributions of each class. Fig. 4.2(b) visually demonstrates the classification result by using the two methods. The shaded circle, which in reality belongs to Class A, will be wrongly classified to class B if minimum distance method is applied, while the classification is correct if Bayesian distance is applied.

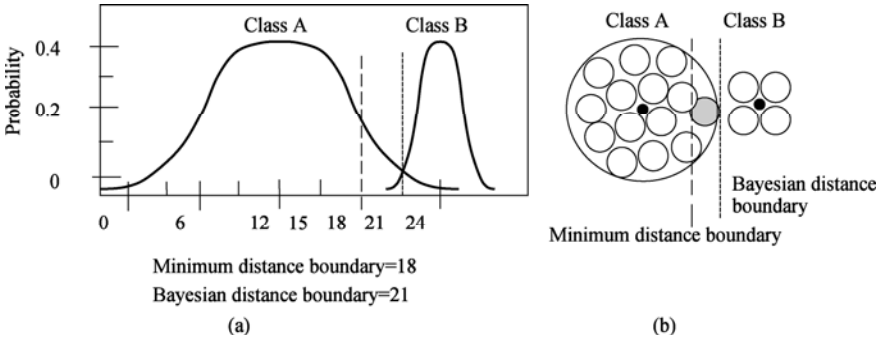


Fig. 4.2 Comparison between Bayesian distance and center distance

For testing purposes, we have chosen 10 pairs of classes. Each pair contains two classes, and each class represents the face database of one person. The cross similarity between both classes in each pair is chosen to be relatively high, meaning that each pair contains two similar-looking people. To simulate the case shown in Fig. 4.2(b), one class is set to be significantly larger (9 face shots) than the other (4 face shots). The Euclidean distance in each class is calculated and a new face shot (corresponding to the shaded ball) is chosen according to the following conditions:

- The new face shot in reality belongs to class A.
- The new face shot is classified to be class B according to Euclidian distance measure.

The same procedure runs for the FaceVACS classifier. The Euclidian distance method fails for all the 10 pairs of classes while the FaceVACS classifier successes for 8 pairs. Hence, we can then conclude that the classifier is very likely to be a Bayesian classifier.

Now we can take the classifier as a black box and only examine input/output values and important threshold values.

The classifier accepts a detected face shot as an input. It compares the face image with the existing databases. For each comparison, there is a similarity value assigned to be linked to the corresponding face database. The similarity value is denoted by S_v, which lies between 0 and 1. 0 means no similarity at all between the current face shot and a database, and 1 means there is 100% the same element found in a certain database. S_v can be also explained as the probability of being identified as a certain person. Intuitively, $S_v = 50\%$ can be taken as the threshold for identification. Beyond the threshold, it is identified as a known face; otherwise, it is probably an unknown face. However, the choice of the threshold highly influences two importance parameters: false acceptance rate (FAR) and false rejection rate (FRR). Their relationship is clearly demonstrated in the FAR/FRR curves in Fig. 4.3. The curves are obtained by taking face images under varying lighting conditions, different poses and facial expressions. Consequently, they represent the typical indoor cases. As shown in the figure, the bigger the similarity threshold, the higher FRR and the lower FAR. At the point of similarity threshold equaling 0.5, FAR and FRR have the same error rate. But $S_v = 50\%$ is not preferred here since FAR is much more harmful for the automatic procedure. With only one wrong face shot from another person, a certain database may continuously enroll more erroneous mugshots from the same wrong person. Hence, the threshold should be chosen that a low enough FAR (e.g. smaller than 0.1%) is achieved.

For any other given chassifier, the similarity threshold can be accordingly selected from the typical FAR/FRR curves. Such curves can also be obtained through experiments if they are not available.

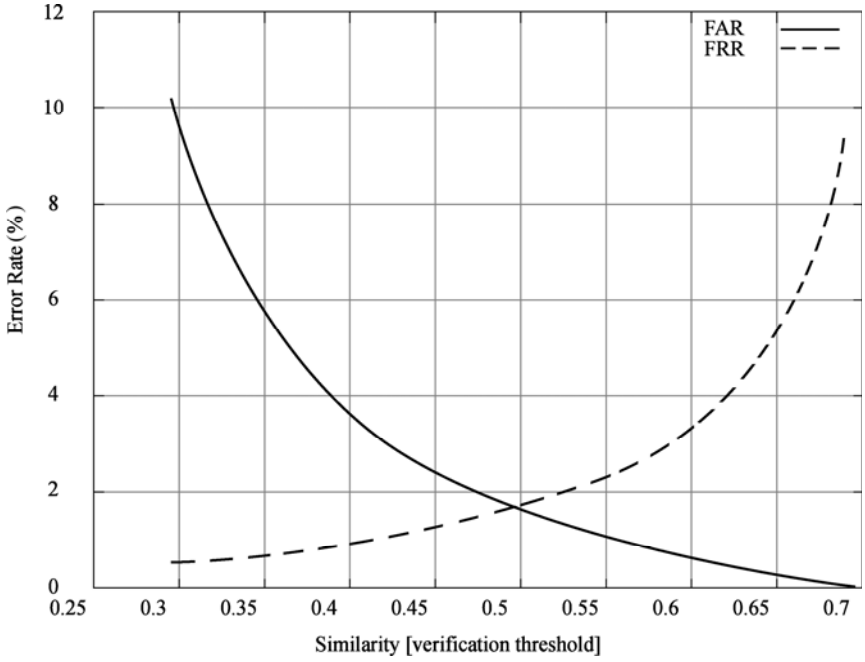


Fig. 4.3 FAR/FRR curves

(Derived from the FaceVACS-SDK Reference Manual, version 1.9, pp.166 with cosmetic changes)

4.3.2 Adaptive Similarity Threshold

In the above, we have discussed the question of choosing a prior value of S_v which can keep the low FAR. As a penalty, the corresponding FRR is unavoidably high. Since a person's database is incrementally constructed, the FRR is too big to be tolerated especially at the beginning of a newly built database. The comparison is shown in Fig. 4.4. FRR16 means that the database is enrolled with 16 different face images of a same person, and FRR1 denotes that only one image is enrolled in the database. Let us assume that the similarity threshold is selected to equal 0.65 as a prior value so that FAR is below 0.1%. In this case, the difference between FAR1 and FAR16 is so tiny that it can be neglected. However, FRR1 and FRR16 are obviously different. FRR16 approximately equals to 1.5% while FRR1 approximates 6.5%. Therefore, we

introduce an adaptive similarity threshold (AST), denoted by Equation (4.1).

$$AST = S_{V0} + a \cdot i, \quad (i = 0, 1, 2, \cdots, N_{\max}, 0 < a < 1) \tag{4.1}$$

where S_{V0} is the minimum value of AST and can be predetermined by the FAR/FRR curve. i denotes the actual number of enrolled face images for a certain person. a is the weight which is smaller than 1. $N_{\max}$ represents the maximum number of enrolled face images for a certain person.

From the above equation, it is obvious that AST should be assigned lower with fewer images enrolled into a database and higher with more enrollments. In the above example, 0.65 is corresponding to the maximum value of AST which equals to $S_{V0} + a \cdot N_{\max}$, where $a \cdot N_{\max}$ can be set to 0.1 such that the minimum number of AST (S_{V0}) equals to 0.55.

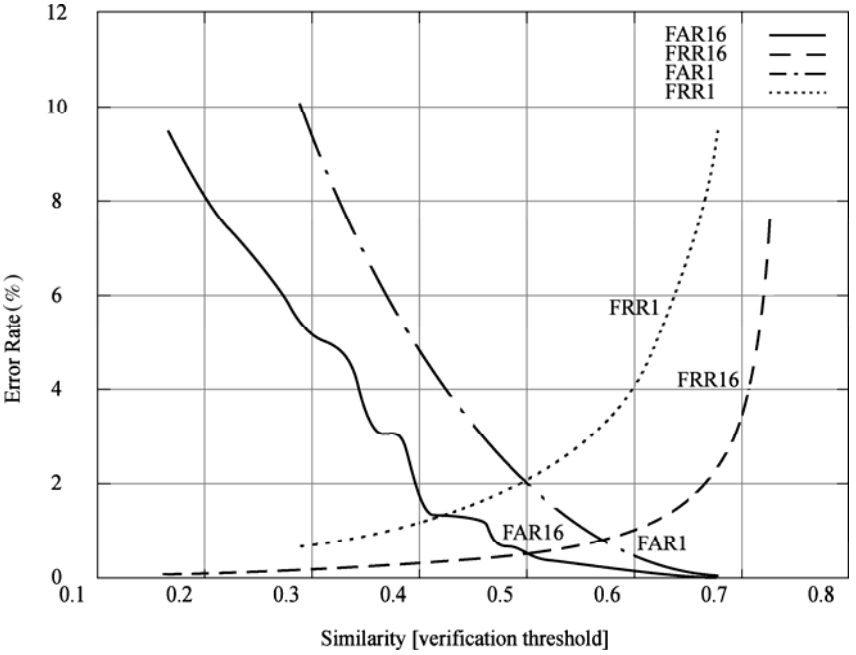


Fig. 4.4 FAR/FRR curves—enrollment from 1 vs. 16

(Derived from the FaceVACS-SDK Reference Manual, with cosmetic changes)

It is to be noted that although the introduction of AST is analyzed based on the FaceVACS classifier, the theoretical fundamentals for designing AST remain the same if we apply another classifier.

4.3.3 Temporal Filtering

Up to now, all the discussions regarding matching assume that only one single frame image is used for comparison. To further decrease the probability of making wrong decisions, taking more frame images is helpful. Thanks to the video context, we could use the valuable temporal information from video sequences. A temporal filter is consequently designed in Equation (4.2):

$$\sum_{i=1}^{n} A_i \;\; S_{v,i} > m \quad AST \tag{4.2}$$

where A_i represents the coefficient with the range between 0 and 1, n is the filter length, m is a factor and can be intuitively approximated to n when each A_i equals to 1. $S_{v,\,i}$ denotes the respective similarity value of the ith frame image compared with a corresponding database.

Applying this filter, the classifier identifies a certain person only if the above condition is fulfilled, instead of using only one image.

Obviously, the filter has a low-pass characteristic, which can deal with sporadically as well as frequently occurring errors. Such typical errors might include:

- False positives from the face detector.
- Falsely accepted frames from the matching procedure.

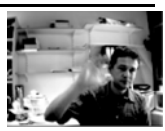
As discussed earlier, there are two categories of false positives in face detection. One is due to some static face-like objects and the other might mainly come from the motion effects. Fig. 4.5 demonstrates such two examples. As shown in Fig. 4.5(a), the region around the monitor is falsely detected as a face. We marked the region in dark curve to show its dimension. But this frame is the only frame out of a 100-frame sequence which leads to the false positive detection. If no filter is applied, the monitor will be saved as a new face. In Fig. 4.5(b), we can observe the motion effects which lead to a wrong face detection in the marked region. Similarly, it is one of the two frames that output the wrong face region in a 100-frame sequence. Obviously, the wrong cases can be filtered by the temporal filter.

(a) False face detection due to a static face-like object

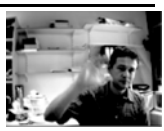
(b) False face detection due to motion

Fig. 4.5 Two examples of false positives in face detection

Table 4.1 Similarity comparison of two similar persons

Person 1's frame images		Person 2's database 
Person 1–image1		$S_{v,1} = 0.72$
Person 1–image2		$S_{v,2} = 0.59$
Person 1–image3		$S_{v,3} = 0.65$
Person 1–image4		$S_{v,4} = 0.51$
Person 1–image5		$S_{v,5} = 0.11$
Person 1–image6		$S_{v,6} = 0.08$
Person 1–image7		$S_{v,7} = 0.65$

Continued

Person 1's frame images		Person 2's database 
Person 1–image8		$S_{v,\,8} = 0.39$
Person 1–image9		$S_{v,\,9} = 0.53$
Person 1–image10		$S_{v,\,10} = 0.50$
Average		$S_{v,\,average} = 0.43$

Regarding the improvement of false acceptance rate, let's take a look at another example. Table 4.1 shows the case of comparing two persons which could confuse the classifier when only one frame is used. It can be seen that there are two face images from person 1 which have pretty high similarity value when compared to some classes of person 2. Assume that *max(AST)* is set to be 0.65. Without the filter, the two persons will be wrongly merged into one database. With a simple average calculation, as long as the filter length is not small, they can be clearly distinguished although they are similar to each other.

4.4 Combined Same Face Decision Algorithms

To further improve the face identification quality, we apply the combined results from MFD (Section 3.3) and from the temporal filter. The sequence context is also included to contribute to same face decision.

As mentioned in Section 3.3, the temporal-based face detector provides the temporal information of a certain detected face in video. Therefore, it behaves

like a face tracker and can be applied as the first processing step to determine a same face. Intuitively, in most cases, when motion pixels are much less than stable pixels in a certain face region, it can be decided to be in the same face state. But the condition may not be fulfilled in the following situations:

- Slow moved faces with occlusions
- Sequence shot change

Fig. 4.6 demonstrates an example of the face occlusion. Every second frame of a 16-frame image sequence is shown in Fig. 4.6 (a)-(h). In the sequence, person A (colored in gray) is rolling his head and person B keeps stationary. In (a), face B is almost completely occluded by face A and in (h), person B is showing up in the same face region as person A. Since the movement is not significant from frame to frame, the temporal-based face detector always assumes that person A is in front of the camera even though in (h), which is a critical failure.

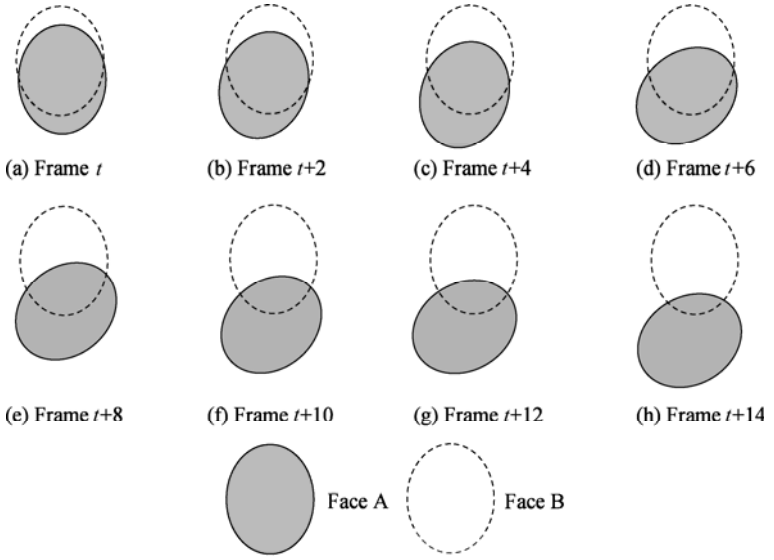


Fig. 4.6 An example of face occlusion with slow motion

For a sudden shot change where the previous face region remains hardly change, similar failure could be made. For example, in TV news, it often occurs that the scene is suddenly changing from one reporter to another with similar background. In that case, the face region is very likely to be detected with no significant motion.

Therefore, the same face decision algorithms cannot only dependent on the output of face tracker. In addition, the filtered output from the face recognition classifier and the output from the image-based face detector are used. For simplicity, we suppose that each of the three components has binary output only, described as follows:

- IFD

$R_{IFD}= 1 \Rightarrow$ A face is detected by the image-based face detector;

$R_{IFD}= 0 \Rightarrow$ No face is detected by IFD.

- Face tracker

$R_{FT}= 1 \Rightarrow$ Small motion is detected in the previous face region, indicating that the face tracker considers the current face to be the same face as existing in preceding frames.

$R_{FT}= 0 \Rightarrow$ Big motion is detected in the previous face region, indicating that the face tracker considers that there is a change in the previous face region.

- Temporal filter

It provides the temporal filtered results from the image-based recognition classifier. The temporal filter has binary results according to Equation (4.2): $\sum_{i=1}^{n} A_i \cdot S_{v,i}$ is either below the similarity threshold $m \cdot AST$ or beyond it. It is the decision from both the image-based recognition classifier and the filter. When beyond the threshold, the current face is very likely to be the same face; otherwise it is likely to be a different one.

$R_{TF} = 1 \Rightarrow$ current detected face is identified as the same face by the filter.

$R_{TF} = 0 \Rightarrow$ if $R_{IFD} = 1$, current detected face is identified as a different face by the filter; otherwise, there is no filtered output, i.e. the image-based recognizer fails to output any S_v. It happens when the extracted face region from the face detector has so low quality that the recognizer cannot take it as a real face shot.

Majority voting is an intuitive way to make the decision out of the above three outputs. That is to say, if two of the parameters vote for "same face", we can finally judge that it's "same face". However, it is much more complicated in reality and the majority voting is subject to lead to failures. For example, when a person's head is turning so that only profile is shown to the system, the detector failures ($R_{IFD} = 0$). Due to small motion, R_{TF} equals to 1. R_{IFD} is 0 in this case. The majority voting decides that it's not a "same face" case. But it would be better to hold on and further examine more following frames rather than deciding that no face exists. Another example is as shown in Fig.4.6, where $R_{IFD} = 1$, R_{FT}

= 1 and $R_{TF} = 0$. Although the majority votes for "same face", it is not true. The result in this case is much more harmful since two persons are to be enrolled into one database. The errors cannot be corrected without any human help.

Therefore, we list in Table 4.2 all the possible combinations of the three parameters to explore all cases. The last two columns in the table are theoretical combinations which do not exist in reality. $R_{IFD} = 0$ means that no face is detected from IFD. As a result, no features are extracted from the current frame image and the recognition filter couldn't recognize it. Consequently, we discuss in the following the other six cases where further video context analysis is applied under some conditions.

Table 4.2 List of all cases for combined same face decision algorithms

Case categories	I	II	III	IV	V	VI		
R_{IFD}	0	1	1	1	0	1	0	0
R_{FT}	0	0	0	1	1	1	0	1
R_{TF}	0	0	1	1	0	0	1	1
Same face	No	No	Yes	Yes	Yes/No	Yes/No	No decision (such cases are not existing)	

Case I $R_{IFD} = 0$, $R_{FT} = 0$, $R_{TF} = 0$. As all the three parameters indicate, there is no face detected and recognized in the current frame. Since a big motion is detected, it might indicate that a new sequence is starting or a sudden lighting change is occurring, etc.

Case II $R_{IFD} = 1$, $R_{FT} = 0$, $R_{TF} = 0$. A face is detected by IFD but not identified by the classifier. It is determined to be not a "same face". However, quite different from case I, this case is a temporal case and can transit into either Case V or Case III or Case VI depending on the future filtered classifier outputs. Since big motion is detected, there mainly exist two possibilities: one is a false positive error in IFD; the other is that a new face might be showing up. If R_{TF} keeps 0 for a certain number of frames, this case becomes Case V. Otherwise, when R_{TF} becomes 1, a new face might start and the condition of Case III or Case VI is fulfilled.

Case III $R_{IFD} = 1$, $R_{FT} = 0$, $R_{TF} = 1$. Although there is a big motion detected, it's still safe enough to use the majority voting in this case to judge "same face". In reality, it might correspond to a rapid head motion or a lighting condition change.

Case IV $R_{IFD} = 1$, $R_{FT} = 1$, $R_{TF} = 1$. As all the three parameters agree with each other, there is no doubt that "same face" is guaranteed. This is a stable state that "same face" is recognized.

Case V $R_{IFD} = 0$, $R_{FT} = 1$, $R_{TF} = 0$. The "same face" decision can be either yes or no, depending on the previous cases. We name the states relationship analysis as video context analysis.

1) When Case I or Case II is the preceding case, the current frame is determined as a static blank background with no face detected.

2) When the current state is transited from Case III, it can be considered as a temporal state where the current face is only out of tracking but still existing. For example, when a face is rotated to profile or is rolled very much, IFD could fail to detect it, small motion condition in face region is fulfilled, and the classifier couldn't recognized it. But it would be better to hold on and further examine more following frames rather than deciding that no face exists. In realistic cases, the profile head is very likely to turn back to his frontal pose.

3) This case can also be directly transited from Case IV. Let's take the example of scale changes shown in Fig.8.7. When a person is slowly leaving from the camera or approaching the camera, there is an upper resolution threshold and lower resolution threshold where IFD couldn't detect the face. Therefore, there might occur a sudden change from Case IV to Case V. The ideal process is to take this situation as "same face".

Case VI $R_{IFD} = 1$, $R_{FT} = 1$, $R_{TF} = 0$. The decision of "yes" or "no" under this condition is also relied on video context analysis.

1) When the previous state was Case II or Case V, it could be a new face showing up, the decision is therefore "no".

2) When the previous state was Case IV, it's held on to search the following several frames. One typical case in reality could be like the following: a person keeps the same, but the filter could not recognize it due to some special head poses or expressions. The decision should be "yes" in this case. FRR (false rejection rate) is accordingly improved by this way.

3) When it has been in Case VI for a certain number of frames, a new person could show up. A typical and real example for this situation is as shown in Fig. 4.6. The decision is therefore "no".

The challenge is to realize the above mentioned state transitions. We have designed hierarchical state machines to deal with it, which will be described in more detail in Chapter 6.

4.5 Summary

As pointed out in Section 2.8, the essence of human brain intelligence regarding face recognition shall be represented by the following abilities: being able to combine all useful information through analysis; learning and running in an unsupervised, completely automatic, and non-invasive way.

From the above discussions in this chapter, we are now ready to summarize how our proposals simulate the human intelligence. The image-based face classifier behaves like the basic function. Like human brains, the proposed system has the ability to be pruning through learning and experiences. The similarity threshold is gradually adapting through learning, which partly demonstrates the pruning process. The temporal filtering is actually utilizing the temporal information and video context for recognition as well as for detection. The combined same face decision algorithm demonstrates the intelligent analysis of all recognition/detection components, which utilizes all possible information including temporal information, video context, and logic analysis etc. for the final conclusion. Moreover, each step is running without any outside help or cooperation, and is therefore automatic, unsupervised and non-invasive.

In the next two chapters, the attempts of simulating human intelligence will be continued for state transition algorithms and for the database construction/update algorithms.

References

[1] http://www.cognitec-systems.de/brochures/FaceVACSalgorithms.pdf, accessed 12 October 2005

[2] P. J. Phillips, P. Grother, *et al.*: FRVT 2002 Evaluation Report, *Technical Report*, March 2003, http://www.frvt.org, accessed 29 Mar 2007

5 Unsupervised Face Database Construction

Abstract In this chapter, we mainly discuss the unsupervised and automatic way to build databases. Followed by the introduction section, the related machine learning background is introduced in section 2, including supervised as well as unsupervised learning methods, and the analysis of typical clustering structures. In section 3, a proposed data structure is described, where the fused clustering structure and its related parameters for building the face database are discussed. The structure proposal is corresponding to the database construction step of the enrollment procedure in Chapter 2.7.4. In section 4, we explore the important features to achieve an optimized database: the mugshot selection criteria and the update rules. This chapter ends by the summary section.

5.1 Introduction

We have so far discussed the combined face detection methods and most parts of the intelligent face recognition algorithms. According to the category of face recognition procedure defined in Chapter 2.7.4, the study of automatic database construction and adaptive database update are still missing.

As pointed out in chapter 2, in traditional face recognition systems, even in the most state-of-the-art automatic systems, database construction and update is considered to be a separate and supervised procedure. What we are expecting is however an automatic and unsupervised procedure. For this purpose, efficient database structure is highly demanding and pre-defined constructing and updating rules are required. In this chapter, we are mainly covering those issues.

5.2 Backgrounds for Constructing Face Databases

5.2.1 Supervised Learning

There are in general two major machine learning methods: supervised and unsupervised. Supervised learning attracts many research interests, among which inductive learning is the most popular. It generalizes from observed training examples by identifying features that empirically distinguish positive from negative training examples, which is defined in [1]. The prerequisite step is to construct classes based on training data. The training process relies on an external human teacher who has the whole knowledge of the training data while the machine learner does not. Both the teacher and the learner are exposed to the training data. The learner applies a certain kind of target function to classify each piece of the training data into an output, which might be either correct or wrong. The teacher is able to provide a desired target response of that piece of data. Accordingly, the teacher either verifies or corrects the output of the learner and feedbacks the informative results to the learner to adjust its target function. In other words, during training, labels are manually assigned to each individual observation of data, indicating the corresponding classes. The process is carried on iteratively through a large amount of data until the learner is expected to be as optimal as the teacher is. After the training step supervised by the teacher, the learner is expected to be able to group an observation of new data into the existing classes. Instance-based approach, decision trees, neural networks, and Bayesian classification and genetic algorithms are the popular examples of inductive learning. Detailed discussions of these traditional methods can be found in many machine learning textbooks such as [1]. For inductive learning, the training data is therefore a key to a successful classification. It should be abundant enough to completely represent all possible distributions of data. An optimum supervised learning method can build such a model based on the training data that each new test observation can be correctly grouped into a certain existing class. However, if the distribution of the data is unknown, it is difficult to determine how many observations of data are sufficient enough and which kinds of data are suitable for training.

Prior knowledge together with deductive reasoning is another way to deal with the information provided by the training examples. This branch of supervised learning is defined as analytical learning in [1]. Since prior knowledge is usually mathematically represented by a set of predefined rules which are not dependent on the training data, we also define this method as the rule-based learning. The rules are used to analyze each piece of training data in order to infer which one is more relevant to the target function and which one is not. In principle, with the help of the additional prior knowledge, this method is much looser conditioned by the training data than the inductive method. In the extreme case, pure analytical learning requires only a tiny amount of training data. It mimics the manner of human learning systems to a much larger extent and is theoretically more promising than inductive learning. However, the rule-based learning suffers a lot from choosing suitable predefined rules. Ideally, the learner's prior knowledge is assumed to be correct and complete, but in reality the assumption is often violated. In most practical systems, there are not plentiful training data available and no perfect predefined rules provided. The reciprocal combination of both prior knowledge and inductive analysis of training data is proven to be a better choice to overcome the shortcomings in both methods [1]: better generalization accuracy when predefined rules are provided and reliance on observed data when prior knowledge is only approximately correct and incomplete.

Another critical assumption in most supervised learning is the distribution of data is not changing from time to time, which is actually frequently occurring in real world. For our specific face recognition problem, it is exactly the case that faces are always looking different from time to time. To deal with such difficulties, update is required. The classes and target functions have to be accordingly adapted to the distribution of new coming data. For example, outdated classes or some outdated parts of a certain class are to be removed and new classes are to be created. Hence, the update is similar to the training process and should also be supervised by a teacher. This updatable learning is defined as reinforcement learning.

5.2.2 Unsupervised Learning

For unsupervised learning, the class labels are unknown since no external

teacher is available to help for the class labels. During training, the learner has to discover and generalize the distribution and characteristics of the observed data, making classifications of the data by itself. This kind of self-classification is also called clustering. A good clustering method is expected to produce high quality clusters with high intra-class similarity and low inter-class similarity. In other words, the distance between any pair of elements inside a class is small; and the distance between any two elements from different classes is big.

Fig. 5.1 lists some of the popular cluster distance measure methods. Minimum distance and maximum distance presentation are the two easiest ways to calculate the cluster distance. The distance between clusters is represented by the distance of the nearest elements from different classes. It is also defined as single linkage. For maximum method, the cluster distance is represented by the maximum distance between two elements, one from each cluster. It is also called complete linkage. Minimum distance is often used to calculate the intra-class similarity $min(d_{intra})$ and maximum distance is often for inter-class similarity $max(d_{inter})$. Ideally, for any given clusters, the $max(d_{inter})$ should be always smaller than $min(d_{intra})$. However, the assumption is too restricted for real world application. For face recognition, the feature vectors inside a cluster are expected to represent the encoded face shots from the same person with similar facial expressions, head poses and lighting conditions. When a live face is given, it is compared with all existing clusters and finds out the corresponding match group. Unfortunately, the similarity between faces from the same person with different views is often bigger than the faces from different persons with same poses and facial expressions. It therefore seems not possible to apply only the minimum/maximum distance methods to cluster the training sequence sets.

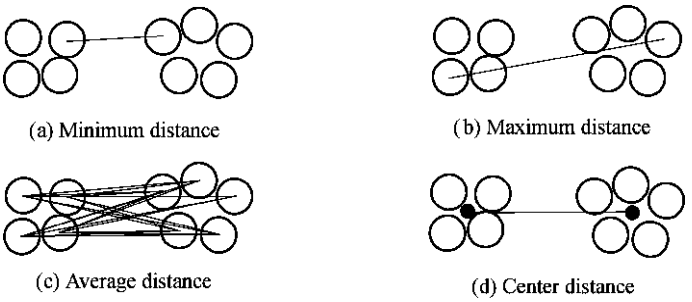


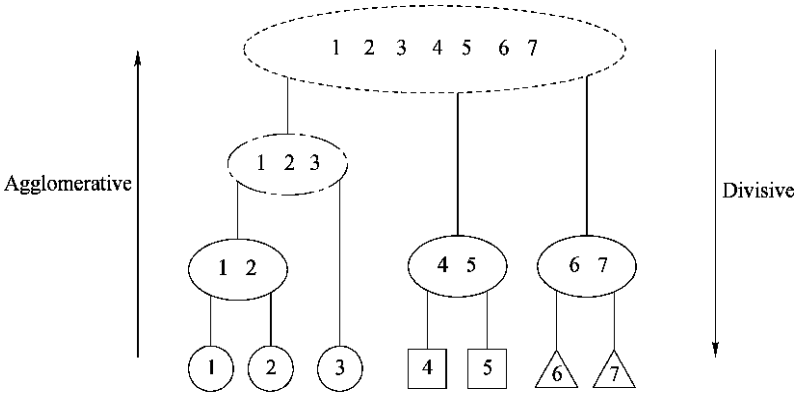
Fig. 5.1 Four different ways of distance measure

Average distance and center distance are the two better ways of representing distances. The average distance method calculates the average of all distances between each pair of the elements. For center method, centroid is calculated from all elements from one cluster. Cluster distance is then denoted by the corresponding centroid distance. Although precise, these two methods suffer from big computational efforts. When a new observation is added to a cluster as an element, the cluster distances related to this cluster have to be recalculated. As mentioned in Section 4.3, the Bayesian-like FaceVACS classifier is applied, which can be considered as an improved center distance method. In the following, we apply this representation to illustrate different clustering algorithms and structures.

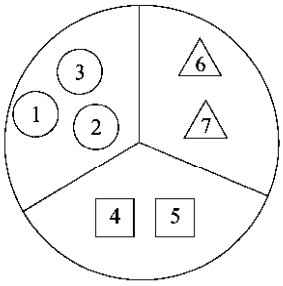
5.2.3 Clustering Analysis

There are in general two traditional clustering techniques: hierarchical algorithms and partitioning algorithms. In hierarchical method, the data are classified into clusters by different layers rather than only one hierarchy. The clusters therefore have a tree-like structure. The tree can be built by either a bottom-up order or a top-down order. The former case is defined as an agglomerative clustering and the latter one is defined as divisive clustering. Agglomerative clustering starts with each observation taken as one individual cluster. Clusters with the minimum distances are then merged into bigger classes. The merge is iteratively made until all data are finally combined as one single cluster. On the contrary, divisive clustering begins with one big class and gradually divides into smaller and smaller clusters. This process continues until each cluster contains only one observation. The two hierarchical clustering methods are illustrated in Fig. 5.2(a).

In partitioning method, a prescribed number k of classes is given. Training data are grouped into k classes according to their mutual distances. The grouping procedure is iteratively carried out until some partitioning criterion are fulfilled, which optimizes the homogeneity of a certain cluster. Fig. 5.2(b) illustrates the partitioning structure of the same observation data as in (a).



(a) Hierarchical clustering method, represented by a tree-like structure



(b) partitioning clustering method, represented by a pie-like structure

Fig. 5.2 Sketches of the two clustering methods

Since both agglomerative and divisive procedures are executed incrementally, the produced clusters contain less failure elements compared to the partitioning method. Another advantage is that the number of clusters does not need to be defined in advance. However, the hierarchical algorithms are of no use in their final procedure. All observations are finally combined into one single cluster for agglomerative method, and every observation is eventually taken as one cluster for divisive method. Therefore, the major challenge in hierarchical algorithm is to determine that at which hierarchy the clustering procedure has to be terminated. For partitioning, the data are grouped into parallel clusters. It is optimum if certain criteria are clearly defined to be appropriate for all data, as shown in the example in Fig.5.2(b). But the major challenge in this method is to determine the number of clusters and their boundaries. In real world, the probability to produce cluster overlaps for this method is unavoidably high. Once there are wrong clustered elements, the partition procedure might not converge

any more. That is to say, the overlapped boundary can be possibly gradually enlarged until two clusters are eventually wrongly merged. Additional efforts have to be made for compensation which is a difficult task.

5.3 Database Structure

5.3.1 A Fused Clustering Method

Fig. 5.3 shows an abstract example of applying clustering methods for constructing the face databases. There are three groups of data observations available as shown in Fig. 5.3(a), denoted by circles, squares and triangles, respectively. Each group represents a set of face images from one person. Every single observation is actually a face shot. Person 2 is somehow similar to person 3. The cluster distances between several face shots from person 2 and person 3 are quite small. In other words, these face shots from different persons are even more similar than an inter-person similarity. As we mentioned earlier, the case frequently occurs when two persons have similar face shapes, head poses, and facial expressions. Fig.5.4 shows a realistic example of such a case. Applying the FaceVACS technology for the comparison between person A and person B, the similarity of the Person A(1) − Person A(2) pair is only a little bigger than the similarity of the Person A(1) − Person B pair, even though the two images of person A are recorded within one second. Since the FaceVACS ranked the first in the FRVT 2002 test, the comparison result is quite representative.

Now we are looking back to the example in Fig.5.3. Partitioning method is advantageous over the pure hierarchical method since it can clearly cluster three persons. Fig.5.3(b) illustrates the partitioned structure. However, as can be seen from the figure, there is an obvious overlap between boundaries of person 2 and person 3. The overlap can be explained as the statistical number of false accepted face shots. The bigger the overlap, the higher the false acceptance rate for recognition. Since there is no teacher or supervisor, the overlap is too harmful to be tolerated.

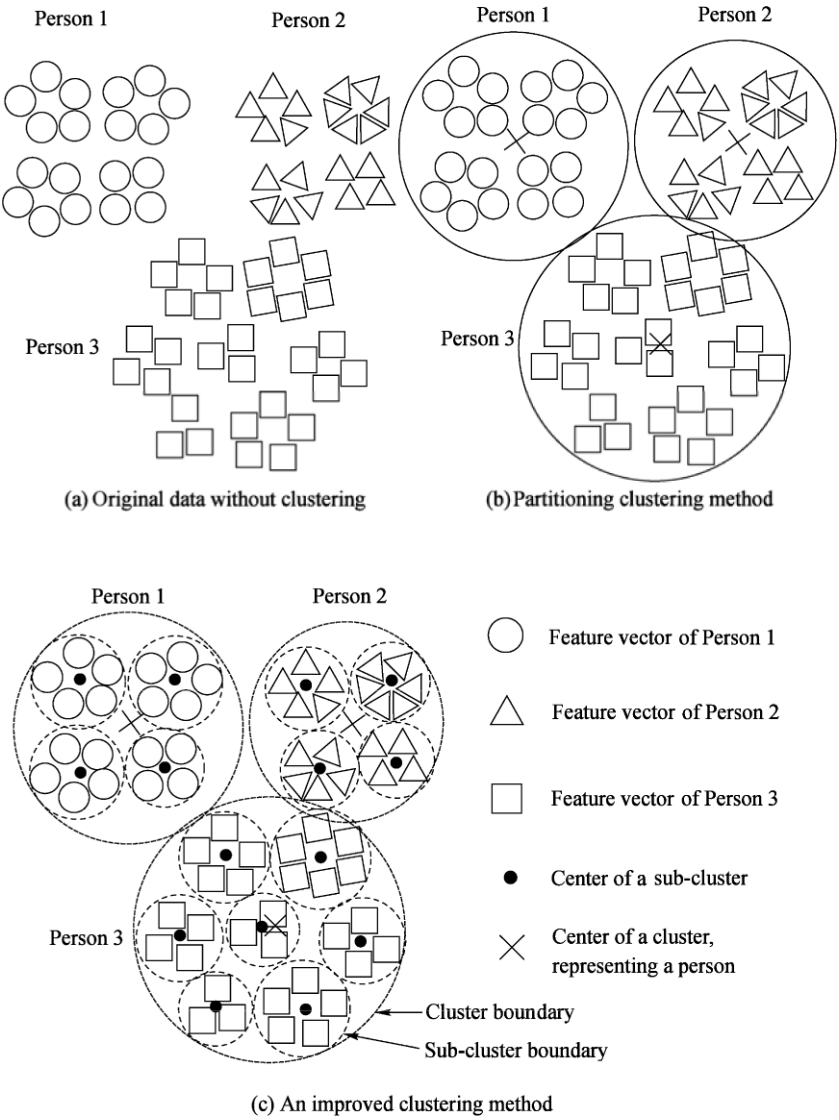


Fig. 5.3 Two structures of a face database

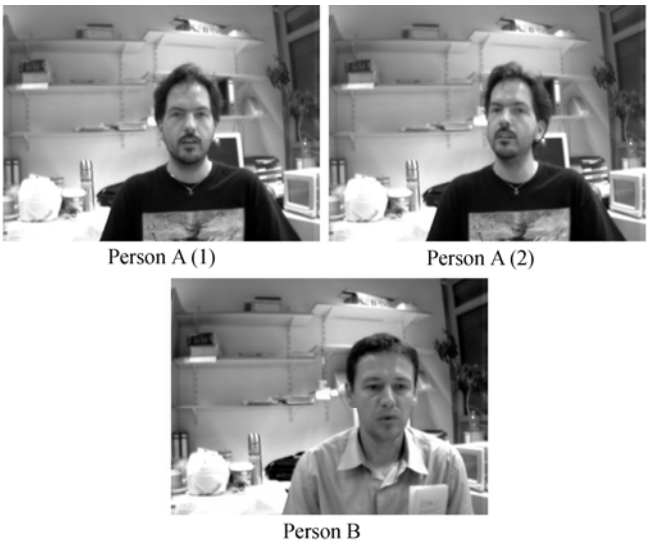


Fig. 5.4 An example of different people having a big similarity

To decrease the possibility of producing overlap, we propose an improved structure which combines the hierarchical and partitioning method, as shown in Fig.5.3(c). We call this method a fused clustering method. At the beginning of online data acquisition, there are few or even no observations. When a face is detected, its face shot is then taken as one observation. If the previous classifier decides that it is a new face, the observation is the only one at this moment. The following observation has to be compared with the first one to determine whether their similarity is big enough. Under this circumstance, the database can only be constructed by taking the bottom-up method. That is to say, the very few observations are clustered into sub-clusters by the agglomerative method. But we stop further agglomerating sub-clusters into larger clusters. In other words, every person contains only a set of such sub-clusters rather than hierarchical clusters. By this way, those sub-clusters which are close to each other are assigned special labels, indicating that they all belong to a certain upper cluster. In Fig.5.3(c), the cluster boundary of a certain person is marked with dotted circle, meaning that it is only a virtual one to demonstrate which sub-clusters are contained. It does not exist in reality and therefore has no overall centroid to be compared to other virtual clusters.

The second step is to apply partitioning method. We predefine the maxi-

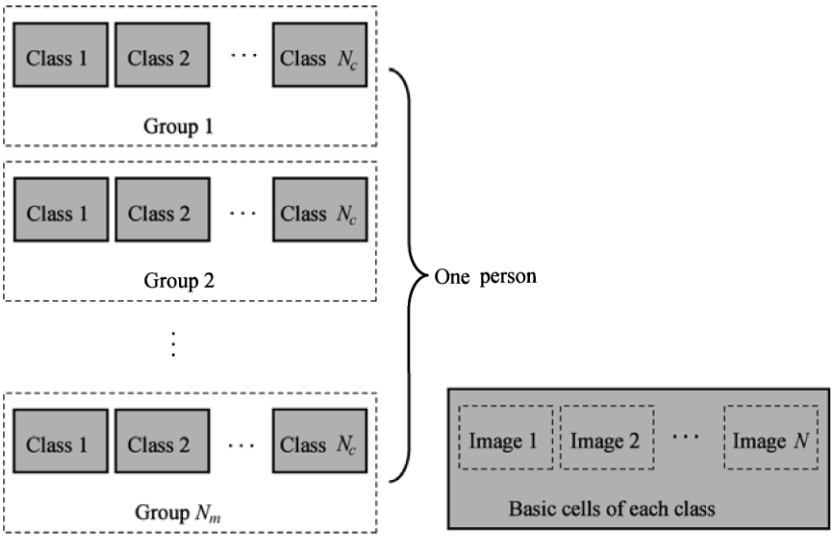
mum number of observations (face shots) per sub-cluster and the maximum number of sub-clusters per person. When a new observation occurs, it is to be compared with all the sub-clusters from all persons. If a match with a sub-cluster is found, the new face shot is added to the corresponding sub-cluster. When the current sub-cluster is full, a new sub-cluster is created. The new sub-cluster is labeled that it belongs to the same person of the matched sub-cluster. In the extreme case, if the maximum number of the sub-clusters for a certain person is reached, the new observation is added to replace an old one.

As can be seen in Fig. 5.3(c), the fusion method has a significant improvement regarding boundary overlap. As a penalty, however, the false rejection rate is pretty big. There are many blanking regions between the sub-clusters and the virtual clusters. Those blanking areas are the possible regions that false rejection occurs. Intuitively, the blanking regions become much smaller when more sub-clusters of a certain person are created. It corresponds to the stable state that one person has long enough interactions with the face recognition system. However, when observations from a new person occur, the false rejection rate is so high that face shots of the same person are very likely to be clustered into different virtual clusters, i.e. enrolled as different persons. We denote this condition as the unstable state.

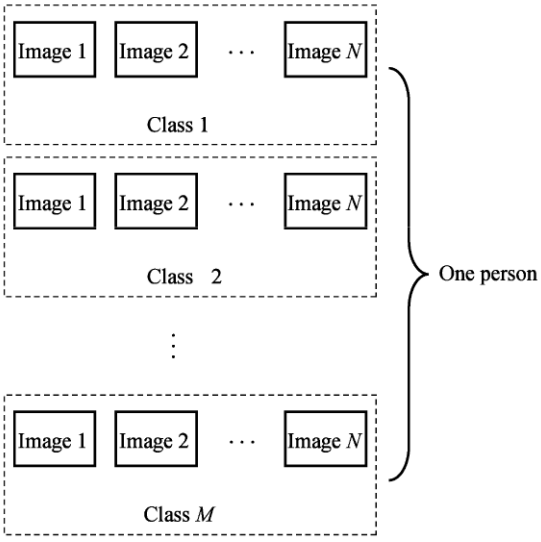
Following the above discussion, the structure of a face database is redrawn in Fig. 5.5. For the simplicity of expression, we use the more general terms "class" instead of "sub-cluster", and "group" instead of "cluster". In the unstable state, the database is a hierarchy of three layers, which are images, classes, and groups respectively from bottom to top.

As shown in Fig. 5.5(a), each group contains a maximum of N_c classes, and each class includes a maximum of N_i face images. When the number of all groups reaches the predefined threshold N_m, their intra-distances are calculated to determine whether they are close enough to belong to the same person. Near-distanced groups are merged to make a new larger group, which is corresponding to one person (a same face). This condition can be defined as a stable state, where the "group layer" can be actually removed. In other words, the structure of the stable state has only a two-layer hierarchy, with one person being composed of M classes and each class consisting of N images. The structure in stable state is illustrated in Fig.5.5(b).

From the previous analysis, we can draw the following conclusions regarding our proposed structures:



(a) Structure of one person's database (in unstable state)



(b) Structure of one person's database (in stable state)

Fig. 5.5 Face database structure in two different states

- The purpose of the proposed structure is mainly to improve the false acceptance rate, compared to the traditional partitioning method.
- In reality, the existing unstable state makes it difficult to achieve the goal.

The transition between the unstable state and the stable state not only produces many computational efforts but also brings in significant recognition difficulties. The most important challenge is that it might increase the false acceptance rate during group merge.

- Same face decision algorithms can significantly contribute to solving the difficulties produced by unstable state. Actually, the unstable state as well as the state transition derives from the following fundamental: each newly created group statistically has much higher false rejection rate than a relatively larger group which contains more classes. Inspired by the prior knowledge and deductive reasoning method in supervised learning, we have proposed the same face decision algorithm rather than the cluster method alone to compensate for the high false rejection rate, which has already been covered in Section 4.4. Let us look back to the example of Fig.5.3(c). With the compensation, observations in the blanking regions will be hardly clustered into a new person. Accordingly, each sub-cluster (class) inside one person is internally linked by special labels so that they can be distinguished from sub-clusters (classes) from another person.
- From the third conclusion, the three-layer structure for the unstable state is not necessary any more since the same face decision algorithm is robust to deal with this state. Therefore, Fig.5.5(b) can be taken as the general face database structure for the stable and the unstable states. With the single structure of only two layers, all the computational efforts and difficulties resulted from the unstable state are removed, which improves the calculation efficiency of our proposed clustering method.

5.3.2 Parameters in the Proposed Structure

In the generic face database shown in Fig.5.5(b), we have to examine how to choose the optimum values of M (maximum number of classes per person) and N (maximum number of face images per class).

N is chosen by following the recommendation of the FaceVACS technology. In Fig. 4.4, we have already shown an example of FAR/FRR curves which applies FaceVACS recognition classifier. There are two groups of data tested for comparing the FAR and FRR. FAR1 and FRR1 represent the case in which only one image is stored as the person's database. FAR16 and FRR16

belong to the group in which 16 images are saved as the database. It can be clearly seen that, both FAR and FRR are greatly improved by taking a higher value of *N*. But in principle, it does not mean the larger *N* the better. There will be too high redundancy when a huge number of *N* is chosen. Moreover, a too large *N* can be harmful to encode the image data into a class. Therefore, the FaceVACS classifier suggests that *N* in the range between 8 and 12 should be suitable. It is declared that the FAR/FRR curve remains nearly the same as *N*=16 when *N*=8~12 is chosen. We keep this range in our implementation.

More generally, *N* can be always roughly determined by two ways. If a certain recognition classifier provides the FAR/ARR curves of different number of enrollment, as FaceVACS does, the range can be easily obtained. But if the required data or suggestions are not available, we can also use some testing sequences to draw the FAR/ARR curves. There are also standard image-based face databases available [2, 3, 4, 5, 6], which are suitable for achieving the FAR/ARR curves as well.

In principle, with *N* face images which are manually selected to be different enough, the database should work well for recognition. But we lack the manual selection, the goal to maintain a high recognition rate is therefore achieved by increasing the redundancy. Hence, we require many classes rather than a single class to represent a person. Here the range of *M* is empirically estimated as 6~12.

We have made the following experiment as a support. Two image sets are recorded for one person. The first set S1 contains 392 face images. They are different in time (more than half a year interval), in lighting conditions, in resolutions and in expressions, which can simulate most of the possible face images of the person except the aging conditions. This image set is used for testing the recognition quality. The other image set S2 contains only 64 face images, which is recorded within several minutes. The sequence can simulate the initializing state. This set is used for enrollment procedure.

Fig. 5.6 shows the two enrolled classes in graphics. Those images are continuously selected from S2.

Fig. 5.7 lists some of the example images in set S1. With two classes enrolled, the recognition result with S1 is,150/392 images that are correctly recognized. The recognition rate is 38%. It indicates that 2 classes are still far too few.

Fig. 5.6 Two classes in the testing set S2

With 8 classes enrolled, however, the recognition result improves a lot:209/392 images that are correctly recognized, the recognition rate is 53%.

With 12 classes enrolled, the recognition rate is only further increased by about 3%. With more than 12 classes, improvement of the recognition rate can be negligible. In Section 7.4, we will provide the empirical choice of the actual M value in the software implementation.

Fig. 5.7 Examples of the testing set S1

It has to be noted that the recognition rate is still quite poor due to the following reasons:

- To observe the influence of *Nc* on recognition quality, the same face decision algorithms are better not included.
- Many of the enrolled images are too similar to each other—too much database redundancy.
- Online update of the database is missing. When a face is recognized and fulfills the update rules, it is added to the database and can therefore contribute to the future recognition.

In the following, we will discuss the general features of an optimum database, including how to decrease the big redundancy and how to update the database.

5.4 Features of an Optimum Database

Regarding mugshot selection, following features are proposed for an optimum database:

- Purity, no face shot from any other person is allowed.
- Variety, only various enough face shots are enrolled.
- Rapidity, at the beginning of building a new database, a rapid growth of the database is important.
- Updatability, the database should be able to keep up with recent views of persons.
- Uniqueness, each person should have one single database to avoid confusion.

Purity is actually another term indicating that a database keeps extremely low false acceptance rate. Our proposed clustering method contributes a lot to this feature.

Variety is important for a database to be complete enough to keep a low false rejection rate. It is crucial for identifying a person in different views, facial expressions and head poses. Two states are distinguished for enrollment. One is the initialized state and the other is the stable state. In the initialized state when a new database is created, a rapid growth is important. In principle, the more numbers of enrolled face shots for one database, as long as they are not the same, the lower FRR is achieved. More face shots are hence to be en-

rolled in this state. In the stable state, however, the selection of enrolled face shots should mainly concentrate on their variety. An adaptive updating threshold (AUT) is used to guarantee the selection in such a floating way. One mugshot is enrolled if the following equation is fulfilled:

$$AST_2 < S_v < AUT, \text{ and } n_e < n_{th} \tag{5.1}$$

where AUT is decreasing for each database growing from initialized state to stable state, AST 2 denotes a threshold which is slightly smaller than AST, n_e is the current number of enrolled face shots in a certain database, and n_{th} is the threshold number of enrolled face shots which indicates the saturation of a database. The database in saturation is supposed to have enough enrolled face shots for identifying a certain person.

AUT is also used to discard the static face-like object which causes the false positives from the face detection procedure. Although detected as one face, such a static object has a hardly change for a long time and can be enrolled into a database with only one mugshot. It can be therefore obviously distinguished from a real face in video. The database automatically removes such databases.

AST_2 contributes to the database purity, which is the most important feature of the databases. Hence, when the image-based face recognizer fails with faces still identified, the enrollment should be careful enough. Face shots with S_v much smaller than AST are discarded for enrollment to avoid bad quality face shots. As mentioned in previous section, the filter with several S_v buffered for recognition also assists the purity.

Another important feature of a successful database is the updatability. Face shots tested from different days statistically have more difference than those from the same day. Enrolling a few known face shots from every day can improve the FRR. With n_e bigger than n_{th}, Equation (5.1) is no longer fulfilled, and a time parameter is therefore introduced to trigger the update of a saturate database. To keep the information from old days, only part of the databases is updated, i.e. only a certain number of face shots are selected for substitution. The face shots to be replaced should from the oldest days and have the most similar values when compared to others.

Since the violation of uniqueness is less harmful than that of the purity, databases tolerate it during the construction. But those databases are to be merged after careful calculation. The mutual similarity values (MSV) between each database pairs are computed. If MSV is bigger than AST, each face shot from one

database is further checked. When a certain enough percentage of face shots in one database is identified with the other, the two databases are merged. This check avoids a wrong merge. Because the step needs relatively large processing power, it is only enabled during an idle period when no face is detected for a certain length of time.

References

[1] Tom M. Mitchell: *Machine Learning*. Singapore, McGraw-Hill, 1997

[2] NIST Face Database, http://www.nist.gov/srd/nistsd18.htm, accessed 10 December 2006

[3] BioID Face database, http://www.humanscan.de/support/downloads/facedb.php, accessed 10 December 2006

[4] CMU/VASC image database, http://vasc.ri.cmu.edu/idb/html/face/index.html, accessed 10 December 2006

[5] AT&T face database, (formerly The ORL Database of Faces), http:// www.uk.research.att.com/facedatabase.html, accessed 10 December 2006

[6] MIT Face database, http://cbcl.mit.edu/software-datasets/FaceData2.html, accessed 10 December 2006

6 State Machine Based Automatic Procedure

Abstract This chapter introduces a state machine based method to put all the detection, recognition, and database construction/update procedures together as a whole automatic and self-learning system. The major points are how to define different states and how to define their corresponding transitions. By utilizing the combined information such as video context, temporal information, and logic analysis, an efficient and stable state machine structure is proposed.

6.1 Introduction

In the previous chapters, we have explored the algorithms that are suitable for each individual steps to implement the whole automatic face recognition procedure, including face detection, face tracking, enrollment, matching and update. But how to put them together to automatically work is still a challenge.

Firstly, we have to consider all possible situations that might occur in every step. Any state that is overlooked can make the state machine stop working.

Secondly, since each step is closely related or is even decisive to the following steps, the following questions have to be answered: under which condition a state should transit; how many states are linked to a certain state, etc.

State machine is the appropriate tool for this task, which can easily define the complete states and the corresponding conditions for state transition. During the discussion in the next section, we might notice that, all useful informa-

tion such as the results from the detection components, the results from the recognition components, video context, temporal information, and logic analysis, is considered for designing the states and their transitions. This further represents the intelligence of the proposed system.

With the help of state machine, the whole intelligent, automatic and unsupervised face recognition system is completed. The implementation issues and performance analysis will be covered in the next two chapters.

6.2 States Explorations

To demonstrate the hierarchical relationship among all states, we use a tree-like state diagram, as shown in Fig.6.1.

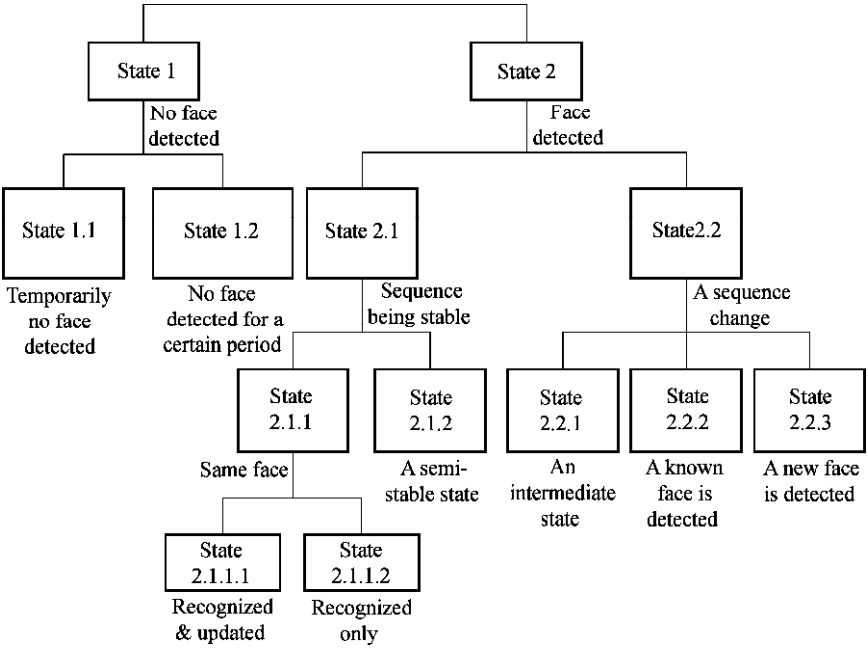


Fig. 6.1 Definition of all possible states for the face recognition procedure

State 1 is the case when no face is detected. **State 2** represents the contrary situation that a face is detected.

Two states are divided for the non-face case. **State 1.1** is a temporary state that no face is detected. It includes the following two major possibilities. One

is due to the false negatives in face detection. In reality, there is a face existing, but the system cannot find it. Therefore, this state should be ready to transit back to its previous state. Another case is that, no face actually exists. It happens when a person starts leaving from the camera. Together with the output from the motion-based face detector, the system can decide under which situation the current state is. Small motion indicates the former case and big motion means the latter case. For the latter case, the system then waits for several frames to see whether the person is really leaving the camera. If so, **State 1.1** is transited to **State 1.2**, where a stable state is defined which indicates that no face is detected for a long time. This state is also called an idle period. Complex computations such as combining databases can be executed in this period.

The system is always in **State 1.2** until a face is detected. Then it transits to **State 2.2**, a new sequence state, meaning that a new sequence starts. At any time, when no face is detected, **State 2.2** transits to **State 1.1**. **State 2.2** consists of three further states. **State 2.2.1** is a temporary state. Since we introduce a temporal filter with the maximum filter length n in Equation (4.2), it requires some time for the filter length l to increase from zero to n. This intermediate procedure is defined as **State 2.2.1**. It can either remain the current state or goes to **State 1.1** when the filter length is below n. Once l reaches n, **State 2.2.1** is ready to transit to either **State 2.2.2** or **State 2.2.3** according to the filtered similarity value $\sum_{i=1}^{n} A_i \ S_{v,i}$. For the simplicity of representation, we denote $\sum_{i=1}^{n} A_i \ S_{v,i}$ from Equation (4.2) as S_{vF}. When S_{vF} is beyond $m \bullet AST$, the current face is identified as a known face and **State 2.2.1** is going to **State 2.2.2**; when S_{vF} is below $m \bullet AST$, the current face is detected as an unknown face and **State 2.2.1** is going to **State 2.2.3**.

The next possible transition state of both **State 2.2.2** and **State 2.2.3** are **State 2.1.1**, which is the same person state. In this state, the current face is expected to remain the same for at least n frame. It includes two different states. **State 2.1.1.1** is the state which fulfils the update rules. In other words, the current face is selected to be enrolled into databases. On the contrary, **State 2.1.1.2** indicates that the current face is only recognized. It is too similar to existing databases and will not be enrolled. The transition between **State 2.1.1.1** and **State 2.1.1.2** is dependent on Equation (5.1).

But there is also a possibility that multiple people exist. It is then quite

normal that one person is approaching the camera as well. When two faces are slowly but alternately occluding one from the other, it is still in the same sequence state, but not in the same face state any more. Therefore, we define it as **State 2.1.2**, meaning that another face is possibly showing up. The transition between **State 2.1.1** and **State 2.1.2** is based on the combined same face decision algorithms described in Section 4.4. When **State 2.1.2** is kept for a certain time, indicating that another face is really showing up, it transits to **State 2.2.1**, the temporal sub-state of the new sequence state.

The hierarchical state structure of the state machine is shown in Fig. 6.2.

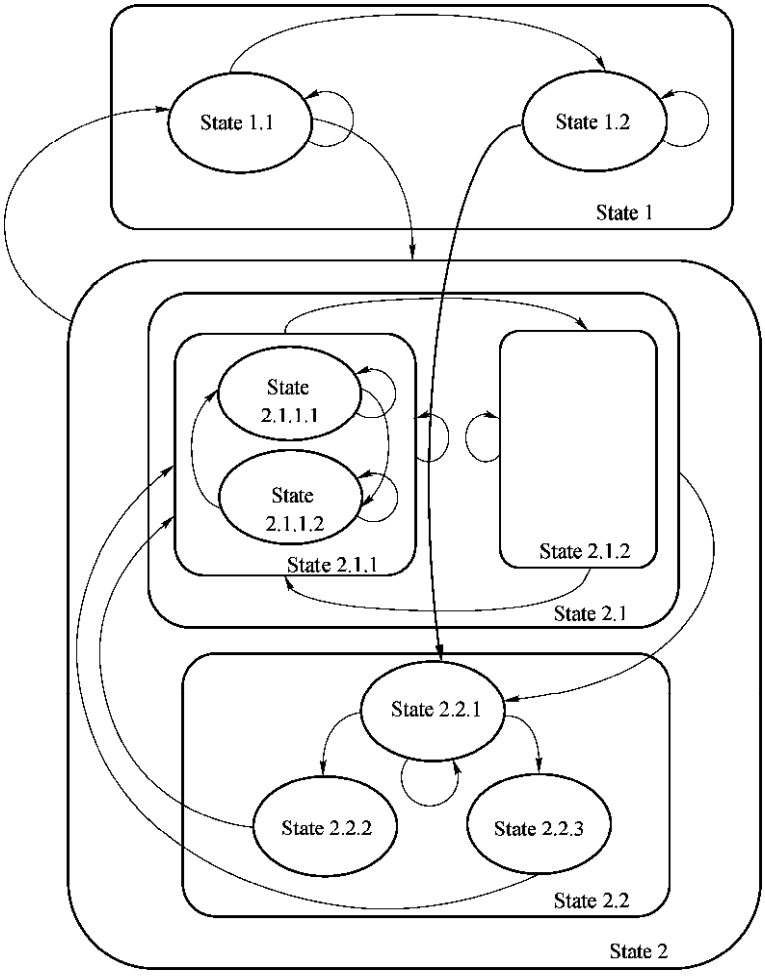


Fig. 6.2 Definition with a hierarchical state machine

7 System Implementation

Abstract This chapter covers the implementation issues of the proposed intelligent, automatic, and unsupervised face recognition algorithms. Followed by the introduction section, in section 2, the typical required hardware configuration is presented. Section 3 is regarding the software implementation, where two versions are designed for different demonstration purposes: online and offline versions. Technology dependent parameters which have to be considered for implementation are discussed in section 4. Finally, this chapter ends with the summary section.

7.1 Introduction

In the previous chapters, the ideas and methods on achieving an intelligent face recognition system are provided. To examine its robustness, an efficient implementation is required. Moreover, we are also expecting to demonstrate the pervasive application prospects of the proposals. As a consequence, two major points have to be considered: typical and cheap hardware configuration is sufficient for running the implementation; high speed or even real-time software design is preferred. In this chapter, we are mainly focusing on those implementation issues, which are not only based on the generic algorithms but also the specific detection and classification techniques.

7.2 Typical Hardware Configuration

The principal criterion to select required hardware is to keep as simple and generic as possible. We would like to demonstrate that the algorithms we have discussed are not dependent on any special hardware and can be easily embedded to the applications we have mentioned in Section 1.3. This is very important, especially for the application in consumer electronics industry, a system with cheap and generic setups is always attractive and promising.

Thanks to the Moore's law, the processing speed of hardware is dramatically increasing from year to year. The corresponding hardware configurations should also adapt to the change. In this section, we only provide hardware platforms as examples to demonstrate how the automatic procedure is running. The readers are encouraged to consider their own preferred platforms.

A generic hardware set-up is shown in Fig. 7.1(a), which has three major components: Video sources which can be from TV channels or from live videos, interface, a PC or a laptop. The interface can be any video capturing interface that can grab frame images from the camera to the PC/Laptop, for example, a PCI TV capture card connected to a PC, or USB/firewire interfaces to PC/Laptop.

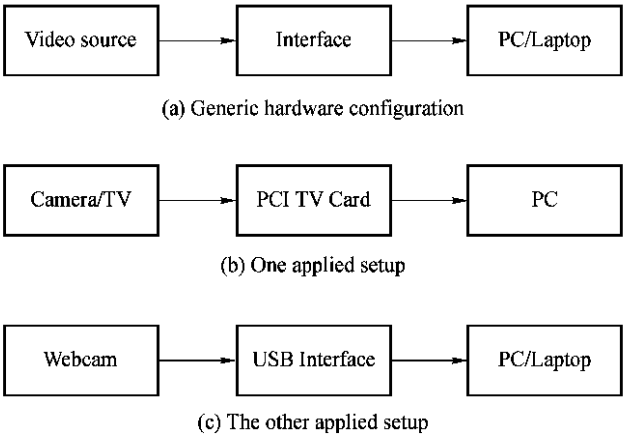


Fig. 7.1 Hardware configurations of the system

Fig. 7.1(b) and (c) show two different setups we have used for our system. In Fig. 7.1(b), the video source can be from either a video camera or from TV sequences. It is connected through a standard PCI TV capturing card to a PC.

The camera and the capturing card can provide up to 768×576 pixels at 12 fps and 384×288 pixels at 25 fps. The PC is used to run the software based on Microsoft Windows. This configuration therefore supports a real-time running of the system from either cameras or TV programs.

In Fig. 7.1(c), a webcam is connected to PC/Laptop through a USB 2.0 interface. At the time when we start the hardware configuration for collecting the test sequences in 2001, the webcam supports the resolution of 320×240 pixels with only 12 fps. This setup is hence more oriented to demos running offline.

7.3 Software Implementation

7.3.1 Overview

The algorithms are implemented in Visual C++, supported by libraries from the DirectX SDK tool [1] for video capturing and playing back issues, libraries from the OpenCV tool (Open Source Computer Vision, [2]) for image processing, libraries from the FaceVACS SDK tool [3] for image-based face detection and image-based recognition classification.

With available PCs in the year between 2001 and 2004, our algorithms based on the OpenCV and DirectX tool have no big difficulties to be implemented for running at nearly real-time. They can meet the requirement of the temporal-based face detector of 10-20 fps processing speed. But the speed bottleneck could come from the classifier running with a low-speed CPU, which can not achieve the 10-20 fps for feature extraction and encoding steps in the enrollment procedure. This limitation hence doesn't fit well to the tracking requirement. Therefore, we have designed two versions to demonstrate the algorithms: online and offline versions.

Online version processes the video from camera or from TV programs. The actual processing speed of the system is around 1-2 fps for a 384 × 288 resolution on a 2.4 GHz PC, which does not fulfill the 10-20 fps assumption. Therefore, it can only achieve an ideal performance when people behave 5-10

times slower than normal. Although not optimum, the attractive advantage of the online running version is that it is much more intuitive to evaluate the system performance by changing scales, head poses, facial expressions and even lighting conditions.

With nowadays new dual-core CPUs, for example, with 2.66 GHz or above, the online version has no problem to fulfill the 10-20 fps assumptions. We can take the advantage of higher processing speed to better evaluate the system performances.

Fig. 7.2 shows a screenshot of the system running with online live video.

Fig. 7.2 Screenshot of the system running live

The top right part of the user interface is the live video window, where the processed video is played back. People in front of the camera can hence observe their own face movement online. The middle right part is the visual thumbnail of the databases which are updated online. It shows how many persons have been stored in the database with one mugshot displayed. For such an unsupervised and automotive system, it is not possible to display the peoples' real names. Therefore, every enrolled person is assigned a person number and lined up in the order of the enrollment time. A known person which is currently recognized is overlaid by a red frame on the mugshot of the corre-

sponding face. Bottom right is a text window which is updated in every processing frame of the video. It includes the following information: whether a face is detected; the person's number if recognized; the similarity Sv between the current face and the corresponding databases. On the left part of the interface, all mugshots in the database of the current recognized person are displayed. It is then intuitive to evaluate the database quality of a certain person. In the following example, people 0 through person 17 are enrolled from live TV broadcasting news, and person 18 is enrolled from the source of a live video camera. A more detailed performance analysis is presented in Chapter 8.

The other is the offline version, which is fed by image sequences stored in hard disks. For this version, it is required to record sequences before running. Since it processes every frame of a given sequence, it can simulate any processing speed dependent on the recording frame rate. This version is useful to demonstrate the nominal performance of the proposed algorithms.

7.3.2 Implementation Efforts

The software implementation of both the online version and the offline version are basically the same in terms of recognition algorithms. But the online version has to additionally deal with real-time capturing and playing back videos. We therefore take the online version to demonstrate the C++ based implementation efforts. About 2800 command lines (excluding the comment lines) in the source code of Visual C++ are written for this online version implementation.

The detailed implementation functional components are listed in Fig. 7.3. There are two main units: the user interface unit and the algorithms unit.

Three main tasks are concerned for the first interface unit:

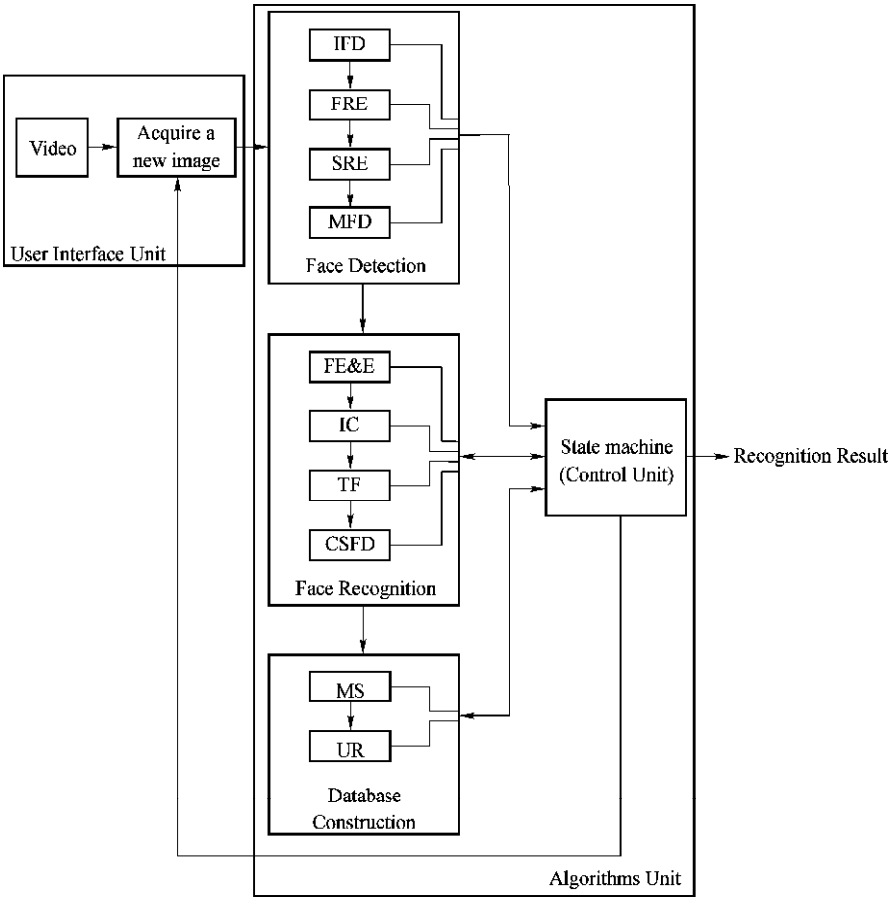
- Real-time video capturing. The DirectX SDK is applied to deal with this task.
- Real-time displaying of the captured video. This task is also implemented based on the DirectX SDK.
- The user interface, as shown in Fig. 7.2.

For the first task, the captured video has to be saved as frame images. DirectX only supports the .bmp file format in which the pixels are true-colored 24 bits. For the algorithms unit, however, .pgm file format is required. It is to be noted that the processing speed for each frame is about 3 times faster if the

greyscaled .pgm file format is used instead of the 24-bit .bmp file. Such an image format conversion function is included right after capturing.

For the second task, a challenge exists since the capturing and displaying should be accomplished at the same time. Multi-thread method has to be applied in the source code.

Regarding the user interface, it is also a big effort to fast overlay different database thumbnails. Therefore, 2/3 of the source codes are dealing with the interface unit.



IFD: Image-based Face Detection, FRE: Face Region Extraction; SRE: Search Region Extraction; MFD: Motion-based Face Detection; FE&E: Feature Extraction and Encoding; IC: Image-based Classifier; TF: Temporal Filtering; CSFD: Combined Same Face Decision; MS: Mugshot Selection; UR: Update Rules.

Fig. 7.3 System implementation components

A large amount of computing power is required in the interface unit as well. The video displaying function itself consumes half of the computing power of the overall system, which explains why the online version is running much slower than the offline version.

One third of the source codes (about 1000 command lines) are contributed to the algorithms unit, where around 250 lines are written for face detection, 360 lines for face recognition, 180 lines for database construction, and 210 lines for the state machine based central control unit. The control unit delivers the final recognition output.

7.4 Technology Dependent Parameters

Up to now, we have mainly discussed the general algorithms for building an automatic and adaptive face recognition system. The discussion is supposed to be as independent as possible from the image-based face detection method and image-based face classifiers. But we still have some specific parameter settings that are technology dependent, which are required to achieve better performance of the whole system. As a part of the system implementation, those parameters are interesting to be described. In the following, we summarize those important parameters.

- Database construction parameters, the number of images per class N and the number of classes per group M.
- The minimum value of AST (adaptive similarity threshold), S_{V0}, introduced in Equation (4.1).
- Which one is better for feature extraction, using the whole frame image or only a face region (face shot)?

The first two parameters have been already briefly discussed in Section 5.3.2. As illustrated in Fig. 4.4, the higher the S_{V0}, the lower the FAR, and the higher the FRR. However, the whole number of enrolled face images for one person $M \times N$ can also influence FRR. As shown in Fig. 4.4, in principle, the bigger $M \times N$, the less FRR. With given N and M, we can correspondingly judge an optimum range of S_{V0}. In Section 5.3.2, the ranges of M and N were empirically determined through the tests with only part of the system implemented. To further determine the exact values of those parameters, we are now able to apply the system with all components implemented. Only the parameters to be

tweaked are left open. As test resources, we have chosen video sequences from 30 people, with at least 100 frames per sequence and at least one sequence per person. The optimized values are finally determined as: M=8, N=8 and S_{V0} = 0.71. It is to be noted that S_{V0} is bigger than the suggested value in the FAR/FRR curves of Fig. 4.4, which are due to the following two main points. Firstly, the FAR/FRR curves are normally obtained based on the manual mugshot selection procedure. Therefore, the database is constructed even without errors. For our system, however, mugshot selection is an automatic procedure with no guarantee of 100% database purity. Consequently, extremely low FAR is required. On the other hand, with S_{V0} = 0.71, the penalty in principle leads to extremely high FRR. For our system, however, much lower FRR is achieved due to the combined face decision algorithms and the adaptive update rules.

The third technology dependent parameter has not been covered in previous chapters. As most face recognition systems do, our original idea was to use the extracted face region from a frame image for feature extraction as well as for database enrollment. In terms of the processing power, the use of face region is obviously advantageous over that of the whole image. Moreover, it is commonly considered that, the background can do harm to the recognition quality if a whole image is enrolled into the database. However, the following experiments demonstrate contradictory evidences. We have selected two out of thirty subjects who are relatively similar to each other. The samples of the whole images and the corresponding face images of the two persons are shown in Fig. 7.4.

Person A

Fig. 7.4 Samples of whole frame images and the corresponding face images

In the first set of experiment, we have tested the cross-similarity between the two persons by using the face images and the whole images respectively. The result is illustrated in Fig. 7.5. Since a higher cross-similarity indicates a higher FAR, the dashed curve achieves better recognition performance. That is to say, the use of whole images achieves lower FAR than that of face images.

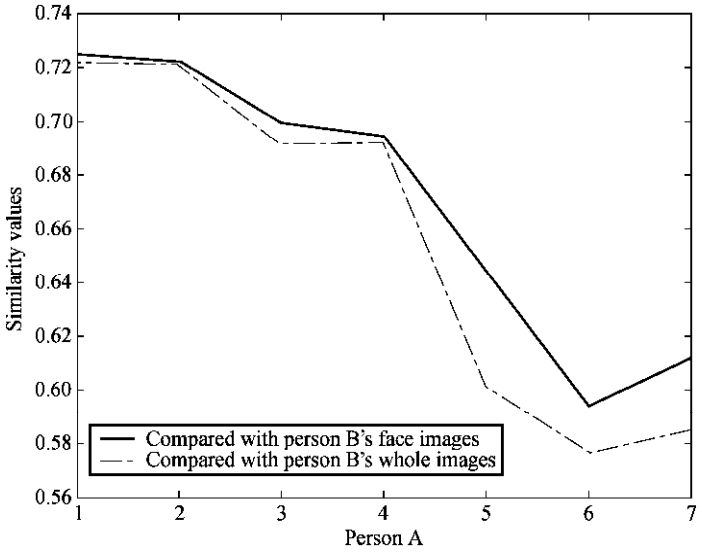


Fig. 7.5 Cross-similarity differences by using face images and whole images

Auto-similarity is tested in the following experiment. Two different sequences from the same person are compared with each other, as shown in

Fig. 7.6. As shown in this curve, the higher the auto-similarity, the better. The use of whole images (dashed curve) is again advantageous in terms of lower FRR.

As FRR and FAR can be both lowered, whole images should be applied instead of face images. This conclusion is due to the specific algorithms applied in the face classifier. But the fundamentals behind it are still unknown since the detailed algorithms are not open to the public.

Although the whole images are applied, we still display the face region thumbnails instead of the whole images to indicate the databases in the graphical user interface, as shown in Fig. 7.2. In this way, two main advantages are obtained. Firstly, it is easier to be understood for normal users because any surrounding background seems to be non-related to the face identity. And secondly, the image overlaying speed is much higher if only face regions are displayed. We also use the term "face shot" to point the enrolled images for database construction to agree with the common way of appellation.

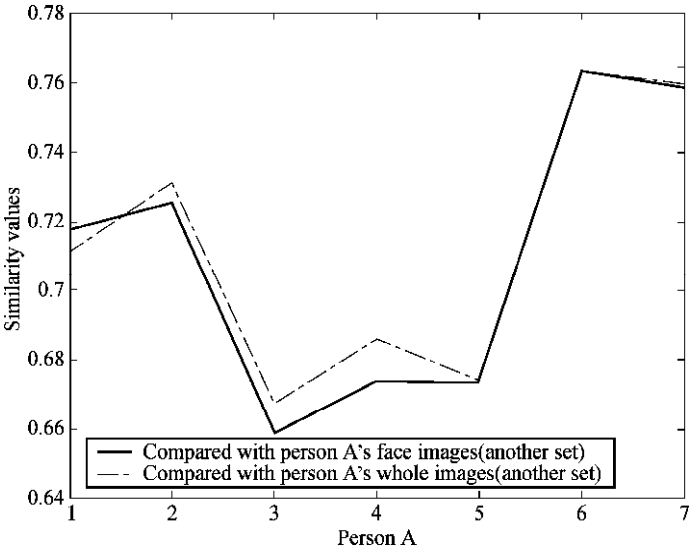


Fig. 7.6 Auto-similarity differences by using face images and whole images

7.5 Summary

In this chapter, we have mainly focused on the software implementation ef-

forts of the proposals. It reveals that, the completely automatic and self-learning system is computationally simple, and achieves the ease of implementation, due to the introduction of state machine. The analysis of fusing multiple methods in state machine contributes to the high speed of processing as well.

While combining multiple methods for the improved recognition, the implementation has been as independent as possible to each single method. In this way, the readers can more easily apply the fundamental ideas in this book for their own research.

References

[1] Freely available tool from Microsoft, http://www.microsoft.com/downloads, accessed 10 October 2006

[2] Freely available tool from Intel, http://sourceforge.net/projects/opencvlibrary/, accessed 10 October 2006

[3] Commercial product from Cognitec Systems GmbH, http://www.cognitec-systems.de, accessed 10 October 2006

8 Performance Analysis

Abstract This chapter presents the performance analysis of all the proposed algorithms discussed in this book. Followed by the introduction section, in section 2, the combined face detection algorithms are experimented and compared to the IFD, which demonstrates the significant benefit of applying the fusion methods. In section 3, the algorithms designed for the recognition procedure are evaluated, which demonstrates the robustness against the change of scale, facial expression, lighting, glass, and occlusion. In section 4, the database construction and update rules are verified. Then in section 5, the overall performance of the system is presented with both the online and the offline version. Finally, this chapter ends with the summary and discussions of the performance analysis issues.

8.1 Introduction

In this chapter, we are analyzing the performance of the proposals applied for and contributed to the whole intelligent face recognition procedure. Separate components like the face detection algorithms, the recognition algorithms, and the database construction/update rules will be evaluated respectively.

Regarding face detection, as mentioned in chapter 3, the combination of IFD and MFD shall show significant advantages. We should accordingly make the performance comparison between the combined method and the single IFD.

Similar comparison is also made for demonstrating the face recognition performance. The ability of dealing with typical recognition difficulties such as occlusions, pose variations etc. shall present the success of the algorithms.

Experiments are designed to verify the expected features of building/updating databases, including the variety, rapidity, updatability, and uniqueness.

The whole system evaluation is required to demonstrate whether the proposed face recognition system is intelligent, and whether it can run in a fully automatic and unsupervised way.

8.2 Performance of Face Detection

To test the performance of the proposed algorithms, we have made experiments with about thirty subjects. Each subject has been recorded with several sequences under very different conditions. The sequences are then processed by IFD only and our system for comparison. Each video sequence is captured at 25 fps.

As discussed in Section 3.2.4, head poses, scales, facial expressions, lighting conditions, motion blurs are the most critical parameters to evaluate the face detection quality. For video-based face detection, occlusion is another crucial parameter especially for tracking the faces.

We begin to analyze the detection performance with the most critical case: pose variations. The three dimension changes including head roll (example is shown in Fig. 8.1), head yaw (examples shown in Fig. 8.2) and head pitch (example is shown in Fig. 8.3) are respectively tested.

With the help of temporal information applied in the MFD, the system successfully complete the all the three detection procedures, regardless of the head motions in each different dimension. Compared with the 100% detection rate of the proposal, the IFD achieves only 66% for head pitch, and unfortunately 55% for head yaw and 51% for head roll.

Fig. 8.1 Test of 3D head motions—head roll

Fig. 8.2 Test of 3D head motions—head yaw

Fig. 8.3 Test of 3D head motions—head pitch

As a further step, we are interested in the performance with freely 3D head motions. Fig. 8.4 shows such a sample sequence, where the person is significantly moving his head in all three dimensions. Frame t1 to t5, frame t6 to t9 and frame t10 to t13 illustrate the head yaw, pitch, and roll respectively. Frame t14 to t20 show the free rotation of the head in all three dimensions.

In this example, when a head rotates to a certain degree or shows only a certain degree of profile, the IFD fails as many image-based face detection algorithms do. It succeeds with only frame t1, t2, t6 and t10 and fails with all the other frames in Fig. 8.4. But our approach achieves satisfying results with only one failure case occurring in frame t9. In that frame image, the black hair almost occupies the whole face region and some of the facial part is outside the face region but inside the search region. However, when the head is up again

as shown in t8, the face returns to its tracking status.

Fig. 8.4 Sequence of one person with significant 3D head motion

The comparison result of the detection rate for the whole sequence is drawn in Fig.8.5. It can be seen that the IFD can only detect the face in 45% of the whole sequence, while our system achieves as high as 93% detection rate. That is to say, for this particular sequence, 93% of the frame images can be passed through our system for recognition while only 45% of the frames can be used if a single IFD is applied in the face detection procedure.

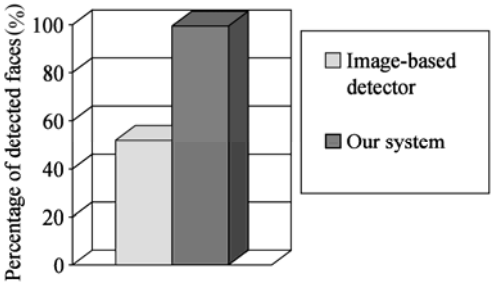


Fig. 8.5 Performance comparison between an IFD and our combined detector

Fig. 8.6 shows one sequence with intentionally different facial expressions. It is captured in the resolution of 192×144 pixels.

Fig. 8.6 Sequence of one person with intentionally facial expression changes

To our first surprise, the IFD behaves even worse with facial expression changes than with pose variations. It fails with 58.5% of the frames (all of which are false negatives) while our system achieves 100%. With more careful examination, we have found out that the failure of the IFD derives from the poor resolution of the eye distance. We have then re-sampled the sequence to be 150% of the original and made the test again. Our approach still achieves 100% detection rate while the IFD detector performs much better with only 2.5% of the false negative rate. As shown in Fig. 8.6, it only fails when the eyes are occluded by hands and arms (frame t14), when the eyes are closed (frame t20), or when the head is rolling with too big angle (frame t17). From the sequence, we can draw two conclusions. Firstly, the IFD is not very sensible to expression variations, and only fails when the eyes are not visible. Secondly, the IFD is very much sensitive to the scale changes especially for too low pixel numbers of the eye distance but our approach is not at all. It is very critical for a video-based processing system to handle low resolution se-

quences to both achieve fast speed and save hardware storage spaces.

Therefore, we have made sequences to further analyze the detection quality against scale changes. Fig. 8.7 shows typical frames of the testing examples. 41.8% of the sequence are failed to be detected by the IFD alone, where 31.8% are false negatives and another 10% are the combined false positives and false negatives. Our method just misses 2 out of the overall 110 frames (1.8%). In Fig. 8.7, frame t5, t6 and t7 are too big for the IFD, and it outputs wrong face positions without eyes detected. From frame t13 through t20, the faces are too small for the IFD to detect any faces. Our approach has no difficulty with t5, t6 and t7 and only fails with t20.

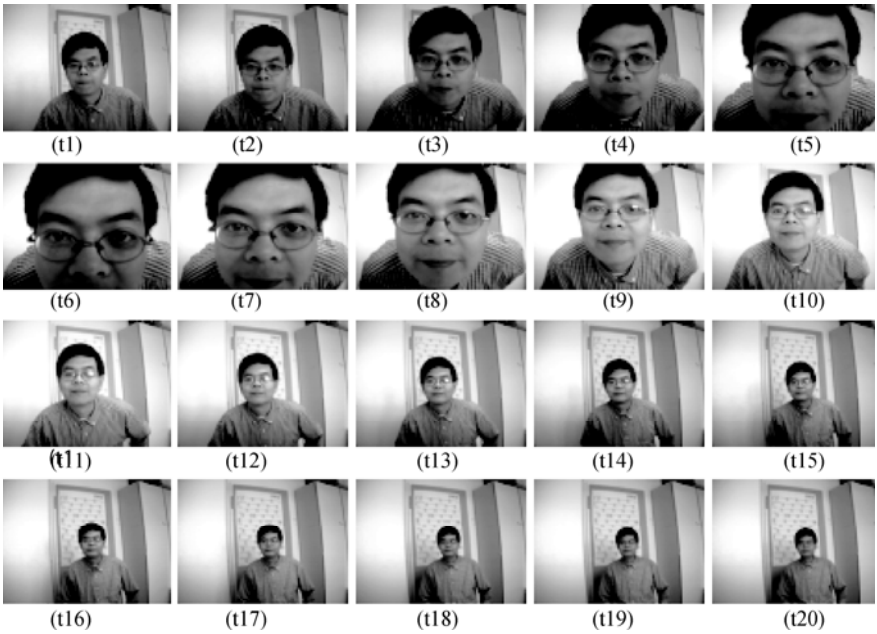


Fig. 8.7 Sequence of one person with significant scale changes

The sequence in Fig. 8.8 demonstrates the cases with sudden and gradual lighting condition changes. Although our proposed MFD has no difficulties in dealing with gradual lighting change as shown from t4 to t5 and from t16 to t17, it in principle cannot handle a sudden lighting change. But the limitation can be well balanced by many successful state-of-the-art illumination compensation methods for image-based face detection [1, 2, 3]. The IFD we have used here is not sensitive to sudden lighting variations. Thus, we are not interested in developing new algorithms to compensate for the illumination changes for

face detection. In this testing sequence, the faces in all frames are correctly detected by the IFD alone (100%) and our combined approach accordingly achieves 100% detection rate as well. As can be seen, the sudden change from frame t6 to t7 and frame t16 to t17 has no influence on face detection.

Fig. 8.8 Sequence of one person with significant lighting changes

The final example sequence shown in Fig. 8.9 is aimed to test the influence of both motion blurs and occlusions. The head is purposely moving fast from left to right and then from right to left. The IFD alone fails with 31% frames and our method fails only with 1 out of 83 frames (1.2%). The failure from the IFD occurs for frame t5, t6, t7, t11, t12, t13, t14 and t15, where t5, t6 and t7 are not successfully detected due to the motion blurs, t11 and t12 are due to head rolling t13, t14 and t15 are due to occlusions. Our approach does not detect the face only in frame t14 where more than 50% of the facial part is not visible.

To summarize the performance, we list the testing result for different parameters in Table 8.1.

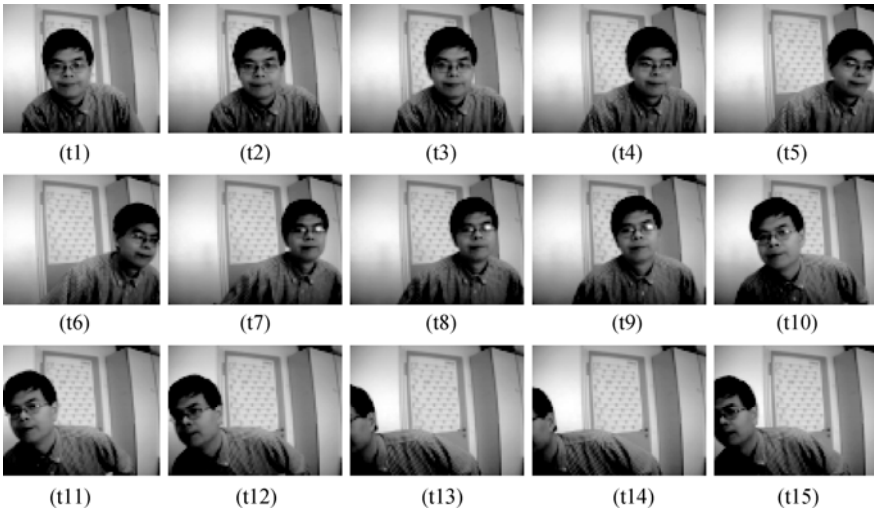


Fig. 8.9 Sequence of one person with fast motion

Table 8.1 Performance comparison between the IFD and our detection method

	Face detection rate under different parameters (in percentage of the whole sequence number)				
	Head pose (Fig. 8.4)	Facial Expression (Fig. 8.5)	Scale (Fig. 8.7)	Luminance (Fig. 8.8)	Motion Blurs and Occlusion (Fig. 8.9)
IFD alone	45%	41.5%	58.2%	100%	69%
Our detector	93%	100%	98.2%	100%	98.8%

As can be seen, our proposed approach significantly contributes to the face detection and tracking in video sequences. The contribution of the high detection rate to the whole recognition system is obvious: much more frame images can be used for recognition with our combined detection algorithms.

8.3 Performance of Face Recognition

The way of evaluating the recognition performance is quite similar to that of the detection quality. Variation of scale, facial expressions, lighting conditions and motion blurs are all critical parameters as well. Since the combined same

face decision algorithms deal with those changes in a better way, we can assume that better performance can be achieved, if compared with the face detection algorithms. We have used the same sequences as for the face detection performance, as shown in Section 8.2. The FaceVACS recognition classifier is taken again as a reference for comparison with our proposed algorithms.

Fig. 8.10, Fig.8.11, and Fig. 8.12 respectively shows the examples of experiments with various face sizes, changing facial expressions, different illuminations, occlusion, and with/without glasses.

Fig. 8.10 Face recognition test—different face scales

The performance of the fusing advantage is further verified through this face scale test. As shown in Fig. 8.10, the face detection even with the help of MFD, fail with 000-003, and 057-063. Consequently, the image-based face recognition also fail with the above 10 frames. However, face recognition is successful in all frames because the same face decision is always successful.

In the facial expression and lighting test, the image-based recognition successes with about 91%, while our proposal is achieving 100%.

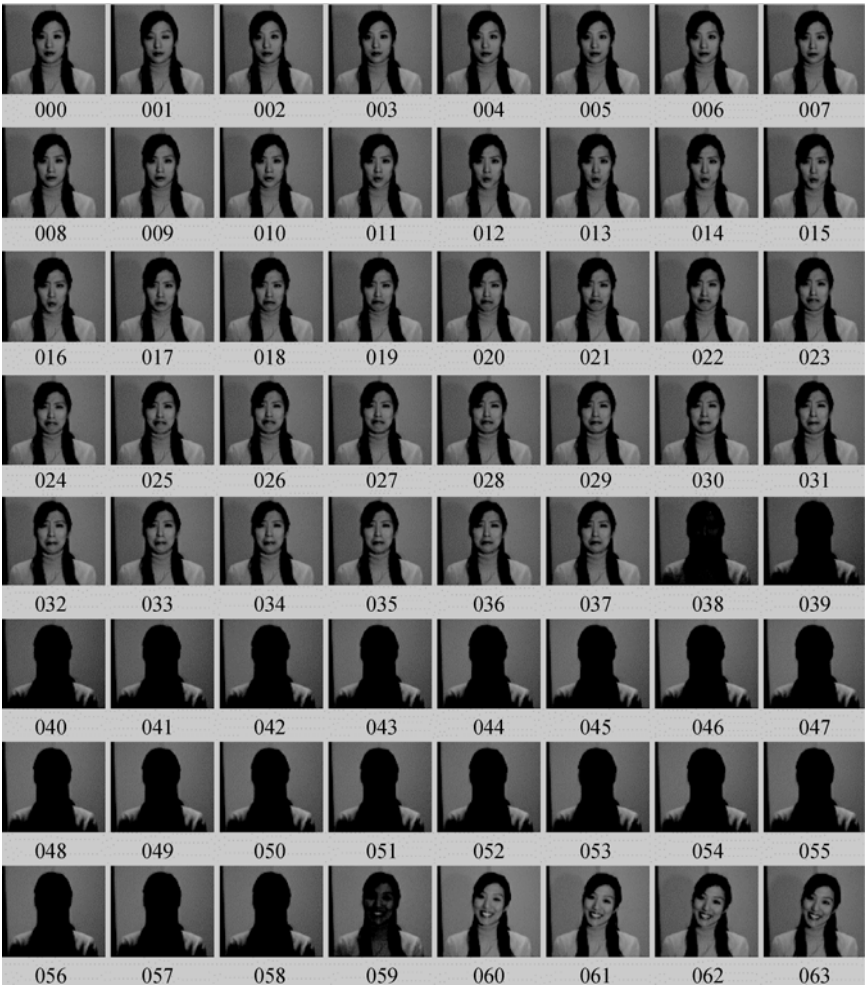


Fig. 8.11 Face recognition test—different lightings and facial expressions

The more difficult case shown in Fig. 8.12 lead to the poor performance of image-based face recognizer, with the recognition rate of only 58%. The pro-

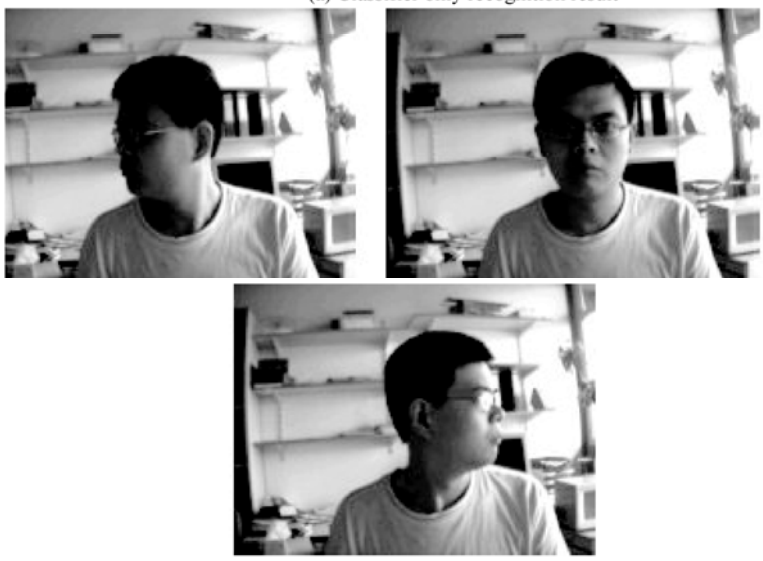
posed system again succeeds in all frames.

Fig. 8.12 Face recognition test—Occlusions and with/without glasses

In the next, we will intensively explore the most critical case: head pose variations. Most recognition systems have problems to deal with this critical change.

The evaluation is done respectively through all three kinds of head pose variations: yaw, pitch and roll, see Fig. 8.13 We list below the corresponding influences for each dimension of head motions.

(a) Classifier only recognition result



(b) Combined same face decision result

Fig. 8.13 Face recognition performance test—yaw angle variance

- **Influence of yaw angles.** The classifier is quite sensitive to this kind of head rotations. Since 2D image analysis is applied and this pose change is actually 3D. It is difficult to show the precise yaw angle limit of the classifier. We hence use a set of sequences to visually demonstrate it. Fig. 8.13 (a)

lists the right and left yaw angle limitations of the classifier. In the 100-frame sequence, the face is turning from frontal to right and then from right to left. The classifier cannot detect and recognize approximately above 30° yawed heads. While our proposed algorithms have no problem for even profiles if the head is rotating continuously. Actually, in the testing sequence, our method achieves 100% recognition rate while the single classifier can only recognize 40% out of all frames.

- **Influence of pitch angles.** The classifier is relatively robust to head pitching. We have applied similar testing sequence as in the yaw angle test, where a person is pitching from frontal to up as much as possible and then from up to down as much as possible. The classifier works well with 85% of the frames and our method fails with only one frame where the face is hardly seen due to too much head bending. It should be noted that, the head pitch can also produce two artifacts that significantly lower the recognition rate of the classifier itself. Through experiments with some subjects, we find out that most people pitch their head with a high speed, producing motion artifacts (e.g. saw edges) on the eyes, as shown in Fig. 8.14 (a). The artifact is subject to be produced due to the poor performance of cheap cameras or cheap video capture cards. Since the eyes are critical for the classifier, the motion artifact hinders it from recognition although the pitch angle is not big enough. Another problem comes from the glisten of the glasses when head rotates. It is also a frequent error for the classifier. However, such artifacts do not influence our proposed method in a video sequence since they normally produce minor pixel differences in the face region.

Fig. 8.14 Examples of artifacts around the eyes which cause recognition errors

- **Influence of roll angles.** Since roll is the in-plane head movement, it can be precisely measured in the 2D image. Fig. 8.15 illustrates the definition of roll angle θ. It is the angle between the connected straight line through two eyes and the x-axis. We have taken several frontal faces from different persons. The frontal faces are manually rolled to achieve exact roll angles by using the matlab image rotation tool (with two different rotation interpolation methods) and the IrfanView tool. The three different rotation methods are applied to make sure that interpolations do not significantly influence the picture quality for recognition. The rolled faces are compared with the original frontal faces by using the classifier alone. Fig. 8.16 shows such a comparison with two face images. From the figure, we draw the conclusion that the classifier cannot recognize the face with θ bigger than 15° when the similarity threshold is set to be 80%. For faces with θ bigger than 20°, the IFD normally fails to detect faces and therefore the classifier fails as well. With our combined face recognition methods, there is no problem to detect and recognize faces even with 90° roll angles in a video sequence.

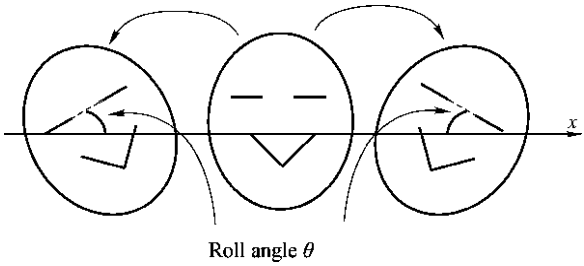


Fig. 8.15 Definition of roll angle

Through the above performance analysis regarding pose variations, we can learn that, each image-based classifier could be more robust in dealing with a certain direction of head motion, while fail frequently with other head motions. Our proposals are independent of those correlations, and can evenly improve the performance, regardless of which kind of head motions.

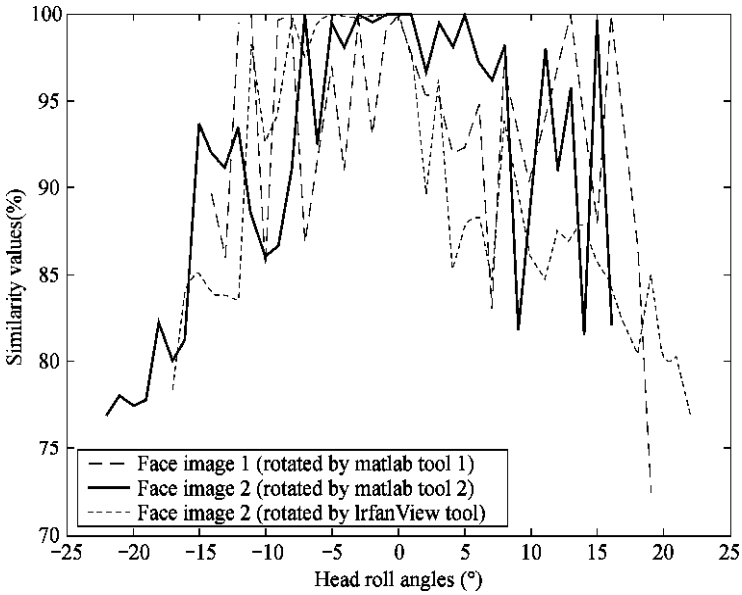


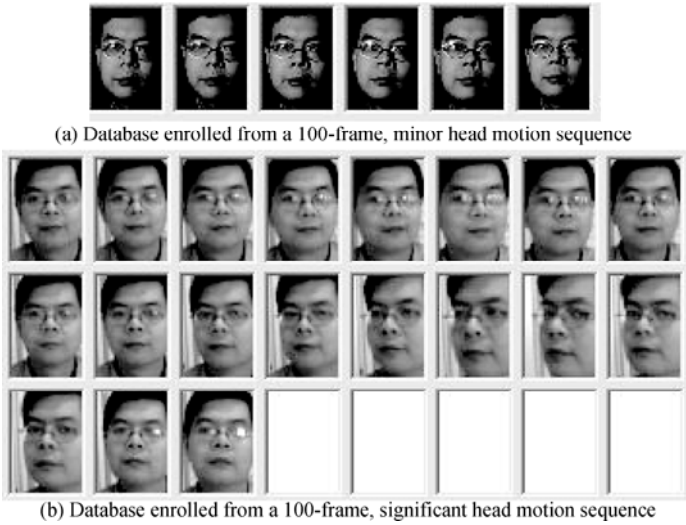
Fig. 8.16 Roll angle influence on the recognition classifier

8.4 Performance of Database Construction Algorithms

The features for an optimum face database have been introduced in Section 5.4. We evaluate these characteristics in this section.

Regarding rapidity and variety, we have made the following experiments. Several sequences with the same number of frames are taken. Some sequences are with minor head motion and others are with significant head motion. They are fed through the proposed automatic system. Fig. 8.17 lists the enrolled databases from two of such sequences. Only 6 out of 100 frame images have been selected by the adaptive update threshold in Fig. 8.17 (a), while 19 out 100 face shots have been selected in Fig. 8.17 (b). At the beginning of the database enrollment, rapid growth is of the most importance. Hence, the first several enrolled face shots are quite similar to each other while the following ones are in much difference. After a while, variety is more concentrated. That is to say, the selection threshold is increased. Since sequence (a) has much less head motion as well as other head variations than sequence (b), there are only few face shots

enrolled in (a) and much more enrolled in (b).



(a) Database enrolled from a 100-frame, minor head motion sequence

(b) Database enrolled from a 100-frame, significant head motion sequence

Fig. 8.17 Variety and rapidity of database construction

We continued the above experiments to test the updatability and uniqueness features. Five sequences have been fed through the system one after another. The sequences are recorded over a span of two years. Based on the constructed database shown in Fig. 8.17(b), we go on feeding through the five sequences. The system can automatically recognize the person and update the databases without errors. The new database is showed in Fig. 8.18. There are quite a lot differences in the source sequences. But the same person is always successfully recognized, demonstrating a satisfactory uniqueness.

The purity performance of the database corresponds to the false acceptance rate and will be covered in the following section.

Fig. 8.18 Updatability and uniqueness of database construction

8.5 Overall Performance of the Whole System

Although there are some standard databases available [5, 6, 7, 8, 9] to evaluate the image-based face detection and recognition algorithms, it is difficult to find freely available databases for testing the performance of video-based methods. Furthermore, since the research in automatic and unsupervised face recognition is still in its infancy, it is even much harder to assess its perform-

ance and make quantitative comparisons with existing methods. Therefore, we have made our own sequences for the overall system evaluation.

8.5.1 Online Version

As the first group of the experiments, we have fed the online TV news and online video from cameras with different resolutions.

Fig. 7.2 demonstrates examples of the test sequences from TV channels. The system had been running online for about several hours in one day and continued with several hours in the following day. As mentioned earlier, the processing speed of the whole system achieves about 1-2 fps. It missed many frames of the live video and therefore cannot learn faces showing up shorter than one second. But the people in small motion can be automatically recognized and enrolled. During the test, the system automatically learned twenty people and had no problem to recognize all the news reporters when they showed up in the news again. Moreover, there was no erroneous enrollment from different people. It can also be seen that although there was a large resolution difference and significant head pose variance of the images, each face was extracted without losing any useful facial information. It is common for TV programs to suddenly change the shot, especially in the news, which is fit for testing the system performance. The result shows no difficulties with the sudden sequence changes. Due to the speed limit, there are many useful face shots neglected and therefore the false rejection rate is high. But the harmful FAR keeps zero, which is quite satisfying.

8.5.2 Offline Version

For the offline version (shown in Fig. 8.19), we have asked more than 30 subjects for recording sequences in a span of more than two years. It is essential to always consider the conditions for creating such offline sequences. During the recording, we did not require any cooperation from the subjects and did not provide any manual operations. The only instruction is as simple as the following:

- Hi folks, there is an intelligent system which can automatically get to know your faces. You are invited to come to visit it and check if it can automatically remember you and recognize you by your revisit.
- You are supposed to be observable for the camera. There is no further requirement for your behavior during the visit. However, you are encouraged to behave facial changes, free head motions in all three dimensions, with or without glasses.

Therefore, the system was completely running passively and automatically. There were also big lighting variations from sequence to sequence. Moreover, we applied different cameras for capturing from time to time. It can be noticed that one camera is with poor quality that the captured image is a little vertically up-scaled, making the captured faces more different from what the other camera captured. But the camera differences are fit for further evaluating the recognition quality.

From the experiment, we can see that, there is also no FAR observed.

Fig. 8.19 Offline performance of the whole system

However, there is one failure. Person 000 and Person 002 have been falsely recognized as two persons. That is due to the same reason as described in Section 8.4. With further experiments, we have observed that, the failure is

due to its too short time of showing up. Person 000 stays within the camera vision for less than 1 second. Person 001 is then coming and Person 000 is going away. The next time when Person 000 shows up, quite different lighting conditions and head poses just occur. We have found out that, when the starting sequence of a certain person can provide more than 30 face shots, FRR can be kept reasonably low. This requirement is not highly restricted. It corresponds to around several-seconds period when a person is talking to someone.

8.5.3 Critical Assumptions

There are two main assumptions to be emphasized for the above performance analysis:

- The system focuses on dealing with videos instead of non-correlated random images. If randomly captured images with no video context are fed into the system, the proposed method could not achieve better performance than the classifier alone. This is reasonable since the proposed algorithms are expected to be used for video-based recognition systems and not as improved methods for image-based face recognition.
- IFD is assumed to work well at the very beginning stage for detecting a new face. For example, if a sequence starts with one face purposely rolled above 20°, there is no video information we could apply to detect and recognize it until IFD-detectable faces are showing up. That is to say, once a face is detected by IFD, the combined face detection and recognition procedure has no difficulty in tracking and recognizing such a face even it is rolled above 20°.

8.6 Summary

In this chapter, we have made experiments to evaluate the performance of face detection, face recognition, database quality, and the whole automatic procedure. From both the online and offline experiments, we have verified the robustness of the proposals against facial expression change, illumination variations and pose changes. It is important to note that, as just mentioned in

Section 8.2, the proposals of dealing with pose variations could fail in extreme cases. However, it is still successfully and fundamentally simulating the process of human brains.

We further implement another feature of human brain, as mentioned in [10]: the temporal information and face motion significantly contribute to achieving better performance of face recognition.

The automatic running of the system, in all cases reveals the robustness of the state machine, which guarantees the intelligent self-learning ability.

References

[1] R-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain: Face Detection in Color Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 24, No. 5, 2002, pp.696-706

[2] T. Kondo and H. Yan: Automatic human face detection and recognition under nonuniform illumination, *Pattern Recognition.* Vol. 32, Issue 10, 1999, pp.1707-1718

[3] M. LaCascia, S. Sclaroff, and V. Athitsos: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 22, No. 4, 2000, pp.322-336

[4] A freely available graphic tool, http://www.irfanview.com/, accessed 07 December 2006

[5] NIST Face Database, http://www.nist.gov/srd/nistsd18.htm, accessed 07 December 2006

[6] BioID Face database, http://www.humanscan.de/support/downloads/facedb.php, accessed 07 December 2006

[7] CMU/VASC image database, http://vasc.ri.cmu.edu/idb/html/face/index.html, accessed 07 December 2006

[8] AT&T face database, http://www.uk.research.att.com/facedatabase.html, accessed 07 December 2006

[9] MIT Face database, http://cbcl.mit.edu/software-datasets/FaceData2.html, accessed 07 December 2006

[10] P. Sinha, B. Balas, *et al.*: Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About, *Proceedings of The IEEE,* Vol. 94, No. 11, 2006

9 Conclusions and Future Directions

9.1 Conclusions

In this book, we have proposed a promising video-based face recognition system, which features the advantages of fundamental intelligent and robustness. The fundamental intelligence is achieved by two ways: self-learning and high-level analysis by intelligently fusing methods for recognition. The proposed system can start with an empty database and get to know the showing up faces of the people in an unsupervised way. When known people occur again, it can recognize them and automatically update the corresponding databases to keep up with recent views. Regarding robustness, experiments show that the system can deal with major challenging limitations in the state-of-the-art face recognition research: the aging problem, the difficulty of no better recognition quality in videos than in images, and the weakness of dealing with head pose variations.

We have also discussed the prospective applications and scenarios, in which intelligent and automatic face recognition could be highly demanded.

The main contributions of the algorithms for such an autonomous recognition system are further summarized in below:

- Novel combined face detection methods are proposed for improving both face detection and recognition rate from video. This is part of the stimulated algorithms to simulate the intelligence of high-level analysis.
- Combined same face recognition algorithms for face recognition from video. This is the second part of the stimulated algorithms on high-level analysis.
- Adaptive face database construction algorithms and database structures for

high quality face recognition. This is the part of the algorithms for achieving the self-learning intelligence.

- State machines contributed to the whole automatic procedure. This is the other part of the research for obtaining the self-learning intelligence.

9.2 Future Directions

We list in the following several proposals for the future exploration.

We have applied the specific image-based face detector and face classifier in our system. Thanks to its high-level intelligence, our proposed system should have no difficulty in integrating other kinds of classifiers. Further experiments are required to support this argument.

Regarding the applications of face recognition, we have not discussed in detail in the already existing applications or systems. In stead, one important purpose of discussing the novel applications is to broaden the visions of the readers. Those application proposals are mainly the conceptual ideas. How to realize the ideas is a prospective research direction.

In Section 5.4, we have mentioned the way of merging falsely separated databases that are from the same person. Since it does not frequently occur if a starting sequence of a new person is not extremely short, the current implementation does not include this function. The merge may also theoretically produce additional false acceptance rate. Further research in this direction could be promising.

Another proposal is also related to the database. Although the adaptive updating threshold decreases the redundancy of the databases, it is still to be improved. For example, the rapid growth of a new database produces the redundancy at the beginning of the database construction. In current implementation, the redundancy remains until the face shots are replaced through the update procedure. In principle, the redundancy can be earlier decreased.

In this book, we are mainly concentrating on face recognition. But the proposed rules and algorithms can be more widely applied for any other biometric recognition techniques which require automatic and self-learning characteristics. A wider range of the application can be further explored.

Index

G

H

I

K

L

M

N

O

P

R

S

T

U

V