Autonomous Face Recognition

Dissertation

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
der Fakultät für Ingenieurwissenschaften
der Universität Ulm

von

Mou, Dengpan

aus China

1. Gutachter: Prof. Dr.-Ing. Albrecht Rothermel
2. Gutachter: Prof. Dr. Heiko Neumann

Amtierender Dekan: Prof. Dr.-Ing. Hans-Jörg Pfleiderer

Datum der Promotion: 22. August, 2005

2005

Abstract

As a booming technology, face recognition has been studied for many years and is expected to be widely used in daily identification systems, communication systems, public security systems, and in law enforcement systems.

Most state-of-the-art machine learning systems are based on the supervised learning theory and image processing techniques, which require separate pre-training procedure for enrolling every new face and updating existing faces. Therefore, an additional human supervisor is normally required. Users' cooperation is expected as well during the training phase. However, a human vision system is far more intelligent. It has no difficulty to automatically memorize the faces they have interacted with for future recognition. All the enrollments, updates, and comparisons have been done completely in the brain without any outside assistance. Although the biological reasons behind it are not clear until now, it is not hard to imagine that the brains can combine all information that is useful for recognition, including "image processing", video context, logic deduction, experiences etc.

Inspired from the human vision system, we combined the conventional learning algorithms and image processing algorithms with predefined rules to increase the intelligence of machine recognition systems. As the first step, face detection is implemented by an industrial image-based face detector combined with novel temporal differencing algorithms. The face detection result, an industrial image-based classifier, temporal filtering and video context related rules are all combined for face recognition. The database can be constructed online and be adapted automatically according to the update rules. State machine is introduced to keep the system running automatically and stably.

The major feature of the system is self-learning. No separate or pre-training is required. It can start with an empty database and get to know the faces of the people showing up in an unsupervised way. When known people enter again, the system can recognize them and adaptively update the corresponding databases to keep up with recent views. The proposed system can find promising applications in many fields especially for consumer electronics.

Acknowledgements

My first and deepest gratitude is to my advisor, Prof. Dr.-Ing. Albrecht Rothermel, for his invaluable supervision and support. His open-minded way of thinking, continuous encouragement and trust make me feel confident to create novel research ideas. His depth of knowledge, brilliant mind and penetrating insight into interdisciplinary research guide my work to be always kept on the right track.

I am deeply grateful to Dr. Rainer Schweer and his group from Deutsche Thomson-Brandt GmbH for providing a wide research topic and for giving me a free hand to explore promising research directions. Moreover, through stimulating discussions, I have been benefited so much from his extensive industrial research experiences.

I would like to express my warmest thanks to Prof. Dr. Heiko Neumann, for his commitment to be the second referee for this dissertation. His earnest review and comments are crucial to shape the thesis.

I am also profoundly indebted to Prof. Dr.-Ing. Hans-Jörg Pfleiderer, the dean of the Engineering School and the head of the Microelectronics department, for providing a superb research environment.

My special thanks also go to all members of the video-group: Dr.-Ing. Roland Lares, Martin Lallinger, Ralf Schreier, Karsten Schmidt, Thomas Kumpf and Christian Günter for their friendly help and inspirational discussions and comments. I wish to thank Lin Wang and Fei Fei for their contributions to the software interface implementation.

My deep acknowledges go to the following current and previous colleagues and friends: Dr.-Ing. Ralf Altherr, Dr.-Ing. Markus Buck, Dr.-Ing. Sviatoslav Bulach, Markus Bschorr, Zhen Chang, Tiefeng Chen, Shan Chen, Turgut Dogan, Richard Geißler, Roland Hacker, Frank Hagmeyer, Stefan Hirsch, Cheng Miao, Oliver Pfänder, Ivan Perisa, Markus Prokein, Xavier Queffelec, Wolfgang Schlecker, Walter Schweigart, Lei Wang, Yi Wang, and Chi Zhang for their generous help to collect the face databases.

I am so grateful to the department secretary, Ehrentraud Höfer for her excellent administrative work, to the system administrator Markus Prokein and Walter Schweigart for their great technical supports.

In particular, I thank my dearest parents for their endless love. I am so lucky as well that, from childhood, I have been influenced by their strong philosophical background. Without these emotional and mental supports, the achievement of the thesis could not be possible.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Face recognition has been studied for many years and is expected to be widely used in daily identification systems (e.g. automatic banking and access control etc.), in communication systems (e.g. teleconferencing and video-phone etc.), in public security systems (e.g. criminal identification, digital driver license etc.) and in law enforcement systems.

Face recognition in a perception system from human beings seems instinctive, but it is really a tough and complex task for a machine-based system. Although the psychophysical and neurobiological background behind human face recognition has been long studied by scientists, their contributions to the mathematical models and engineering solutions for a machine vision system are not satisfying. Therefore, researchers from the computer science are constructing vast numbers of mathematical models and dedicated algorithms, which may cover the field of artificial intelligence, statistical learning, image processing and even video signal processing. We will briefly overview popular methods in chapter 2.4.

Machine vision systems 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 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 manage it during the monitoring. And in this case probably all people under supervision would prefer a machine than a human supervisor. Nevertheless, in terms of recognition precision, even the most successful state-of-the-art face recognition technology in general is still far from the level of our 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 many 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 current machine recognition techniques 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 consumer electronics. Quite recently, there are some advances ([1], [2], [3]) 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 ([4], [5]) 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 features the following:

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

In this dissertation, we are exploring the ways and algorithms to approach such an ideal system, especially for the application for consumer electronics.

1.2 Proposed Approach

The rule-based algorithms for computer vision and pattern recognition used to be popular more than fifteen years ago. For example, to detect faces, people used eyes, noses, mouth 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. Learned 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 dissertation, 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.

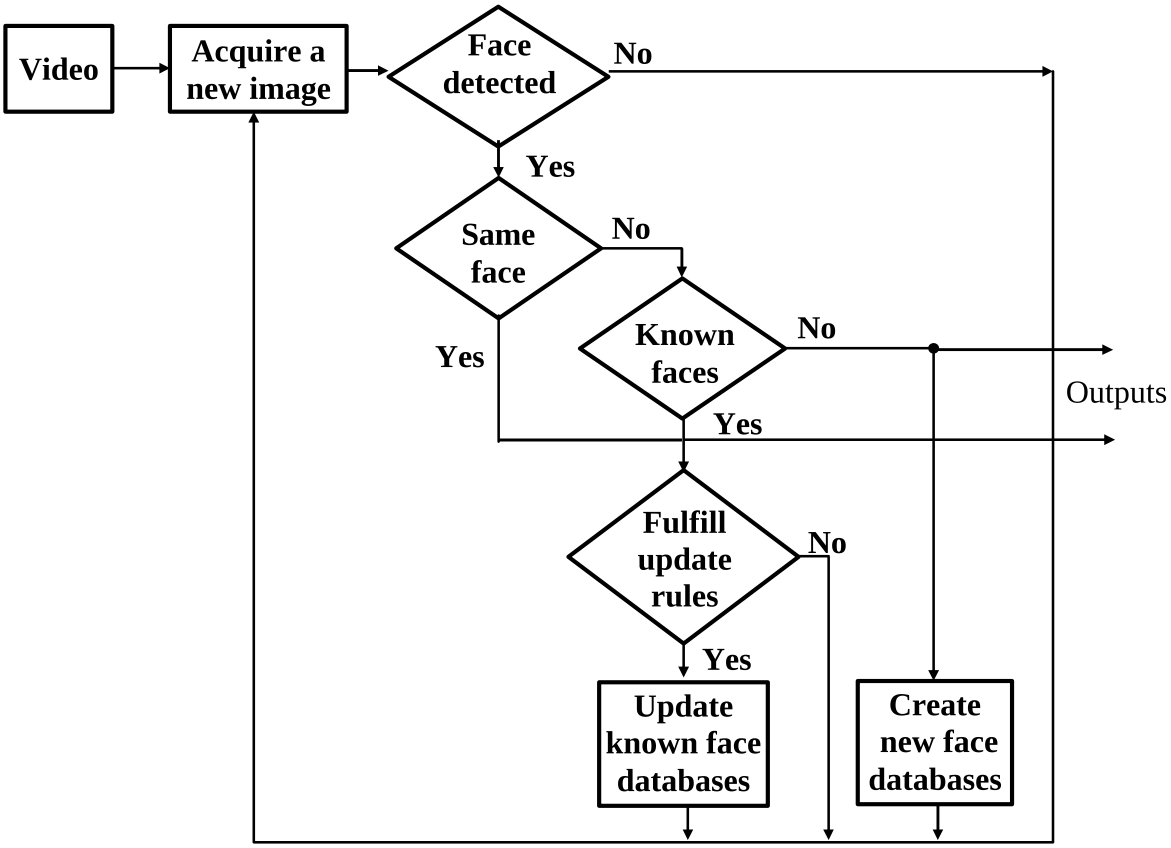


Figure 1.1 Overview of the System Functional Blocks

For the convenience of describing the proposed approaches, we list the overview of the functional blocks of the automatic system ([6]) in Figure 1.1.

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 detection techniques (mostly learn-based) can be applied here. The temporal information [7] derives from the limited day to day human moving speed, a reasonable processing speed assumption and human face dimensions. 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 en rolled, 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 ([7], [8]). The implemented prototype shows robust performance for consumer applications [9].

1.3 Prospective Applications

The proposed non-invasive and unsupervised face recognition system can find prospective applications in consumer electronics. [10] describes a passive person identification system which can automatically memorize frequently visiting customers and welcome them. [11] 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 Home Security

The first one is a home security system which can be mainly integrated on a TV set. 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 telephone calls or internet connections. 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 the alarm message as well as the mugshot of the stranger. It is also possible to obtain a higher resolution mugshot through an email. 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 very crucial to keep a low false alarm rate. For a convenient use, 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.

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 of being stolen, setting of drivers' 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 engine the motor, but also for keeping the intelligent system in a normal mode. When some one 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 Entertainment

A third promising application might be in entertainment, e.g. as a part of an electronic pet. It is shy to strangers but can learn to get to know him with a long enough interaction. It can show some friendly expressions or poses to welcome a known person. [12] 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 be personally on the scene. When multiple users are playing the game, automatic face recognition is absolutely required.

1.3.4 Mobile Phone

Mobile phones are more and more widely used. Many parents would like to have their children taking it for security reasons when 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 Dissertation Outline

The remainder of the dissertation is arranged as follows.

Chapter 2 introduces the background and related research in biometric recognition, face detection and face recognition. We have proposed new ways of classifying face detection and face recognition methods and procedures. The state-of-the-art research of video base-face recognition and unsupervised recognition systems are reviewed in more detail.

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 firstly 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.

Chapter 9 summarizes the contributions of the dissertation and explores the future research directions.

Chapter 2 Background and Related Research

2.1 Biometric Recognition

As a booming technology, biometric recognition of a person has been successfully applied in many security systems. Typical biometric technology can be defined as a 3-step procedure: enrollment, formatting and matching. For enrollment, the users have to cooperate to follow some instructions from a supervisor. The biometric data are then acquired and stored as a certain format in the formatting step. We can denote the data as templates or databases. Matching is actually the comparison between the live biometric data and the databases.

We list several biometric recognition methods in Table 2.1 with the comparison of three important features.

Table 2.1 Comparison between different biometric recognition techniques

Biometric recognition methods	Robustness	Invasiveness	Usability
Iris	+ +	–	– –
Retina	+ +	–	– –
Fingerprint	+ +	– –	–
Voice	–	+ +	+ +
Face	+	+	+

Iris, retina and fingerprint are the most precise techniques for human identification. They are usually used for the systems where the recognition quality is the most important factor and all the other issues such as cost, usability etc. can be negligible. Legal applications or military systems are the two major application areas. But the critical problems are their usability in daily life. The users have to be extremely close -- around 0~30 centimeters to the capturing system in both enrollment and matching procedures. They are therefore too annoying to the users as well. And for iris and retina, a users’ head has to keep stable for a certain time until his/her identification is verified. Fingerprint is a little bit easier for application but it’s the most invasive technology since it is conventionally used for criminals.

Voice is convenient of use and can be passively applied, but the vocal characteristic for a person is changing too much from time to time. So it is normally used for speech recognition.

For consumer electronics, we are more interested in an identification system that ranks high for all the parameters. Face recognition is appealing because it promises high accuracy, requires less user cooperation compared to the iris and retina methods, and is less invasive and is potentially fast and cheap.

In the following, we mainly discuss the person identification by using facial characteristics.

2.2 Face Detection

Face detection is the first and crucial step for any face processing system. 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. 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 [13] and [14]. The categories of face detection methods are summarized in Figure 2.1.

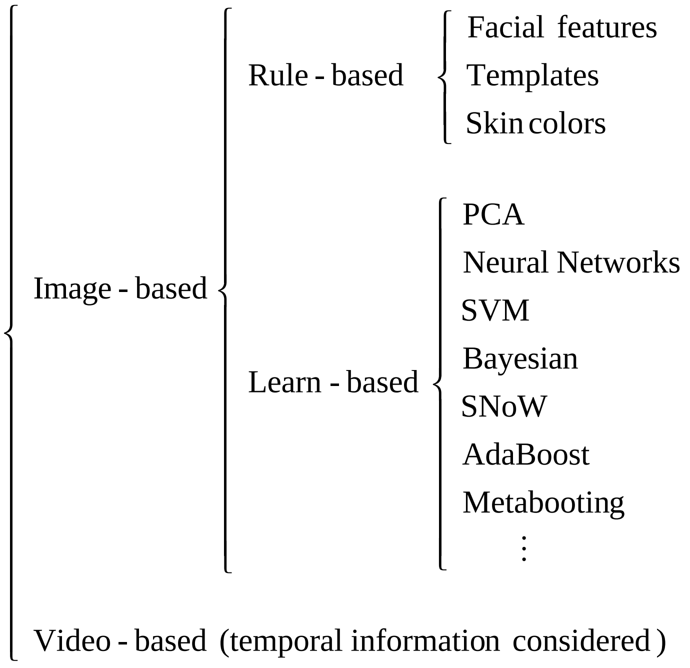


Figure 2.1 Categories of face detection methods.

As an expanding hot topic, numerous algorithms are developed for image-based face detection in the last 20 years. [16] defines the image-based face detection approaches as two groups (see chapter 5.1 of the book [16]): local feature-based and holistic-based. But skin-color approach is not included according to the grouping method. [14] classifies the image-based face detection methods into four categories: 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.

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 [17], 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 [18] 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 [19], [20], 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. Multi resolution, sub templates 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 in [10] for both detection and recognition and further developed in [21]. Neural networks are successful applied in [22], [23] with exciting results. [24] proposes the support vector machines (SVM) algorithm to classify face and non-face patterns. Bayesian approach is explored in [25] with wavelet representations. Sparse network of Winnows (SNoW) is used in [26] to specifically learn very large numbers of features. Recently, [27] applies AdaBoost learning algorithm for real-time frontal face detection and is extended to multi-view detection in [28]. A detector-pyramid architecture is proposed in [29] for fast multi-view 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 differencing and background subtraction are two main cues to detect motion. Some video-based systems even apply the motion information alone to detect faces. [3] subtracts the current image from the initial background to detect a face. This method, however, requires a stable background with no face. [30] and [31] 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 ([10]). 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 ([18], [10], [32], [33]) 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 [14], 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.

2.3 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 popag*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 are claimed in [34], [35], and [36]. But the main problems are their weakness in illumination changes. The combination of different approaches is proposed with improved results. [37] 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. ([38]) 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.

2.4 Face Recognition

2.4.1 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 and in principle can only 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 [30] in 1991, where an eigenface-based system of detecting and recognizing faces was demonstrated. A large number of algorithms have been developed since then. [39] in 1995 and [4] in 2002 are two important surveys of machine-based face recognition. Much of our discussion in this section (Chapter 2.4) is stimulated by the two papers. Additionally, [40] is the most recent paper that explores the 3D face recognition approaches.

The public available FERET (*F*ace *Re*cognition *T*echnology) database and its protocols ([41], [42]) 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--Aug94, second--Mar95, and third--Sep96 (administered in Sep. 1996 and Mar. 1997). The tests were sponsored by U.S. government and appealed many university research groups. The tests do not only provide the possibility to objectively evaluate different face recognition algorithms under almost real-world situations, but also clearly propose 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 two FRVTs (Facial Recognition Vendor Test): FRVT 2000 [43] and FRVT 2002 [5], both 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 [43]. 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, and 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 low-pass filtering probes could increase performance.

FRVT 2002 report [5] 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 ([44], [45]) 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.

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 or 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 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:

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

2.4.2 Recognition Procedures and Methods

[39] 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. Figure 2.2 summarizes the 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, and 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 deemed 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. So it is really a challenging problem for a machine system to manage it by itself.

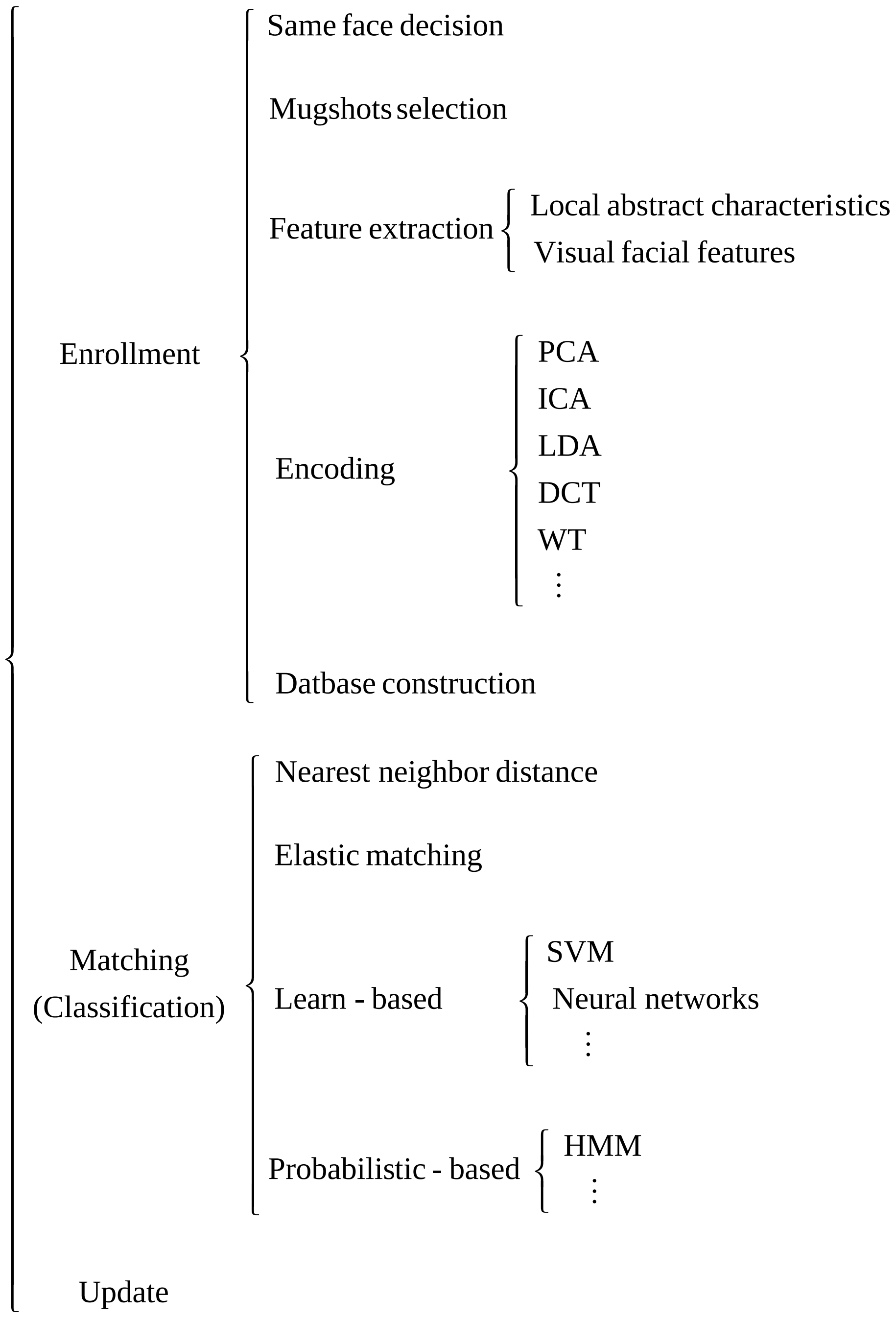


Figure 2.2 Category of face recognition procedures and methods.

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 expression. 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 2.4.4.

Feature extraction and encoding are the third step and fourth step respectively in the enrollment, on which most 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 ([46], [47]) of human perception systems show that both holistic and feature information are crucial for the perception and recognition of faces [4]. 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×n pixels is normally represented by a m×n matrix with each element corresponding to the intensity value of each pixel. The matrix is also called an image space with m×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 [30], which is based on principal component analysis (PCA), independent component analysis (ICA) [48], linear discriminant analysis (LDA, known as Fisher discriminant analysis—LDA) [49], [50], discrete cosine transformation (DCT) [51], [52], [53] and wavelet transformation (WT) [54], [55], [56] are the major well-known encoding methods.

The main idea of PCA is trying to project a m×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” [30]. 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. There are contradictory arguments of whether

PCA or ICA is the better for face recognition, [57] is the most recent paper we could find that more completely compares both methods.

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 [58] 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, non-linear (kernel-based) approaches are introduced quite recently: Kernel-PCA (KPCA), Kernel-ICA (KICA) and kernel-LDA (KLDA/KDA) [59], [60]. There are few papers that make comparison among all PCA/KPCA, ICA/KICA and LDA/KLDA. A recent one we could find is [61], but some of the conclusions don’t agree with [57]. Further explorations are required to tell which one in which case is better. But they are beyond the scope of this dissertation and we are just putting forward this question as a research challenge.

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 [51], [52], [53] thanks to its computational efficiency and ease of implementation both in hardware and in 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.

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 can be taken from pattern recognition theory. 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 [30].

Elastic matching approach together with wavelet transformation is proposed in [55]. It is a more complex method which runs heuristically to find the image graph which maximizes the graph similarity function.

Support vector machines (SVM) [62], [63], [64] and neural networks [65], [66] 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) [51], [52] are successfully applied as face recognition classifiers and [54] 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 most recent FVRT 2002 test [5], 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 have an additional human supervisor, who makes the decision on when and which mugshot to be updated.

2.4.3 Video-based Recognition

There are two main groups of research for face recognition: image-based and video-based. Recognition from video attracts much 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 section, 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 the two and give an overview of the state-of-the-art of research in video.

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 [5].
- **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. 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 U.S.A. 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 [30], [55], although they are using video sequences as inputs, mere 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 [4]. We explore the more recent research here.

[67] proposes adaptive Hidden Markov Models inspired by speech recognition. It requires a supervised enrolment procedure and an unsupervised 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.

[68] 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 [69]. PCA and LDA are chosen as pure image-based methods while HMM [67] was 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.

[70] 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.

[71] 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-mentioned 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 is that every sequence consists of only one person at one time, without moving background objects, which is still far from the complex environments in real world.

2.4.4 Unsupervised and Automatic Systems

The so far discussed face recognition technology requires a separate enrollment (training) step, where face images are collected, selected and labeled manually. 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 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 chosen 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 database 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. [72], [1], [73], [12], [3], [2] are the important scientific papers we could find from state of the art.

[30] 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.

[74] 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 (actually only the verification task here) task is accomplished by comparing with existing face databases. This system is tested on three image sequences: the first was 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 were 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 not a fully automatic system in

broadly speaking. 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 by [72]. 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 try 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.

[1] is an extension of [72]. 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 [72].

[73] 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 identified 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.

[12] presents a humanoid robot which can silently and passively get to know new people who have interactions with it. The system is based on the face detection method in [28] and the face recognition method in [30]. 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 [3]. 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.

[2] is the most recent paper we could find. It 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 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 [75] 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.

[10] is a more recent patent which describes a passive system for biometric customer identification and tracking. 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 welcome frequently 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.

[11] is a recently published patent application. An adaptive facial recognition system and method is demonstrated. It is a more detailed system than [10], which uses 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. [76] 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 [75], the proposed ideas are to be performed with a much higher speed.

In general, most current automatic systems have several common limitations.

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 high efficient and high quality databases is far from adequate.

Chapter 3 Combined Face Detection and Tracking Methods

3.1 Introduction

This chapter explores how to automatically extract face(s) of interest from live video input. It is the first but crucial step for an automatic and unsupervised face recognition system.

As mentioned in chapter 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 [3], it is normally used in the preprocessing step ([77], [32], [33]) 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 [78], 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. Although MFD is computationally simple, which is mainly based on the daily people moving speed assumption and frame differencing, it can significantly improve the false negatives. Figure 3.1 illustrates the procedure for face detection.

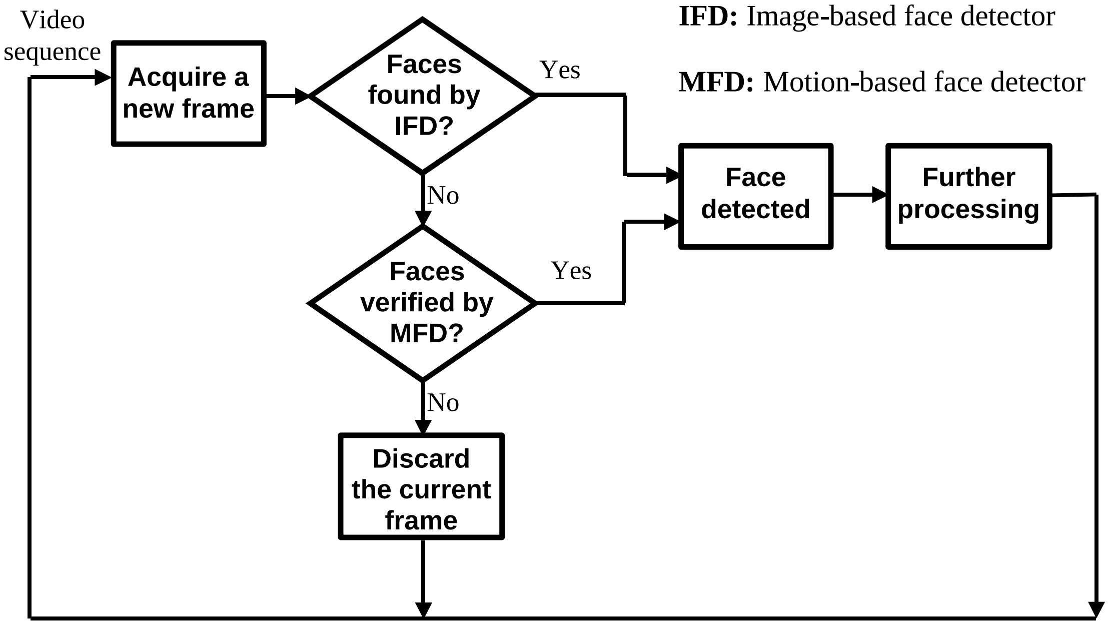


Figure 3.1 Function blocks of face detection

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 [5]. 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 processing 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 [28] and [29].

3.2.2 Overview of the Detection Algorithm

It is a rule-based face detection method by using a predefined face template, according to the face detection categories described in chapter 2.1. The following two paragraphs briefly introduce the detection algorithm from the publicly provided information in [79].

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 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

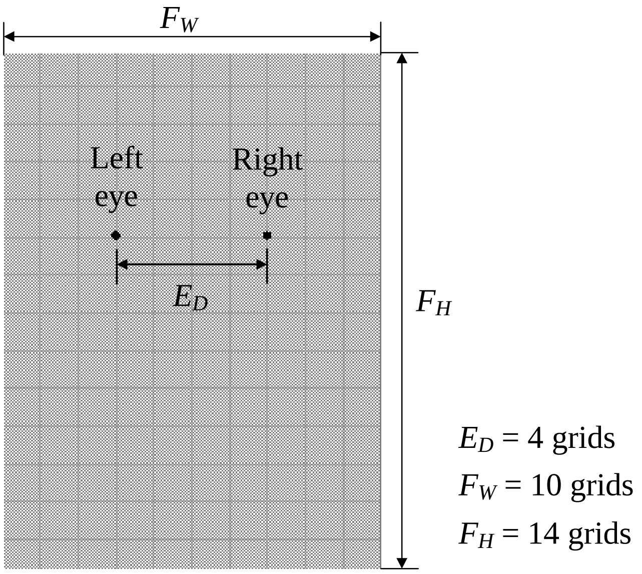


Figure 3.2 Proportional illustration of face region depending on the eye distance.

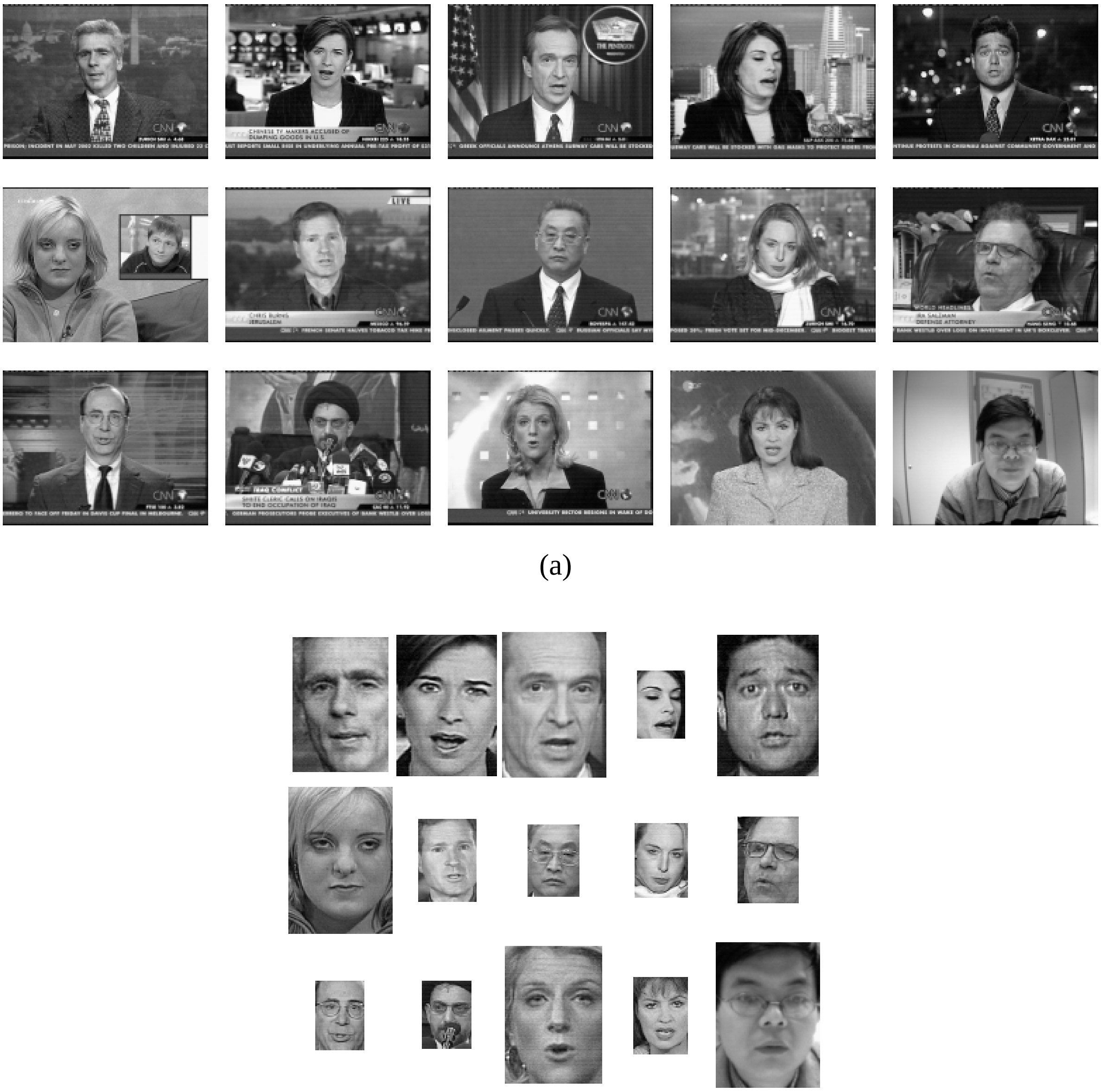
The study of anatomy ([80]) repeals that most human faces follows a certain range of width-height ratio, which can be also represented by the eye distance. In this thesis, we empirically apply the following equation to extract the face region from detected eye positions.

$$\begin{aligned} F_W &= 2.5E_D \\ F_H &= 3.5E_D \end{aligned} \tag{3.1}$$

where E_D is the eye distance, F_W and F_H are face width and face height respectively.

Figure 3.2 uses grids to proportionally illustrate the relationship between the estimated face height, face width and the detected 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 mentioned image-based face detection method to obtain the eye distance of a certain face. Face regions are then extracted by equation (3.1).



(a)

(b)

Figure 3.3 Example of face extraction

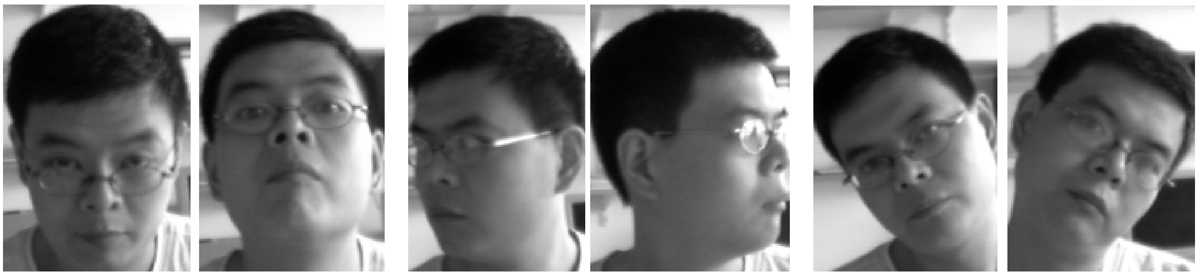
Figure 3.3 shows examples of the extracted faces and the thumbnails of the corresponding original images. For the sake of displaying, the whole images are re-sampled to the same resolution while the face regions are shown in their original size. Thumbnails re-sampled from TV programs and video cameras 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.

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. Figure 3.4 shows examples of those pose changes.



(a) Pitch faces (b) Yaw faces (c) Roll faces

Figure 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 8.2. Although relatively robust 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 Chapter 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 [16], 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 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 Figure 3.1 and can be further decomposed into four sub-blocks as shown in Figure 3.5.

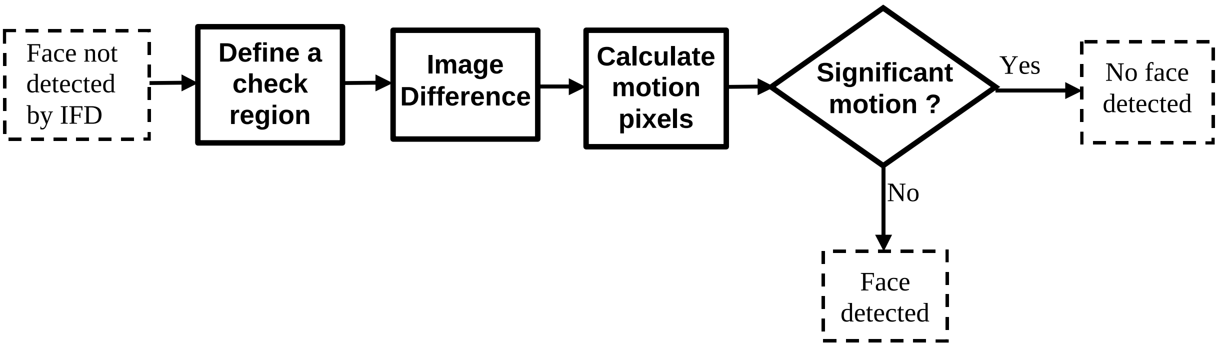


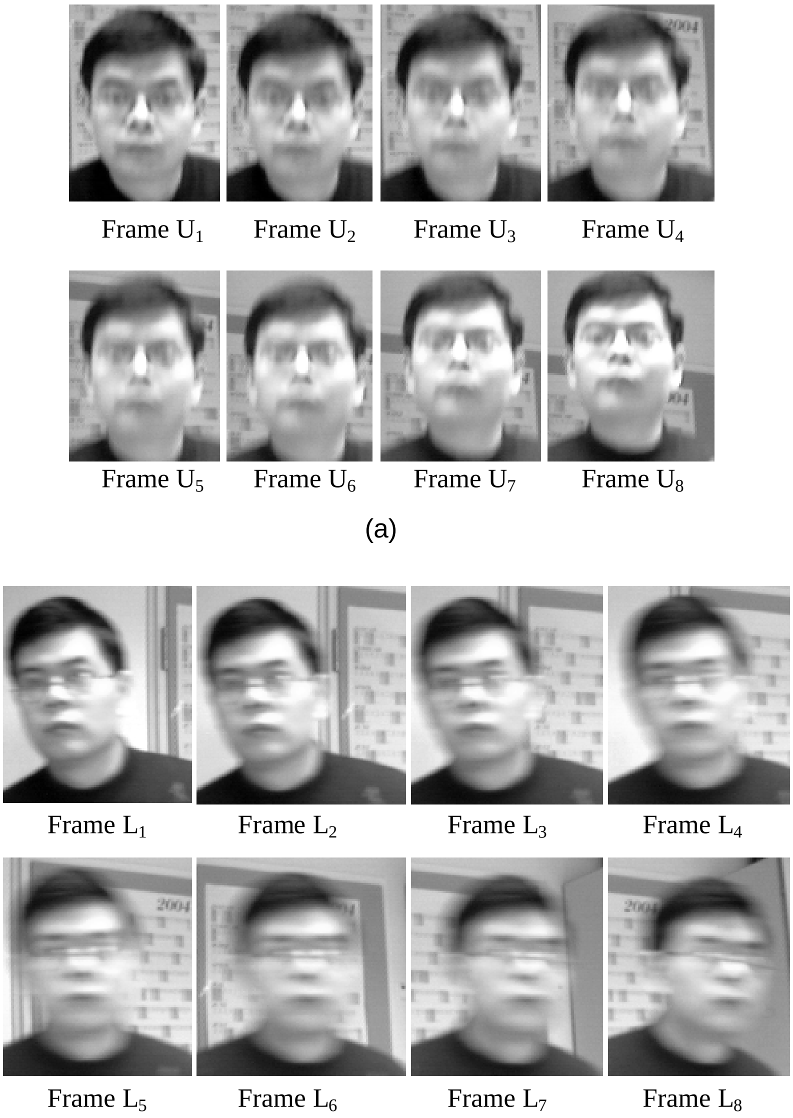
Figure 3.5 Function blocks of temporal-based face detector

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 110cm/s~140m/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 [81] finds out that the vast majority of adults have interpupillary distance in the range 5.5cm~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:

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

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.


Frame U_1 Frame U_2 Frame U_3 Frame U_4

Frame U_5 Frame U_6 Frame U_7 Frame U_8

(a)

Frame L_1 Frame L_2 Frame L_3 Frame L_4

Frame L_5 Frame L_6 Frame L_7 Frame L_8

(b)

Figure 3.6 Two examples of face region fast movement.

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

$$M_F \leq 20E_D \cdot ?t \tag{3.3}$$

where M_F denotes the maximum distance that a face can move between two successive frames and *?t* 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 Figure 3.6 are the main face regions cut from the original images. (a) lists the consecutive frame images with a person moves the head up and down as fast as possible. (b) shows the consecutive frame images with a person moves the head left and right as fast as possible.

Table 3.1 Comparison of M_F and actual speed in the experiment shown in Figure 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		
	Frame L_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. MF 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 ?*t* equals to 1 second, 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 ?*t* 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 ([27], [28], [29]) strongly support the above assumption. Equation (3.3) is converted as the following:

$$M_F \le E_D \tag{3.4}$$

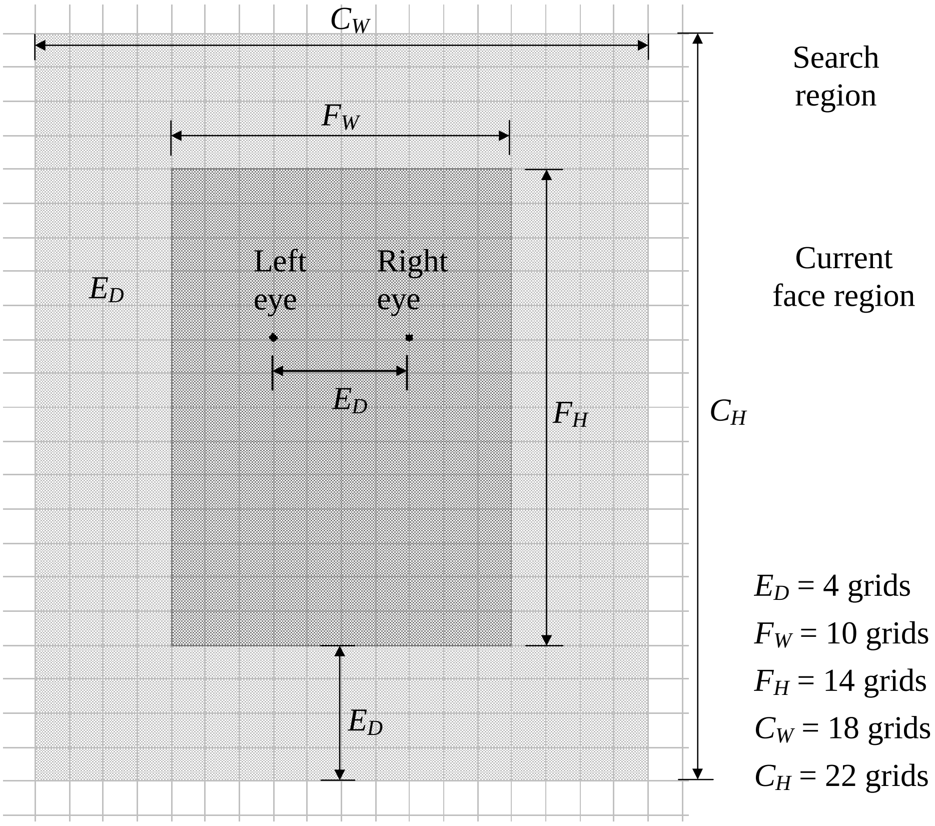


Figure 3.7 Proportional illustration of face region and search region

Figure 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:

$$\begin{aligned} C_W &= 4.5E_D \\ C_H &= 5.5E_D \end{aligned} \tag{3.5}$$

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:

$$\begin{aligned} A_C &= 4.5E_D \times 5.5E_D = 24.75{E_D}^2 \\ A_F &= 2.5E_D \times 3.5E_D = 8.75{E_D}^2 \\ A_C &\approx 2.8A_F \end{aligned} \tag{3.6}$$

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

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 Figure 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 $I_D(x, y)$ is the Nth frame of a certain sequence.

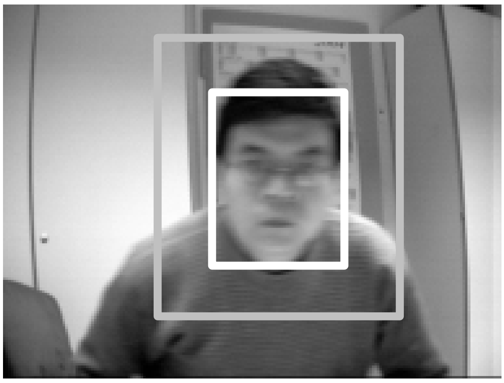
The following equation is used to calculate the motion pixels:

$$
\begin{aligned}
I_M(x,y) &= \begin{cases} 1, & if \;\; I_D(x,y) \geq T_M \\ 0, & if \;\; I_D(x,y) < T_M \end{cases} \\
N_D &= \sum I_M(x,y) \\
M_P &= \frac{N_D}{N_S} \times 100\%
\end{aligned}
\tag{3.8}
$$

$I_M(x, y)$ 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.

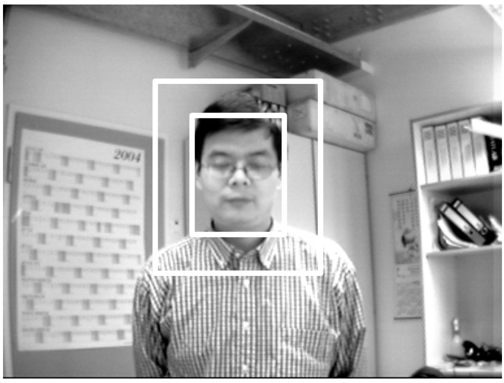
There is a dilemma for the selection of T_M. On one hand, it should not be too small to be influenced by noise. Intensity changes due to noise can vary from one pixel to even more than 10 pixels. On the other hand, it makes no sense if too big value of T_M is chosen. With an extremely big T_M, motion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

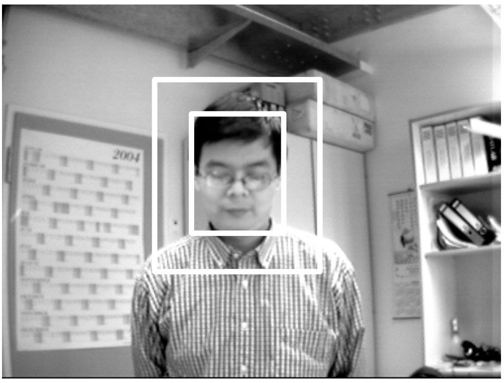


(a1)

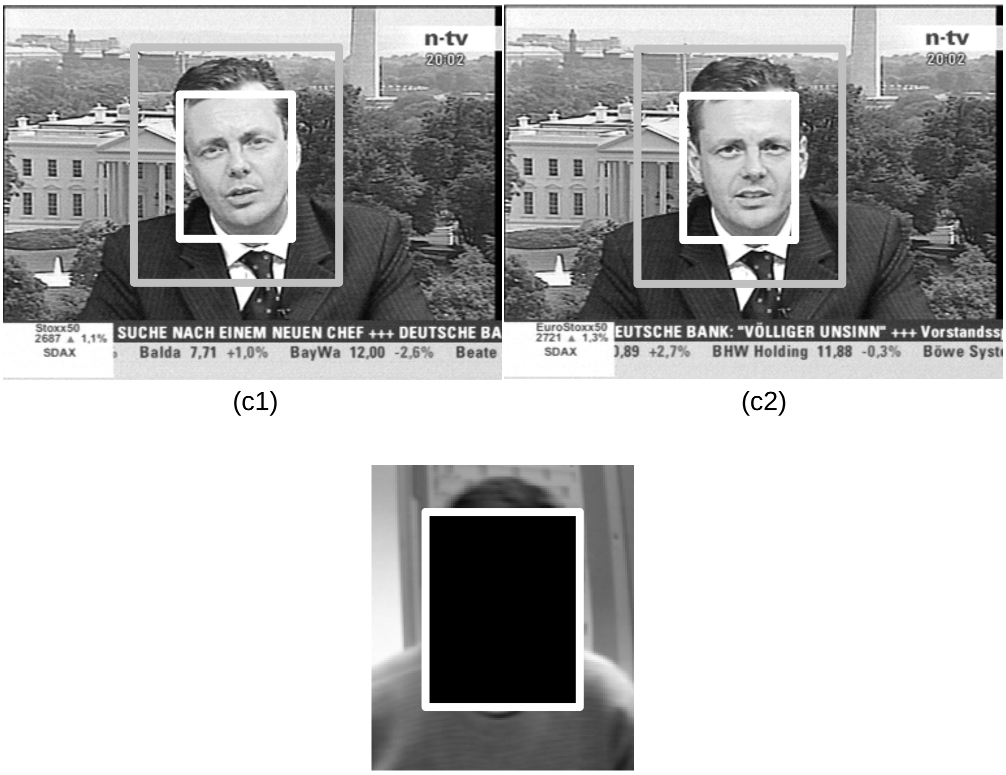
(a2)



(b1)



(b2)


n-tv
20:02
Stoxx50
2687 ▲ 1,1%
SDAX
SUCHE NACH EINEM NEUEN CHEF +++ DEUTSCHE BA
Balda 7,71 +1,0% BayWa 12,00 -2,6% Beate
n-tv
20:02
EuroStoxx50
2721 ▲ 1,3%
SDAX
EUTSCHE BANK: "VÖLLIGER UNSINN" +++ Vorstandss
),89 +2,7% BHW Holding 11,88 -0,3% Böwe Syst

(c1) (c2)

(d)

Figure 3.8 Example images to find the range of T_M.

Figure 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).

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 Figure 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 Figure 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 Figure 3.9. We can estimate the motion pixels from Figure 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 (3.6). When there is a sudden change in the face region, there are roughly 30%~40% pixels of the search region in motion.

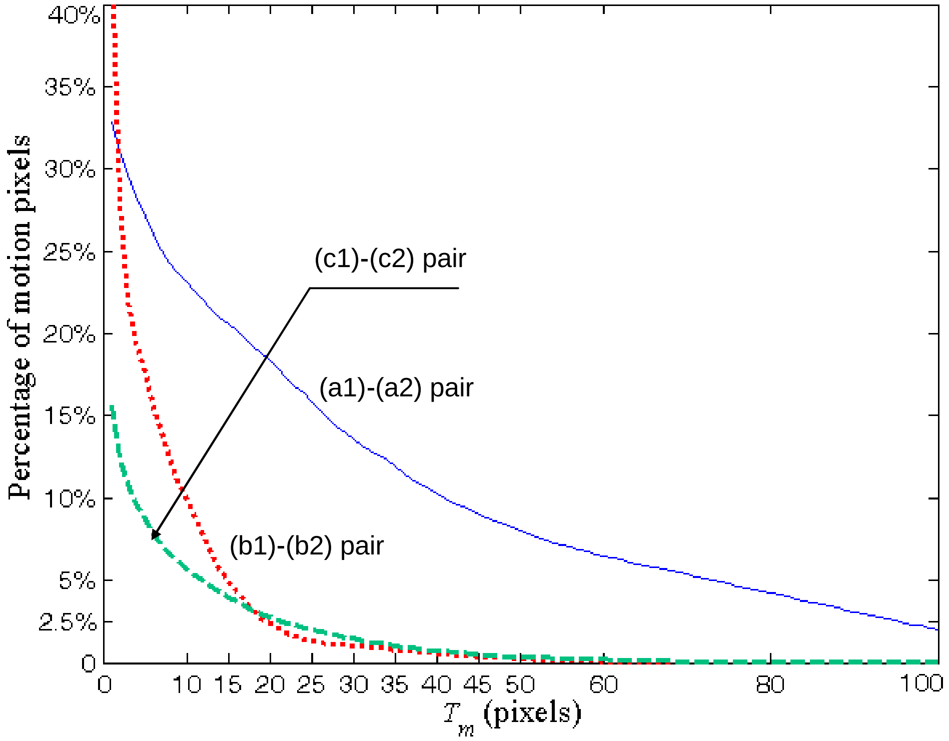


Figure 3.9 Percent of the motion pixels in the search region (face region excluded)

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).

$$
\begin{aligned}
& I_M(x,y) = \begin{cases} 1, & if\ I_D(x,y) \geq T_M \\ 0, & if\ I_D(x,y) < T_M \end{cases}, with\ T_M \in (15,35) \\
& M_P = \frac{N_D}{N_S} \times 100\%,\ with\ N_D = \sum I_M(x,y) \\
& \text{Face verified by MFD} = \begin{cases} True, & if\ M_P \leq M_{th} \\ False, & if\ M_P > M_{th} \end{cases},\ with\ M_{th} \in (30\%, 40\%)
\end{aligned}
\tag{3.9}
$$

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 chapter 8.2, we will elaborate the detection performance analysis. But 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 chapter 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. Chapter 4.3.3 provides such a solution.

3.5 Further Discussions

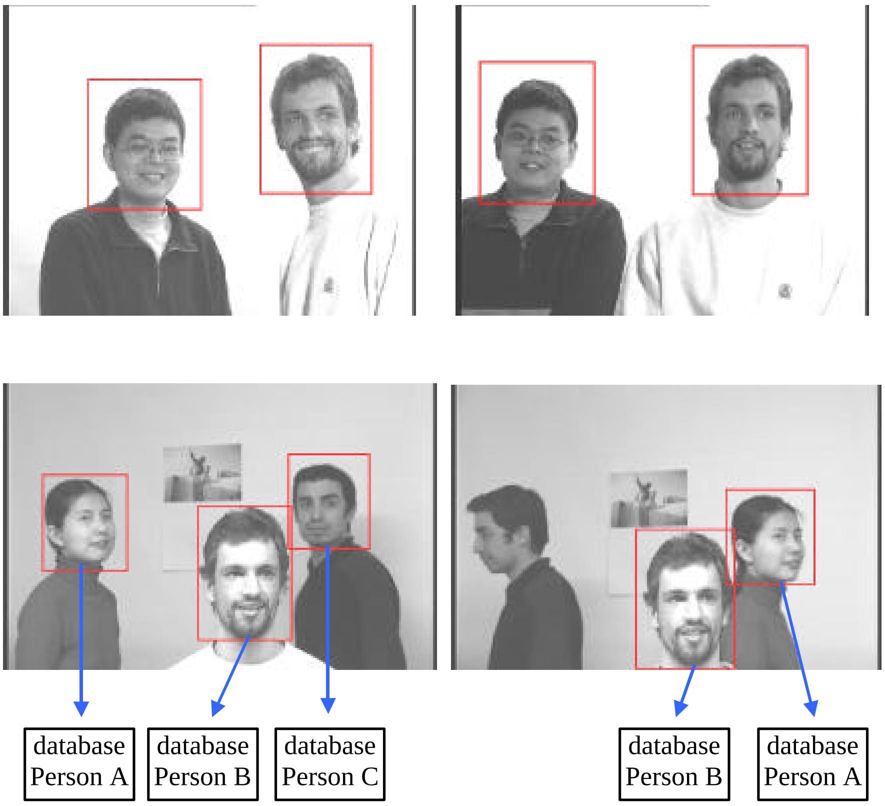


Figure 3.10 Sequences of multiple people

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 Figure 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 Figure 3.11. This method and is a workaround for the specific IFD. Therefore, we do not intend to discuss it in detail here.

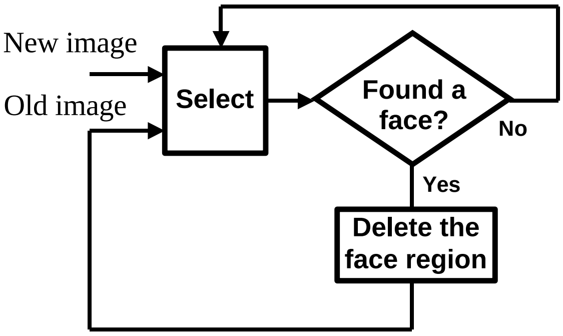


Figure 3.11 Proposal for the multiple IFD

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

Chapter 4 Automatic Face Recognition

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 Chapter 2.4.2, 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.

Feature extraction and encoding methods are taken from existing techniques which will be firstly and briefly covered. Then the matching algorithms are described in detail, which includes an image-based classification method, an adaptive way of selecting the thresholds for the classifier, and a video-based temporal filtering method. Finally, the combined same face decision algorithms are discussed. Each single method is running passively which guarantees the autonomous procedure of recognition.

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 [79].

A face image is firstly divided into an M×N block, as shown in Figure 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 Chapter 2.4.2. 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×N number of vectors are accordingly determined to make a vector set, as illustrated in Figure 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. Figure 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.

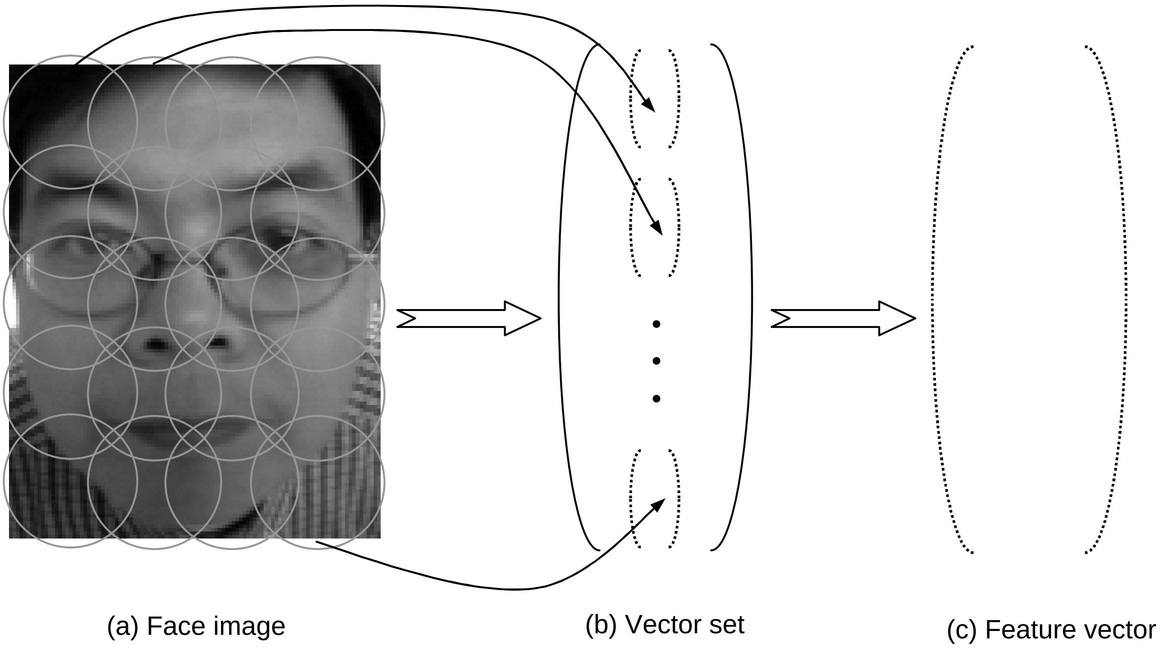


Figure 4.1 Feature extraction of a face region

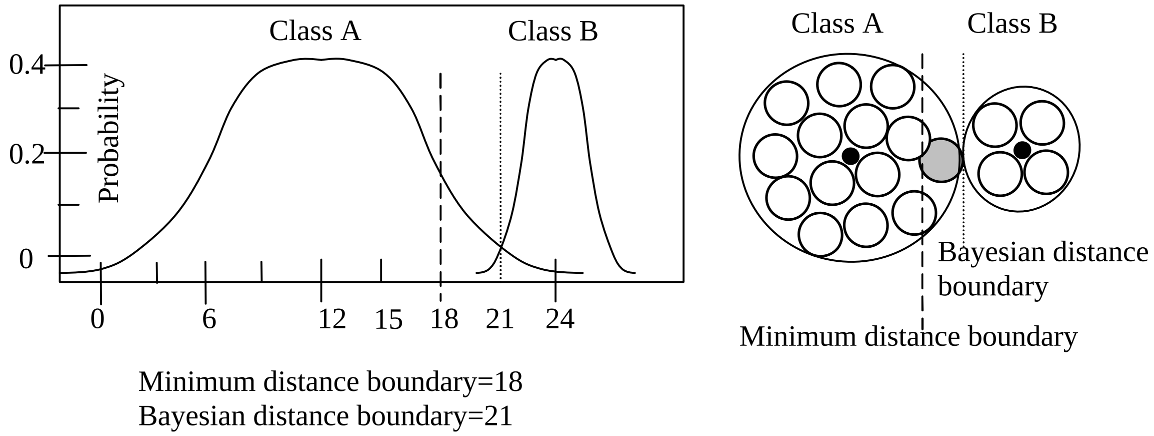
The encoding algorithms from FaceVACS are not open to the public. The procedures we developed are independent from those encoding algorithms.

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 the next chapter.

Suppose that there are already face databases available, the next step after feature extraction is matching. Any classification method as discussed in chapter 2.4.2 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 [5]. 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.


Class A
Class B
0.4
0.2
0
Probability
0
6
12
15
18
21
24
Minimum distance boundary=18
Bayesian distance boundary=21
Class A
Class B
Bayesian distance boundary
Minimum distance boundary

(a) (b)

Figure 4.2 Comparison between Bayesian distance and center distance

Among many distance measure methods, minimum distance and Bayesian decision are more frequently applied in face recognition. Figure 4.2 sketches the differences between them. As shown in (a), while minimum distance method only measures the Euclidean distances, Bayesian distance additionally considers the probability distributions of each class. (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.

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 Figure 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.

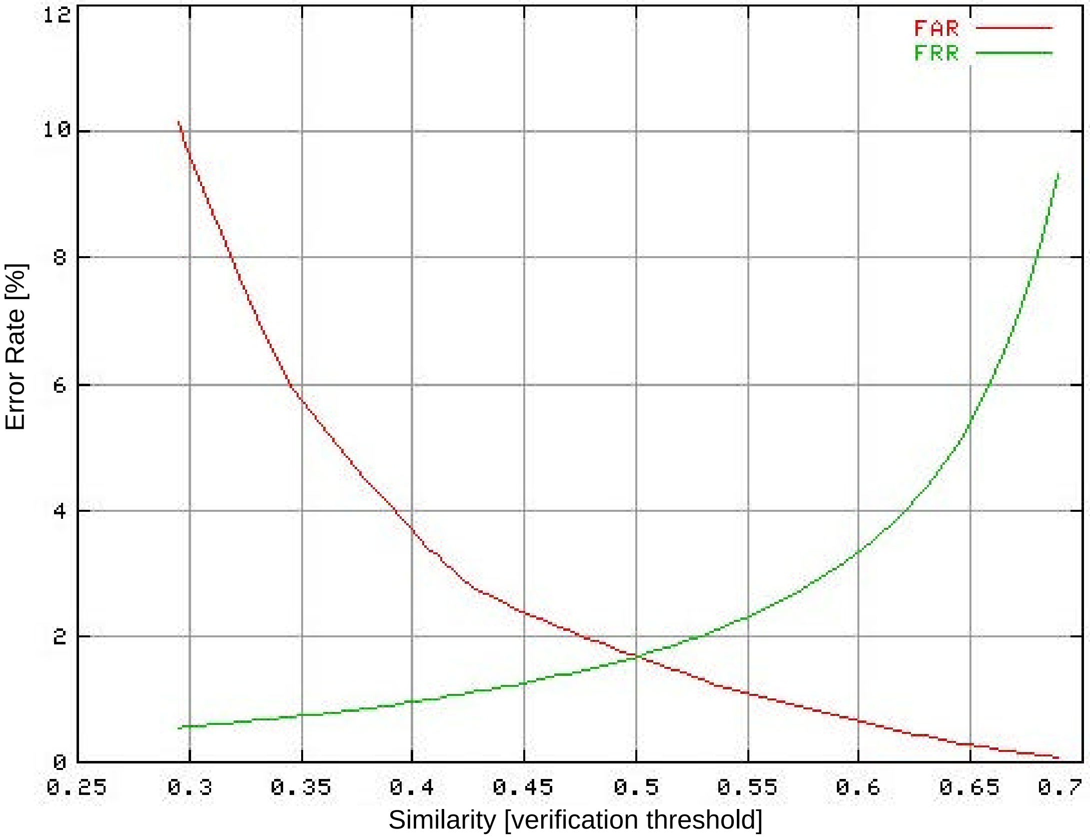


Figure 4.3 FAR/FRR curves

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

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. Zero means no similarity at all between the current face shot and a database, and one 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/ARR curves in Figure 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 our 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 classifier, 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.

4.3.2 Adaptive Similarity Threshold

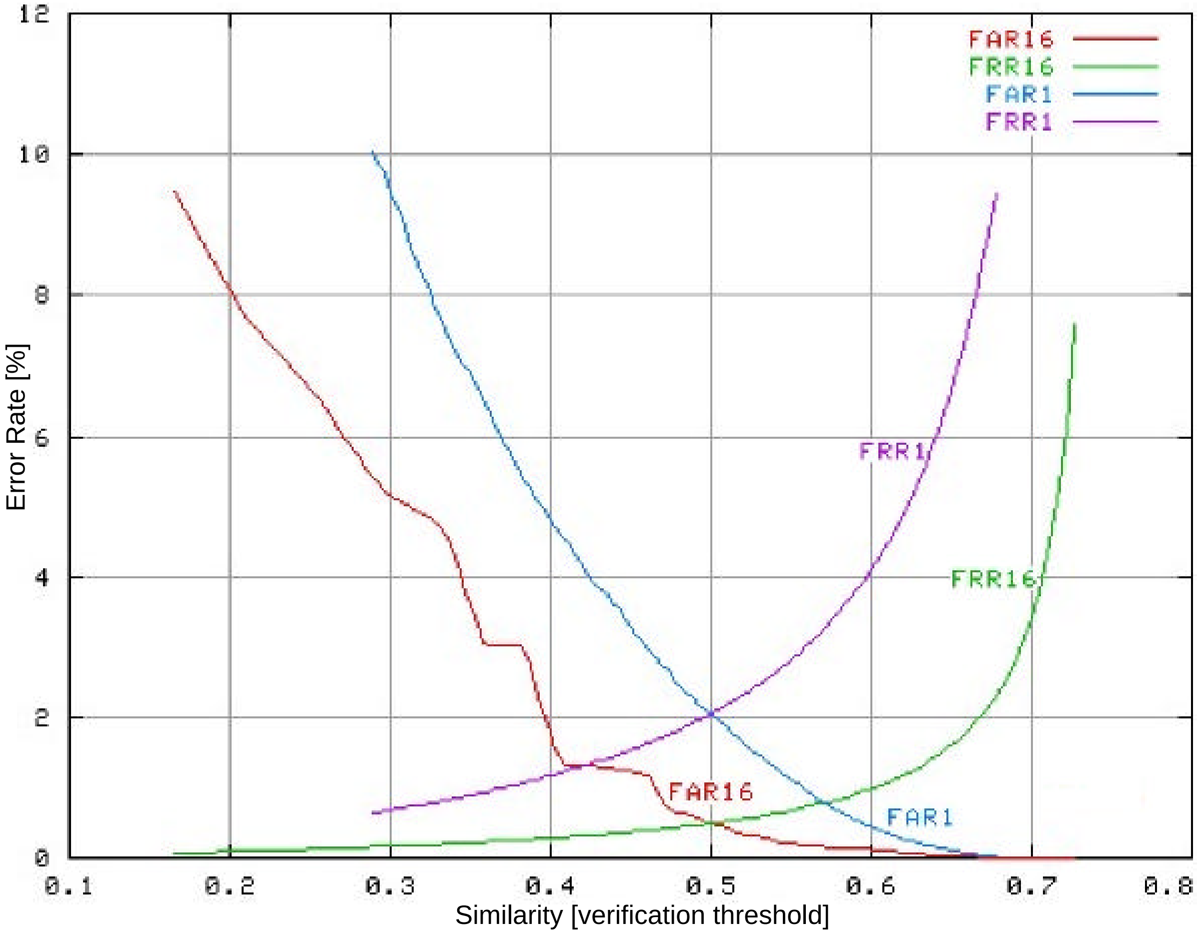


Figure 4.4 FAR/FRR curves – enrollment from 1 vs. 16

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

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 Figure 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).

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

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 one. 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 our above example, 0.65 is corresponding to the maximum value of AST which equals to $S_{V0} + a \bullet N_{\max}$, where $a \bullet N_{\max}$ can be set to 0.1 such that the minimum number of AST (S_{V0}) equals to 0.55.

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:

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

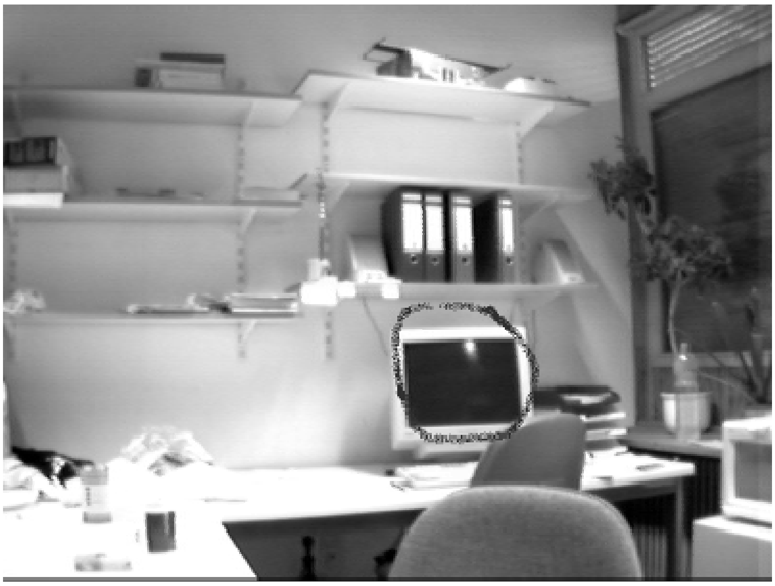
where A_i represents the coefficient with the range between 0 and 1, n is the filter length, m is a factor and can be intuitively approximated to n when each A_i equals to 1. $S_{v,i}$ denotes the respective similarity value of the *ith* frame image compared with a corresponding database.

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

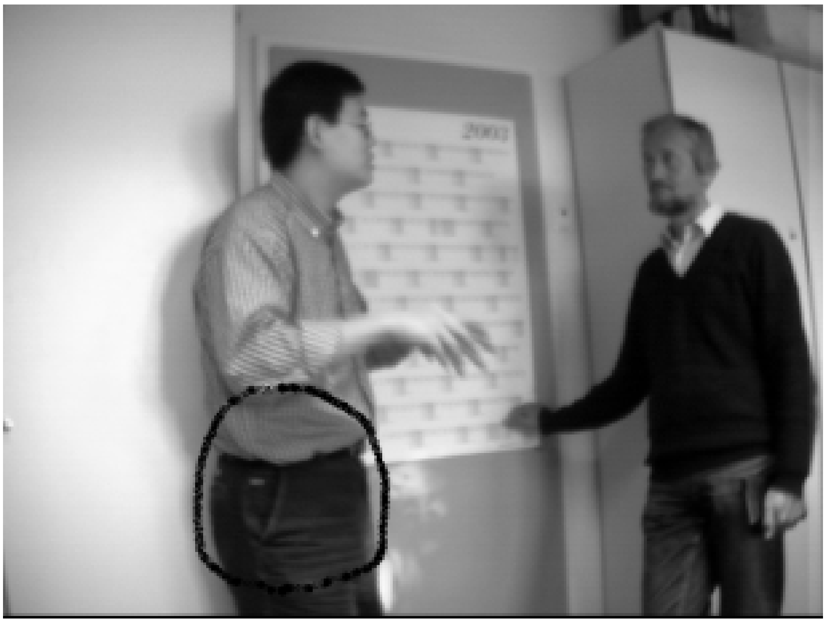
Obviously, the filter has a low-pass characteristic, which can deal with sporadically as well as frequently occurring errors. Such typical errors might include:

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

As we 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. Figure 4.5 demonstrates such two examples. As shown in Figure 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 Figure 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.



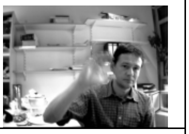
(a) False face detection due to a static face-like object



(b) False face detection due to motion

Figure 4.5 Two examples of false positives in face detection

Table 4.1 Similarity comparison of two similar persons

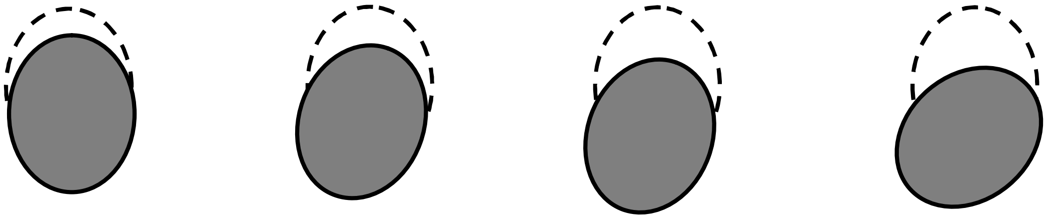
Person 1’s frame images	Person 2’s database 
Person 1—image1	$S_{v,1}$ = 0. 72
Person 1—image2	$S_{v,2}$ = 0.59
Person 1—image3	$S_{v,3}$ = 0. 65
Person 1—image4	$S_{v,4}$ = 0.51
Person 1—image5	$S_{v,5}$ = 0.11
Person 1—image6	$S_{v,6}$ = 0.08
Person 1—image7	$S_{v,7}$ = 0.65
Person 1—image8	$S_{v,8}$ = 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.

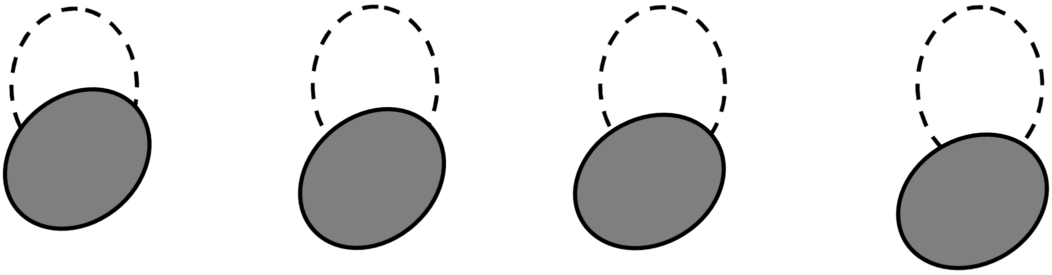
Table 4.1 shows the case of comparing two persons which could confuse the classifier when only one frame is used. It can be seen that there are two face images from person 1 which have pretty high similarity value when compared to some classes of person 2. Assume that *max(AST)* is set to be 0.65. Without 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 (chapter 3.3) and from the temporal filter. The sequence context is also included to contribute to same face decision.



(a) Frame t (b) Frame t+2 (c) Frame t+4 (d) Frame t+6



(e) Frame t+8 (f) Frame t+10 (g) Frame t+12 (h) Frame t+14

Face A
Face B

Figure 4.6 An example of face occlusion with slow motion

As mentioned in chapter 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

Figure 4.6 demonstrates an example of the face occlusion. Every second frame of a 16-frame image sequence is shown in Figure 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.

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 (IFD) 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): $\sum_{i=1}^{n} A_i \bullet S_{v,i}$ is either below the similarity threshold $m \bullet AST$ or beyond it. It is the decision from both the image-based recognition classifier and the filter. When above 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 Figure 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.

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)	

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:

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 of 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. 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 be 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 it 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 Figure 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.

Chapter 5 Unsupervised Face Database Construction

5.1 Overview

In this chapter, we mainly discuss the unsupervised way to build databases. We firstly introduce the related machine learning background, including supervised and unsupervised learning methods, and analysis of typical clustering structures. A proposed fused clustering structure as well as the related parameters for building the face database is then described. That is corresponding to the database construction step of the enrollment procedure in Figure 2.2. Finally, we explore the features that are important to achieve an optimized database, which are basically the mugshot selection criteria as well as the update rules. Those are crucial for the success of the whole automatic system while most of the current state-of-the-art systems are not paying much attention to them.

5.2 Backgrounds for Constructing Face Databases

5.2.1 Supervised Learning

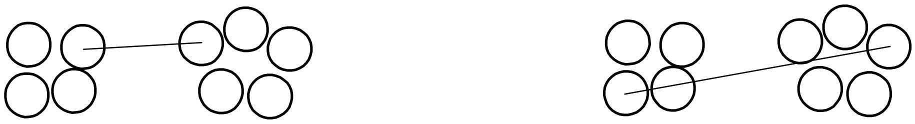
There are in general two major machine learning methods: supervised and unsupervised. Supervised learning attracts many research interests, among which inductive learning is the most popular. It generalizes from observed training examples by identifying features that empirically distinguish positive from negative training examples, which is defined in [85]. 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 [85]. 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 [85]. 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 [85]: better generalization accuracy when predefined rules are provided and reliance on observed data when prior knowledge is only approximately correct and incomplete.

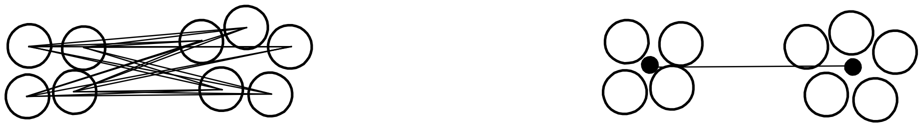
Another critical assumption in most supervised learning is: the distribution of data is not changing from time to time, which is actually frequently occurring in real world. For our specific face recognition problem, it is exactly the case that faces are always looking different from time to time. To 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.



a) Minimum distance b) Maximum distance



c) Average distance d) Center distance

Figure 5.1 Four different ways of distance measure

Figure 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.

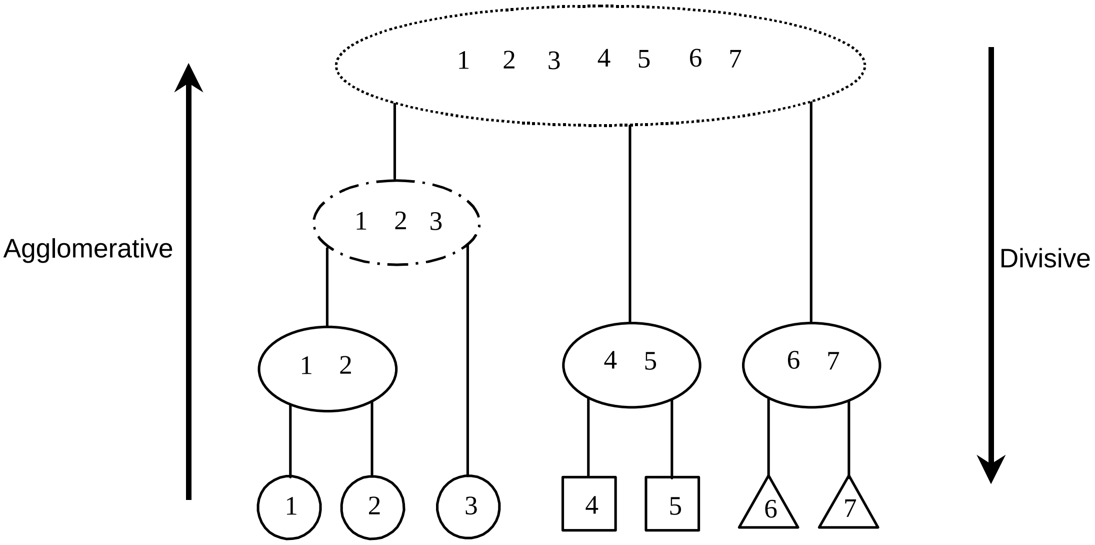
Average distance and center distance are the two better ways of representing distances. The average distance method calculates the average of all distances between each pair of the elements. For center method, centroid is calculated from all elements from one cluster. Cluster distance is then denoted 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 Chapter 4.3, the Bayesian-like FaceVACS classifier is applied, which can be considered as the an improved center distance method. In the following, we apply this representation to illustrate different clustering algorithms and structures.

5.2.3 Clustering Analysis

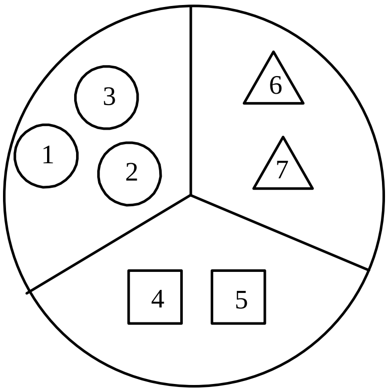
There are in general two traditional clustering techniques: hierarchical algorithms and partitioning algorithms. In hierarchical method, the data are classified into clusters by different layers rather than only one hierarchy. The clusters therefore have a tree-like structure. The tree can be built by 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 it into smaller and smaller clusters. This process continues until each cluster contains only one observation. The two hierarchical clustering methods are illustrated in Figure 5.2(a).

In partitioning method, a prescribed number *k* of classes is given. Training data are grouped into the 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. Figure 5.2 (b) illustrates the partitioning structure of the same observation data as in (a).

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 Figure 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 be converging 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.



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



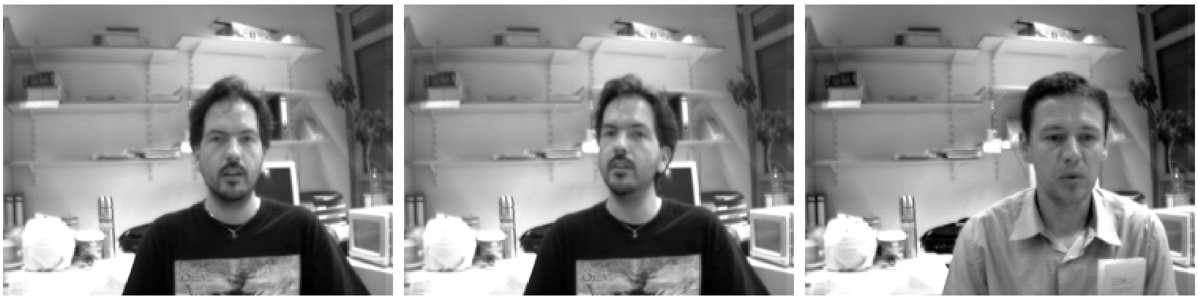
(b) partitioning clustering method, represented by a pie-like structure

Figure 5.2 Sketches of the two clustering methods

5.3 Database Structure

5.3.1 A fused Clustering Method

Figure 5.4 shows an abstract example of applying clustering methods for constructing the face databases. There are three groups of data observations available as shown in Figure 5.4 (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, similar head poses, and facial expressions. Figure 5.3 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.

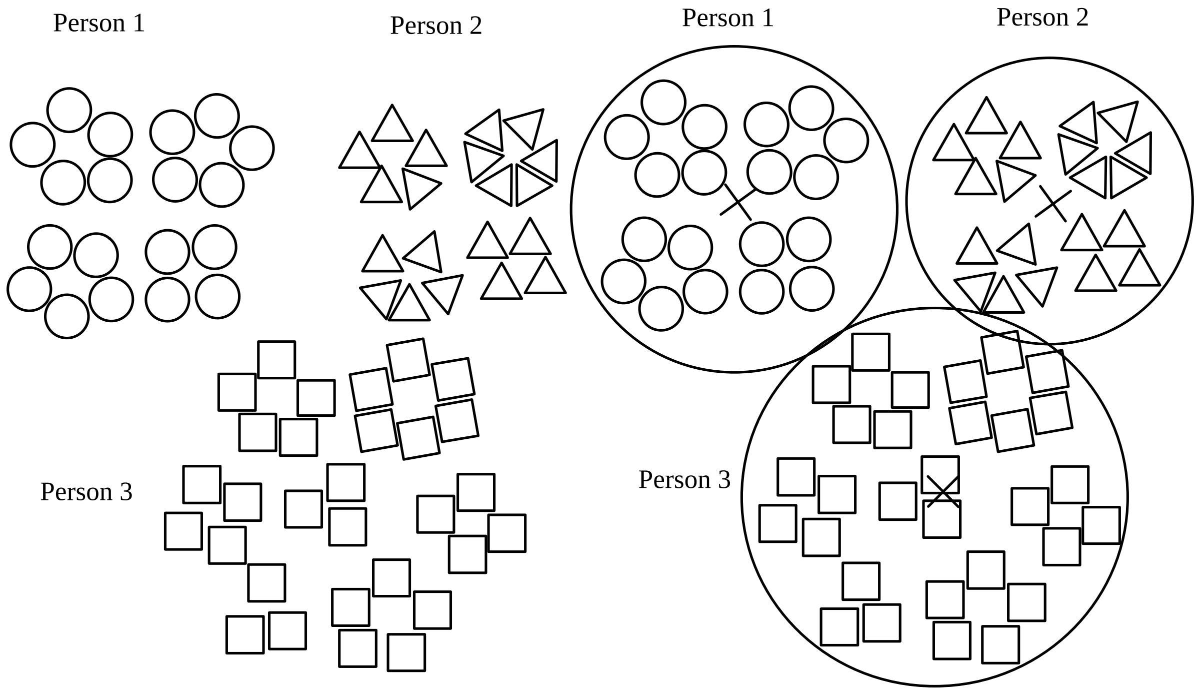


Person A (1)　Person A (2)　Person B

Figure 5.3 An example of different people having a big similarity

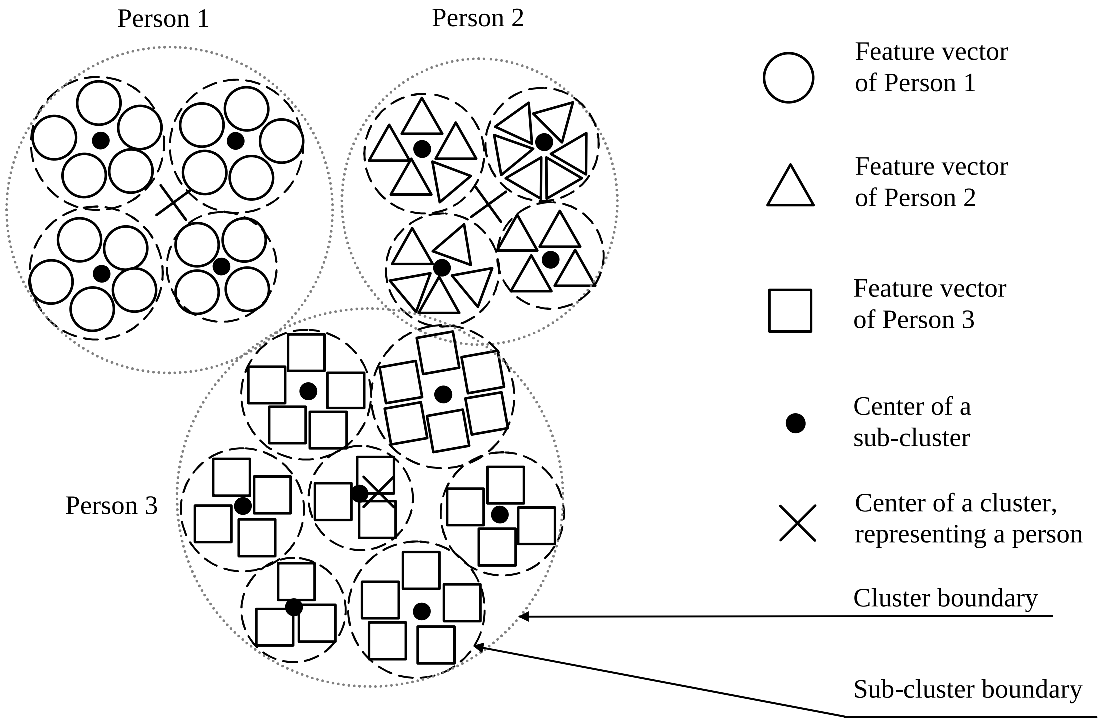
Now we are looking back to the example in Figure 5.4. Partitioning method is advantageous over the pure hierarchical method since it can clearly cluster three persons. Figure 5.4 (b) illustrates the partitioned structure. However, as can be seen from the figure, there is an obvious overlap between boundaries 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.

To decrease the possibility of producing overlap, we propose an improved structure which combines the hierarchical and partitioning method, as shown in Figure 5.4 (c). We call this method the 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 Figure 5.4 (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.



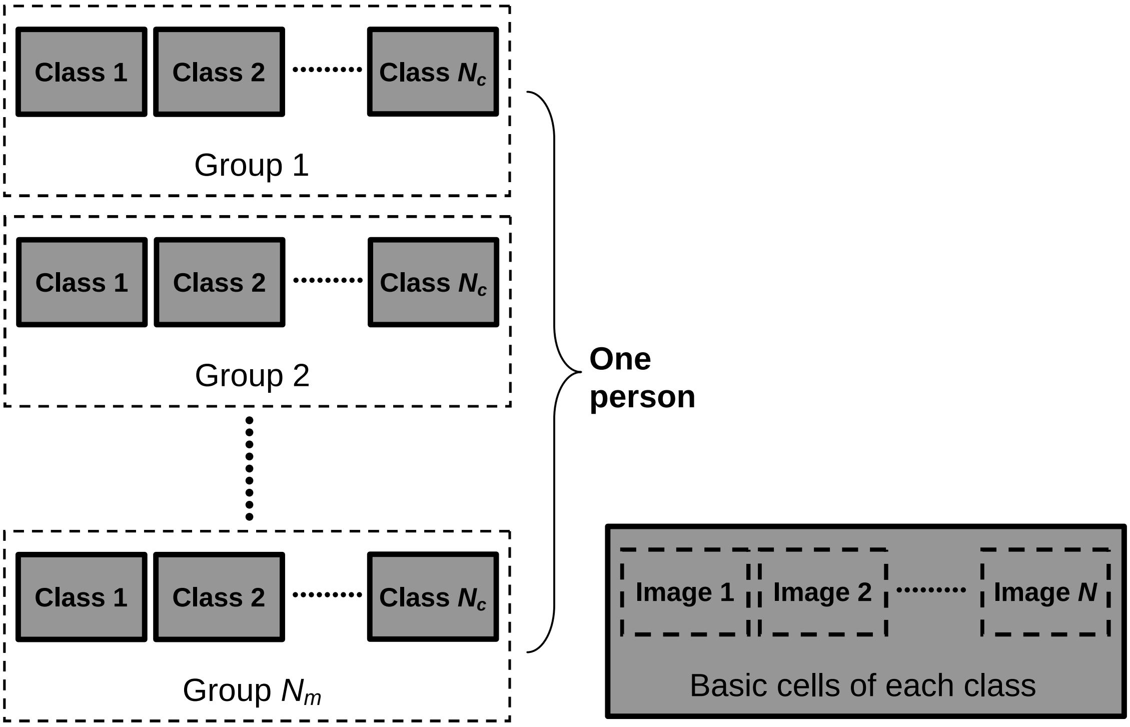
(a) original data without clustering

(b) partitioning clustering method

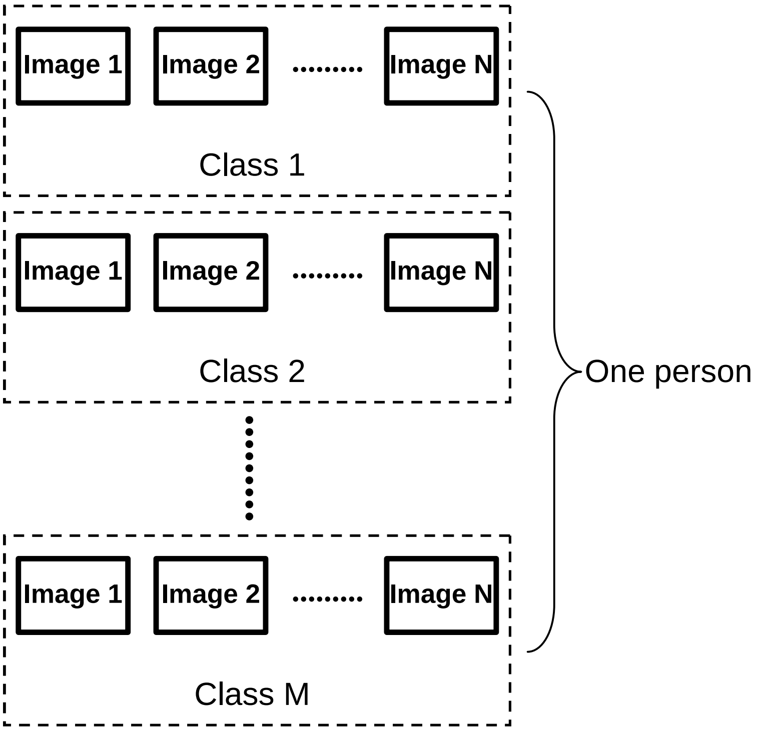

Person 1
Person 2
Person 3
Feature vector of Person 1
Feature vector of Person 2
Feature vector of Person 3
Center of a sub-cluster
Center of a cluster, representing a person
Cluster boundary
Sub-cluster boundary

(c) An improved clustering method

Figure 5.4 Two structures of a face database



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



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

Figure 5.5 Face database structure in two different states

The second step is to apply partitioning method. We predefine the maximum 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 Figure 5.4 (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 Figure 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 Figure 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 Figure 5.5 (b).

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

- 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 chapter 4.4. Let us look back to the example of Figure 5.4 (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, Figure 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 Figure 5.5 (b), we have to examine how to choose the optimum values of M (maximum number of classes pro person) and N (maximum number of face images pro class).

N is chosen by following the recommendation of the FaceVACS technology. In Figure 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 (false acceptance rate) and FRR (false rejection rate). 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 ([86], [87], [88]), which are suitable for achieving the FAR/ARR curves as well.

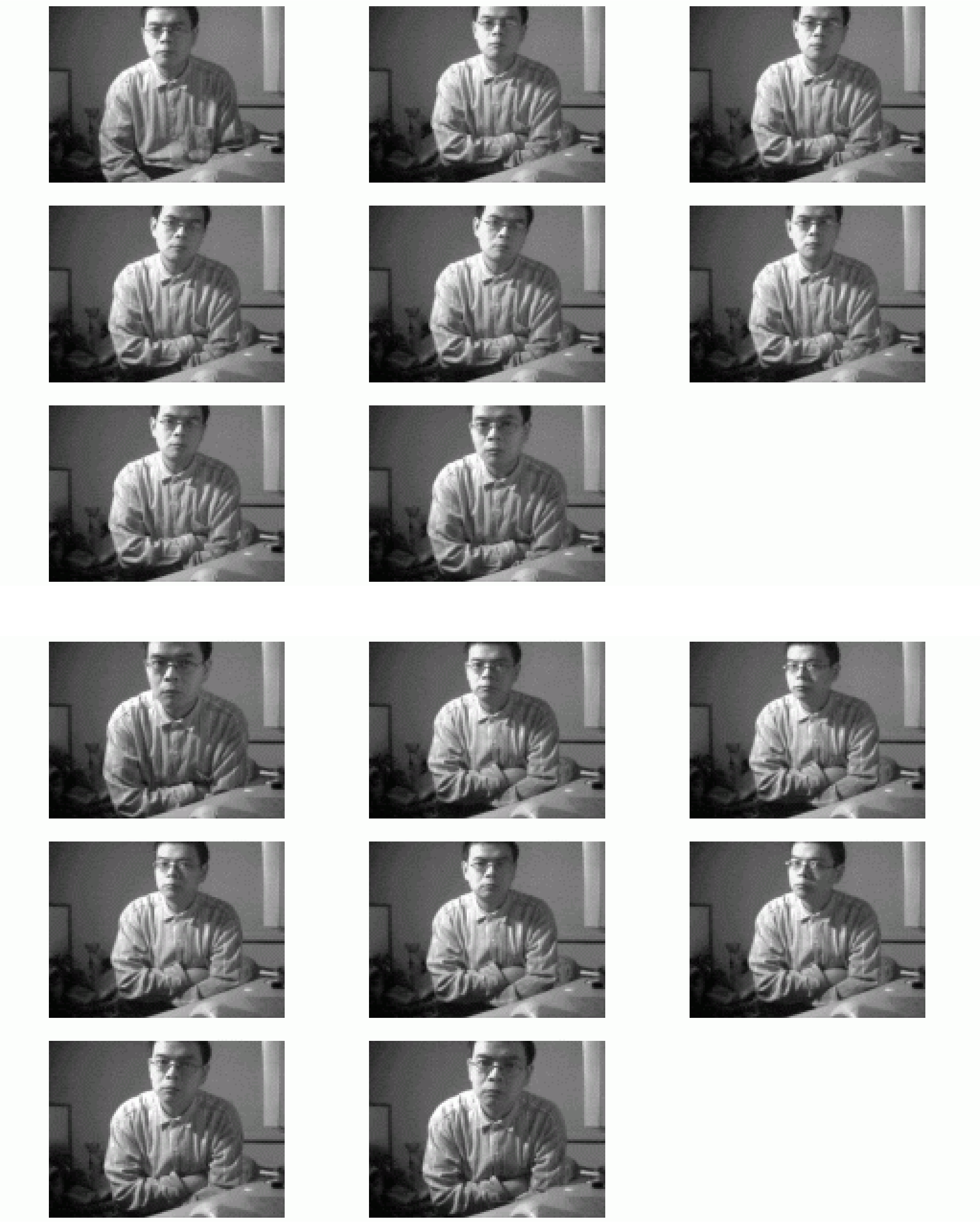
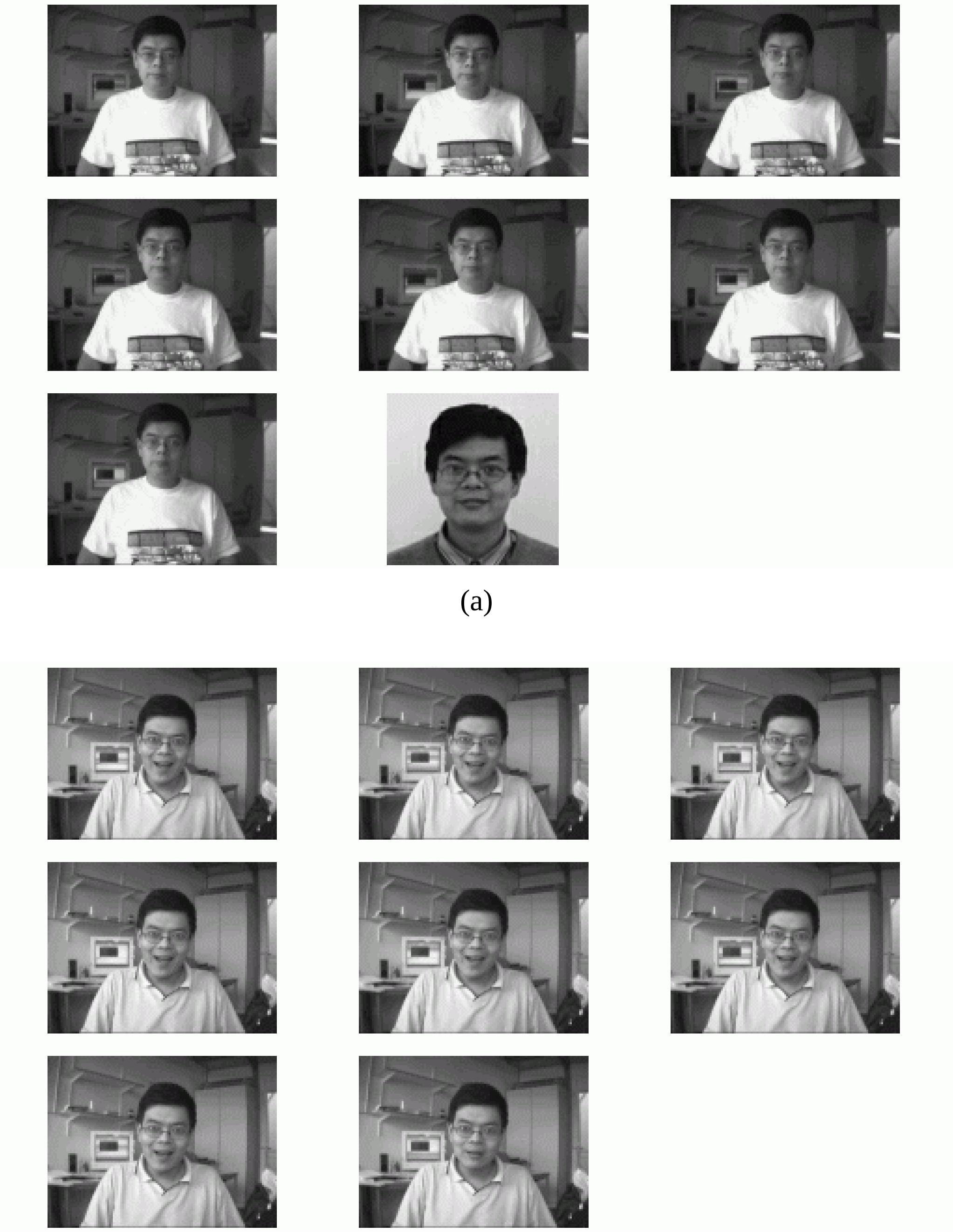


Figure 5.6 Two classes in the testing set s2

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.

Figure 5.6 shows the two enrolled classes in graphics. Those images are continuously selected from S_2.



(a)

(b)

Figure 5.7 Examples of the testing set s1

Figure 5.7 lists some of the example images in set s1. With two classes enrolled, the recognition result with S_1 is:

- 150/392 images that are correctly recognized (38%)

It indicates that 2 classes are still far too few.

With 8 classes enrolled, however, the recognition result improves a lot:

- 209/392 images that are correctly recognized (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 chapter 7.4, we will provide the empirical choice of the actual M value in the software implementation.

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 (FRR). 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 enrolled 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:

$$AST2 < S_v < AUT, \text{ and } n_e < n_{th} \tag{5.1}$$

Where AUT is decreasing for each database growing from initialized state to stable state, AST2 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.

AST2 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 section 3, 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 (6) 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 faces are detected for a certain length of time.

Chapter 6 State Machine Based Automatic Procedure

6.1 Overview

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.

6.2 States Explorations

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

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.

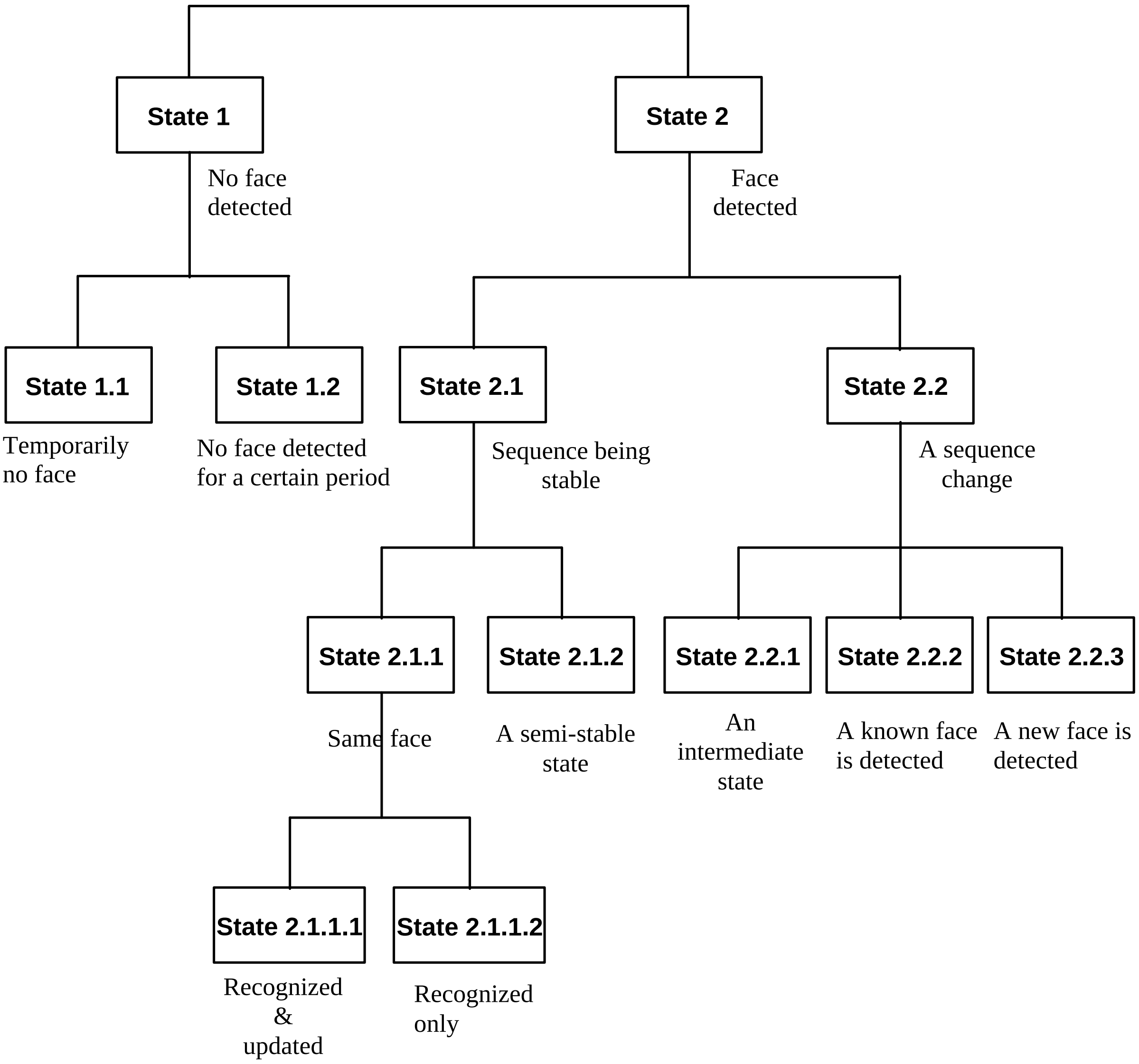


Figure 6.1 Definition of all possible states for the face recognition procedure

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 $\sum_{i=1}^{n} A_i \bullet S_{v,i}$. For the simplicity of representation, we denote $\sum_{i=1}^{n} A_i \bullet S_{v,i}$ 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**.

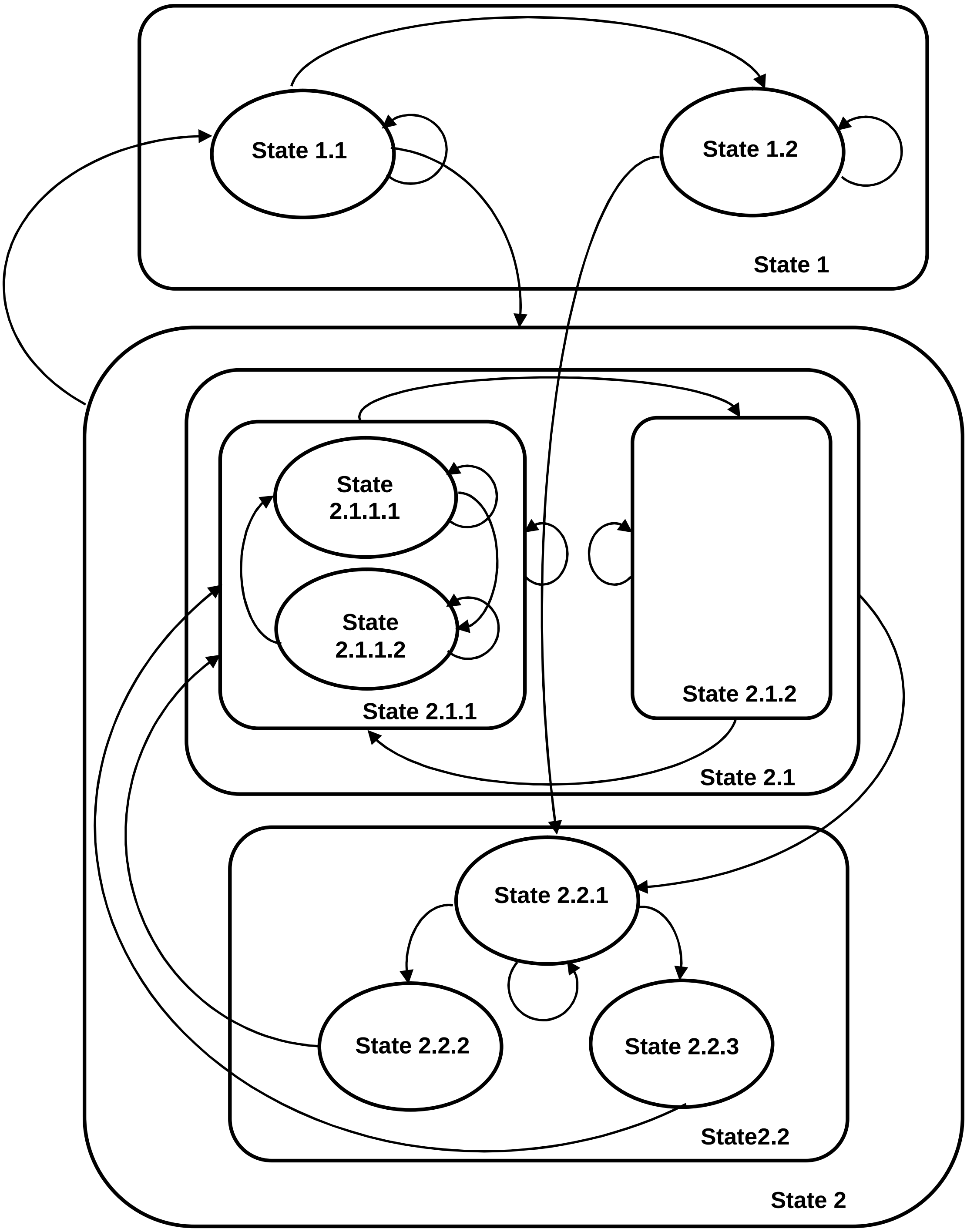


Figure 6.2 Definition with a hierarchical state machine

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 chapter 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 Figure 6.2.

Chapter 7 System Implementation

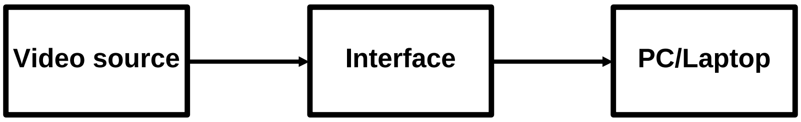
7.1 Overview

In this chapter, we are mainly focusing on the implementation issues, not only based on the generic algorithms but also the specific image based detection and classification techniques. We firstly cover the hardware configurations and then describe the software implementation efforts for the whole system. Additional parameter settings related to the detection and recognition classifier are finally explored for optimizing the system performance.

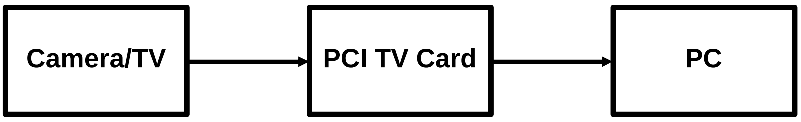
7.2 Hardware Configuration

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 chapter 1.3. This is very important especially in the consumer electronics industry-- a system with cheap and generic setups is always attractive and promising.

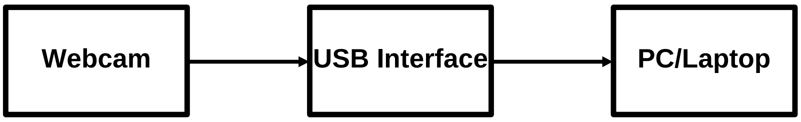
We have implemented the whole system mainly software-based as a demo version.



(a) Generic hardware configuration



(b) One applied setup



(c) The other applied setup

Figure 7.1 Hardware configurations of the system

Video capturing devices including a camera and its corresponding interface, a PC or a laptop are the three main parts of the configuration, as shown in Figure 7.1 (a). The interface can be any video capturing interface that can grab frame images from the camera to the PC/Laptop. It can be a PCI TV capture card connected to a PC, or USB/firewire interfaces to PC/Laptop. Figure 7.1(b) and (c) show two different setups we have used for our system.

In (b), the video source is either a 50-Euro CMOS video camera or a TV cable. 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 our software running on Windows. This configuration therefore supports a real-time running of the system from either cameras or TV programs.

In (c), a 30-Euro webcam is connected to PC/Laptop through a USB 2.0 interface. The webcam can only support 320×240 pixels at 12fps. This setup is more focused on mobile purposes for 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[2] (Open Source Computer Vision) for image processing, libraries from the FaceVACS SDK tool[3] for image-based face detection and recognition classification.

With currently available PC, 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 comes from the FaceVACS SDK tool. It requires around 1/3 ~ 1 second for feature extraction and encoding steps in the enrollment procedure, which doesn’t fit well to the tracking requirement. Therefore, we have two versions to demonstrate our algorithms.

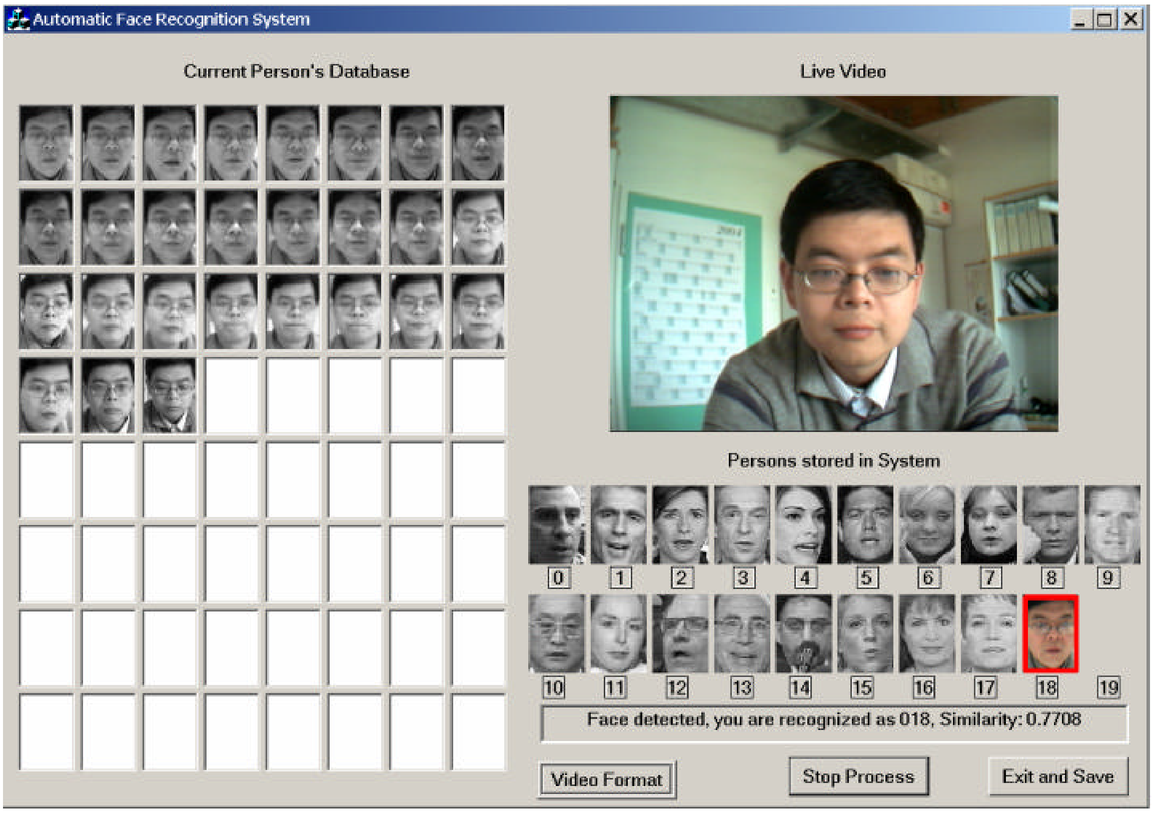
One is online version, which processes the video from camera or from TV programs. The actual processing speed of the system is around 1-2 fps for a 384×288 resolution on a 2.4 GHz PC, which does not fulfill the 10~20 fps assumption. Therefore, it can only achieve an ideal performance when people behave 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. Figure 7.2 shows a screenshot of the system running with online live video.

---

[1] Freely available tool from Microsoft, can be downloaded from http://www.microsoft.com/downloads.

[2] Freely available tool from Intel, can be downloaded from http://sourceforge.net/projects/opencvlibrary

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



Face detected, you are recognized as 018, Similarity: 0.7708

Figure 7.2 Photograph 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 corresponding 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 Figure 7.3. There are two main units: the user interface unit and the algorithms unit.

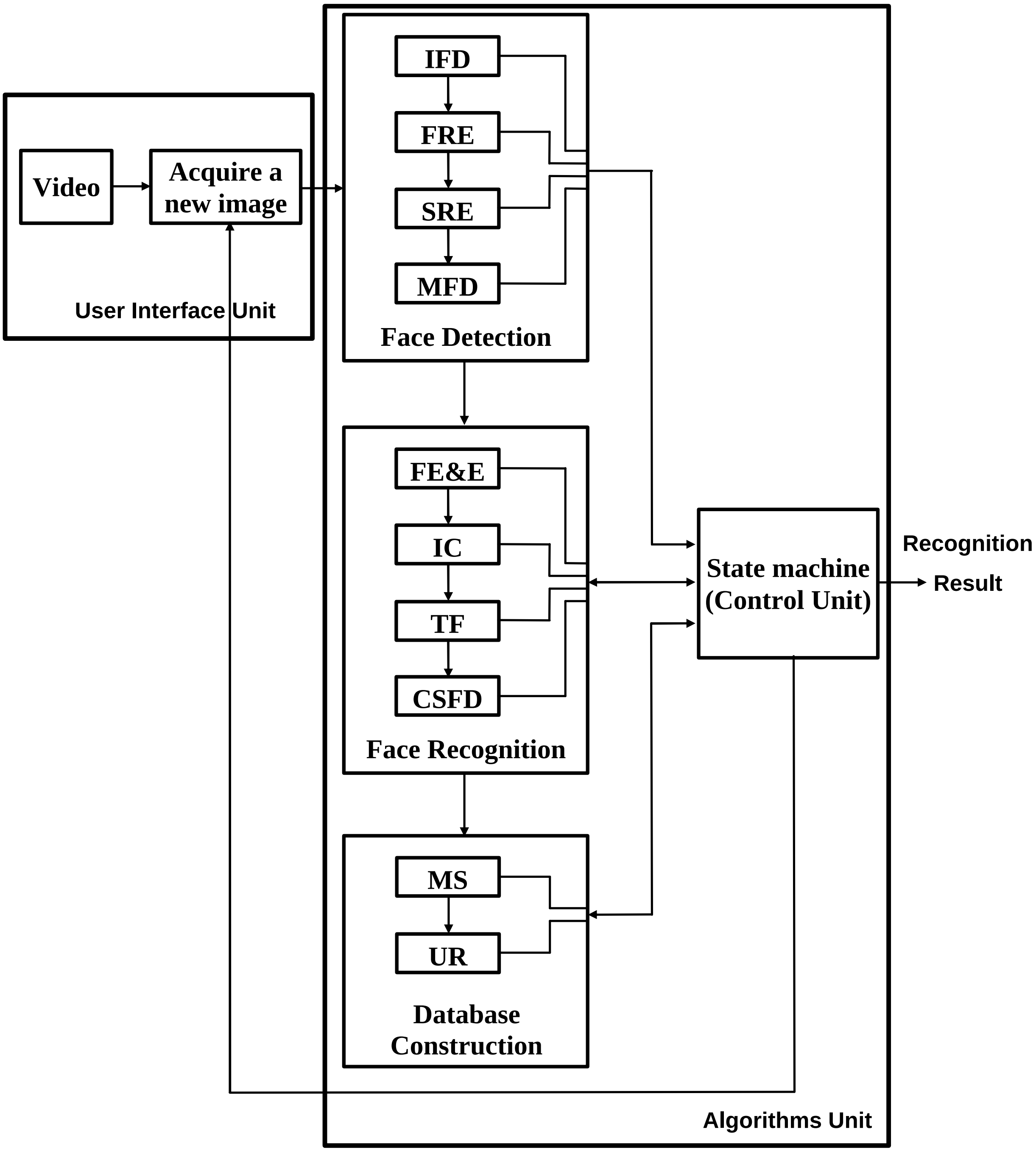
Three main tasks are concerned for the first interface unit:

- 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 Figure 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.

Figure 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.

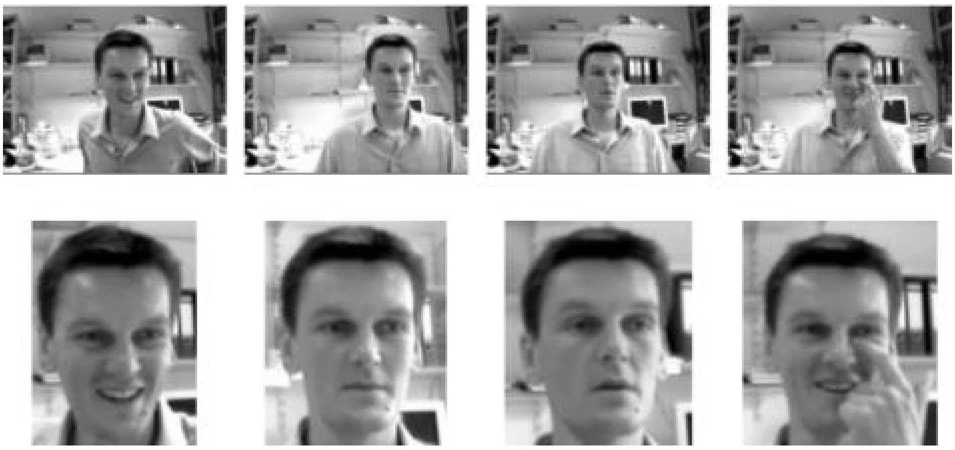
One third of the source codes (about 1000 command lines) are contributed to the algorithms unit, where around 250 lines are written for face detection, 360 lines for face recognition, 180 lines for database construction, and 210 lines for the state machine based central control unit. The control unit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

7.4 Summary of Technology Dependent Parameters

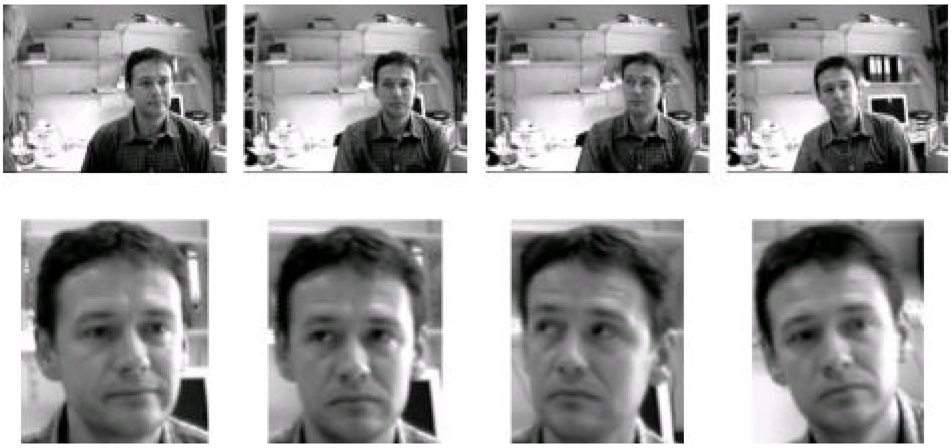
Up to now, we have mainly discussed the general algorithms for building an automatic and adaptive face recognition system. The discussion is supposed to be as independent as possible from the face detection and face classification techniques. But we still have some specific parameter settings that are 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 chapter 5.3.2. As illustrated in Figure 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×N can also influence FRR. As shown in Figure 4.4, in principle, the bigger M×N, the less FRR. With given N and M, we can correspondingly judge an optimum range of S_{V0}. In chapter 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 Figure 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.



Person A



Person B

Figure 7.4 Samples of whole frame images and the corresponding face images

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 Figure 7.4.

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 Figure 7.5. Since a higher cross-similarity indicates a higher FAR, the green-dashed curve achieves better recognition performance. That is to say, the use of whole images achieves lower FAR than that of face images.

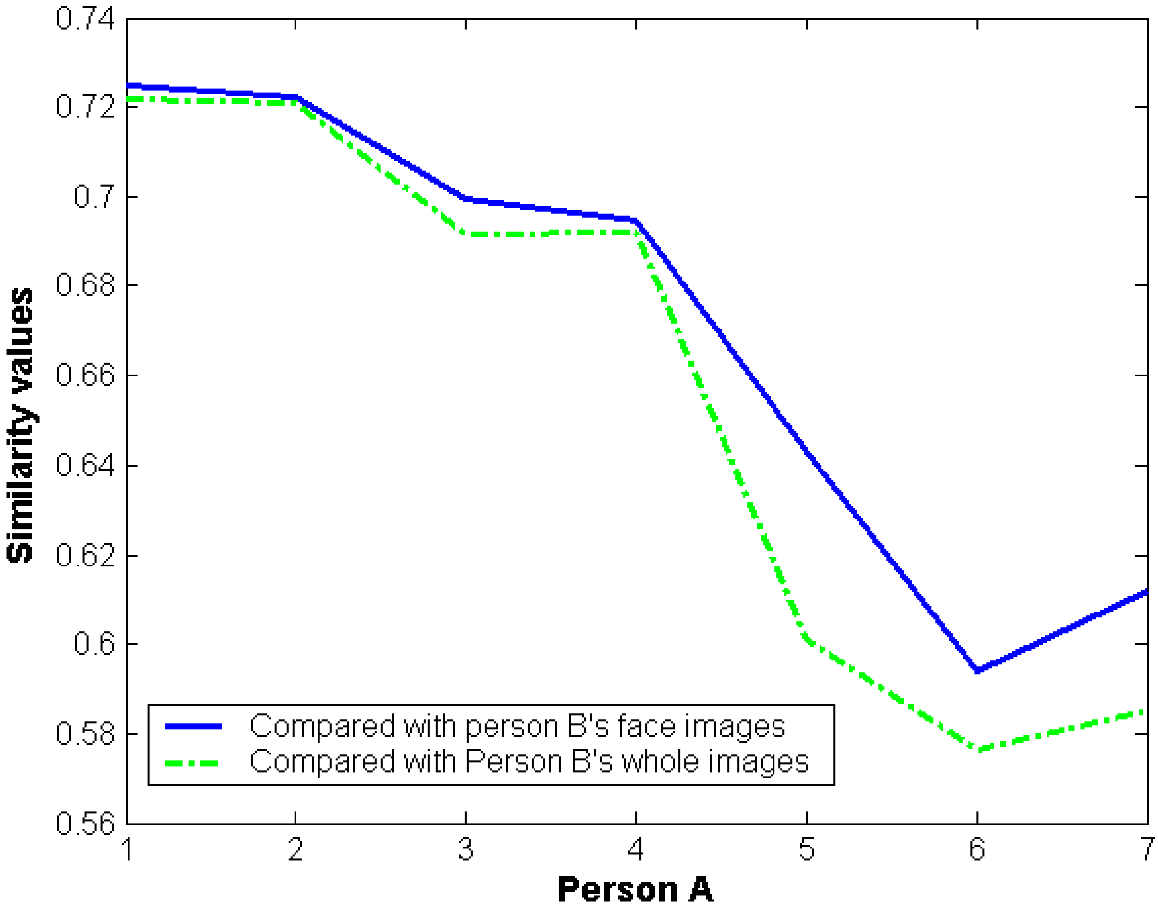


Figure 7.5 Cross-similarity differences by using face images and whole images

Auto-similarity is tested in the following experiment. Two different sequences from the same person are compared with each other, as shown in Figure 7.6. As shown in this curve, the higher the auto-similarity, the better. The use of whole images (colored in green-dash) 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 Figure 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.

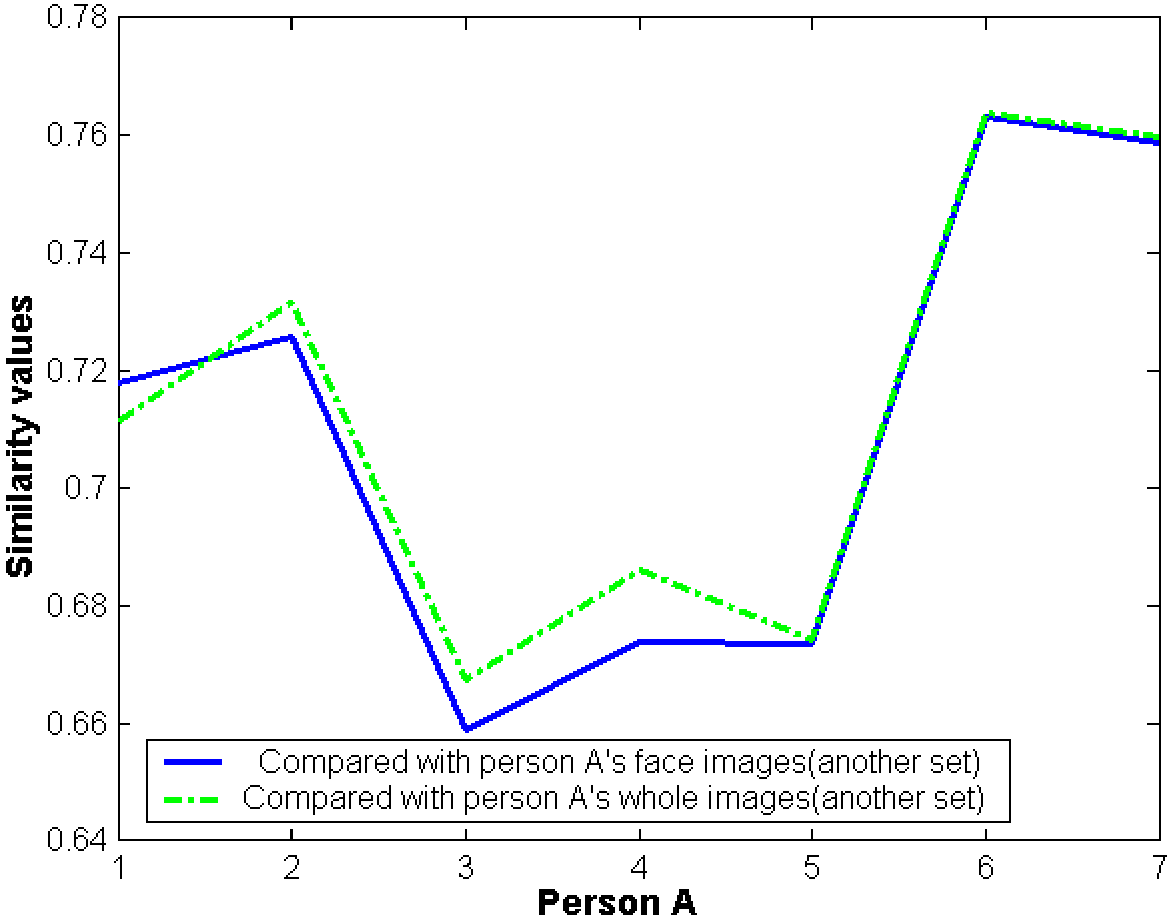


Figure 7.6 Auto-similarity differences by using face images and whole images

Chapter 8 Performance Analysis

8.1 Overview

In this chapter, we are mainly demonstrating the performance of the previously discussed 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.

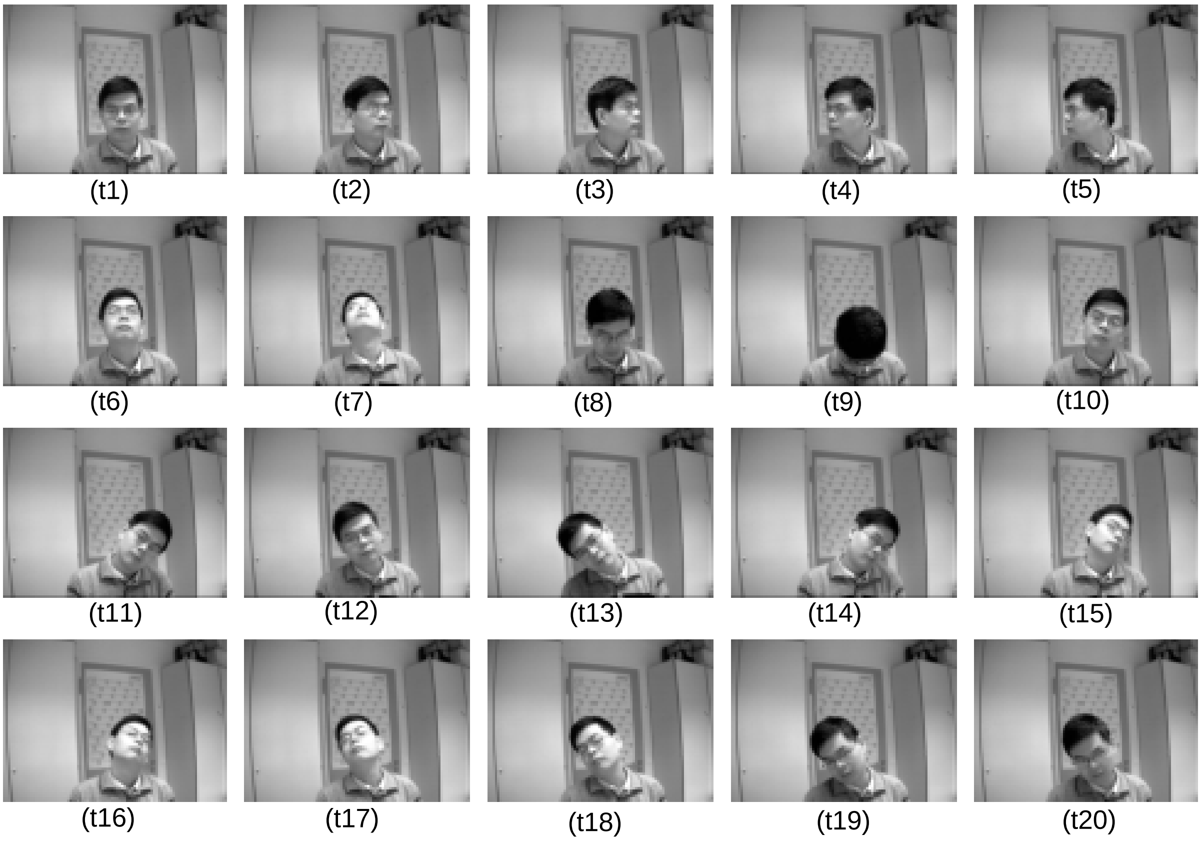
8.2 Performance of Face Detection

As discussed in Chapter 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.

To test the performance of the proposed algorithms, we have made ten video sequences, each of which is processed by IFD only and our system for comparison. Each video sequence is captured in 25 fps.

Figure 8.1 shows the most critical experimental sequence with freely 3D head motions. It is captured at 384×288 pixel resolution. One person is freely moving his head in 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 Figure 8.1. 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. The comparison result of the detection rate for the whole sequence is drawn in Figure 8.2. 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.



(t1) (t2) (t3) (t4) (t5)
(t6) (t7) (t8) (t9) (t10)
(t11) (t12) (t13) (t14) (t15)
(t16) (t17) (t18) (t19) (t20)

Figure 8.1 Sequence of one person with significant 3D head motion

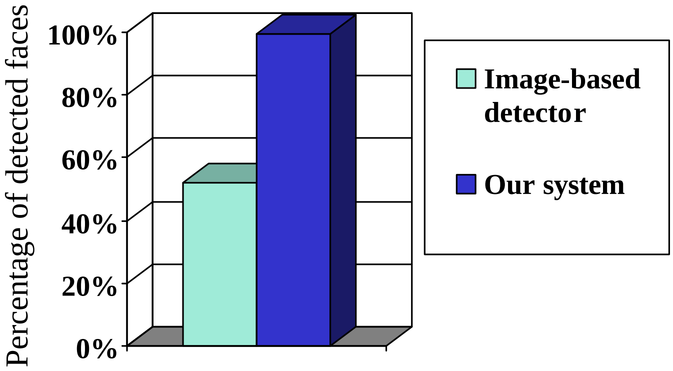


Figure 8.2 Performance comparison between an IFD and our combined detector

Figure 8.3 shows one sequence with intentionally different facial expressions. It is captured in the resolution of 192×144 pixels. To our first surprise, the IFD behaves even worse than the case of pose variation. 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 Figure 8.3, 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 angles (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 sequences to both achieve fast speed and save hardware storage spaces.

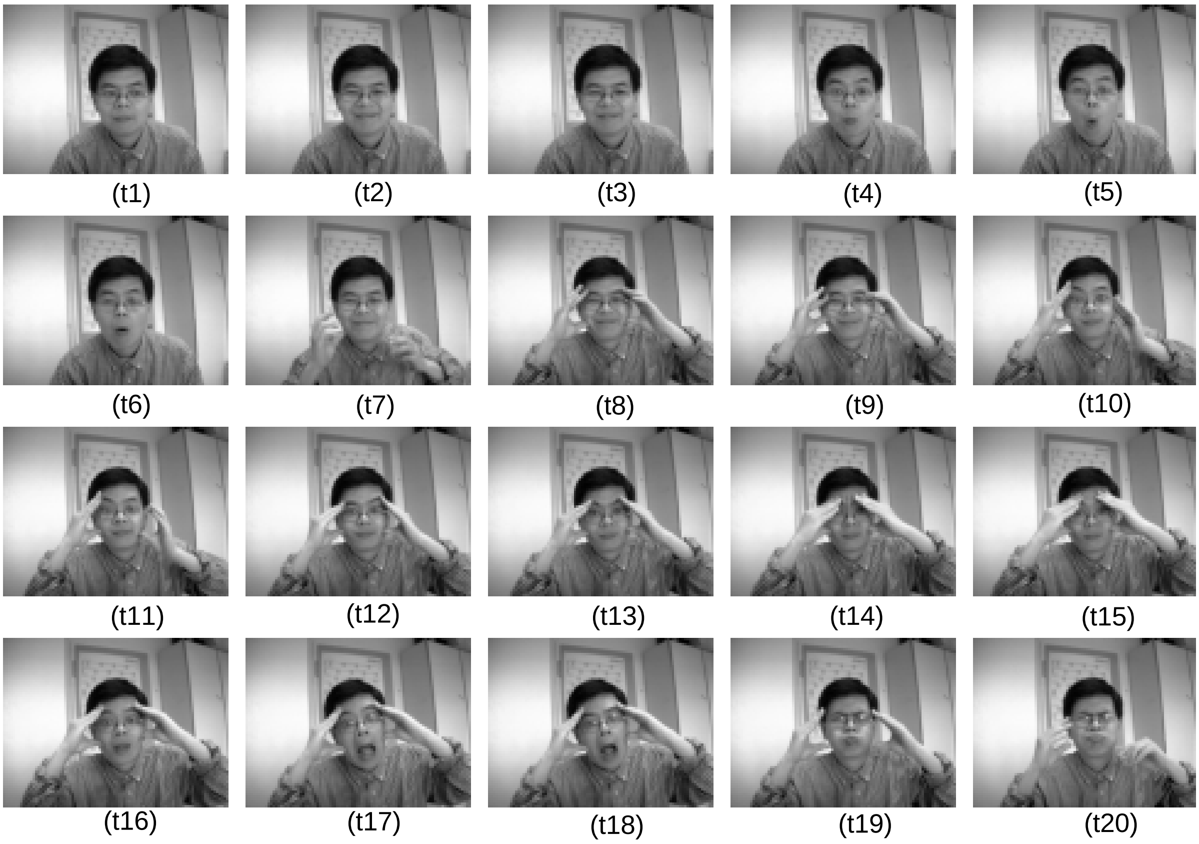

(t1) (t2) (t3) (t4) (t5)
(t6) (t7) (t8) (t9) (t10)
(t11) (t12) (t13) (t14) (t15)
(t16) (t17) (t18) (t19) (t20)

Figure 8.3 Sequence of one person with intentionally facial expression changes

Therefore, we have made sequences to further analyze the detection quality against scale changes. Figure 8.4 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 Figure 8.4, 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.

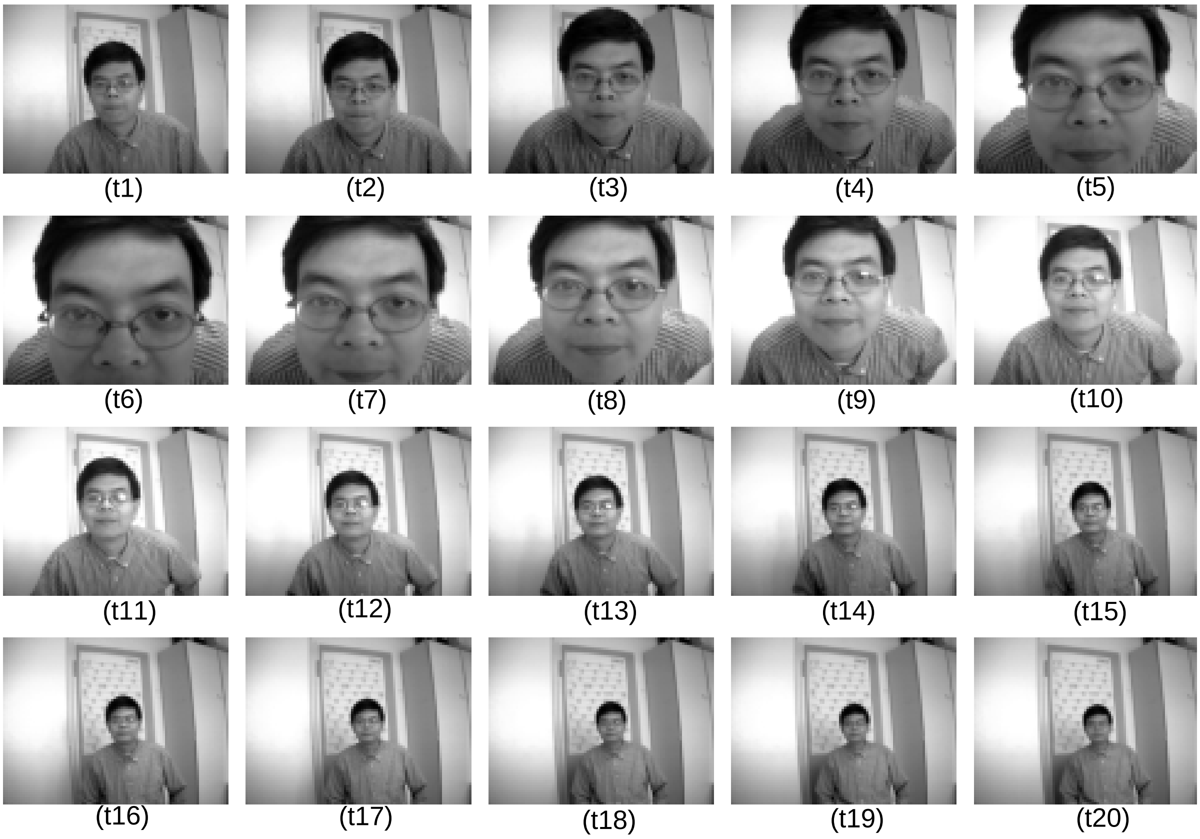


Figure 8.4 Sequence of one person with significant scale changes

The sequence in Figure 8.5 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 ([82], [83], [84]). 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.

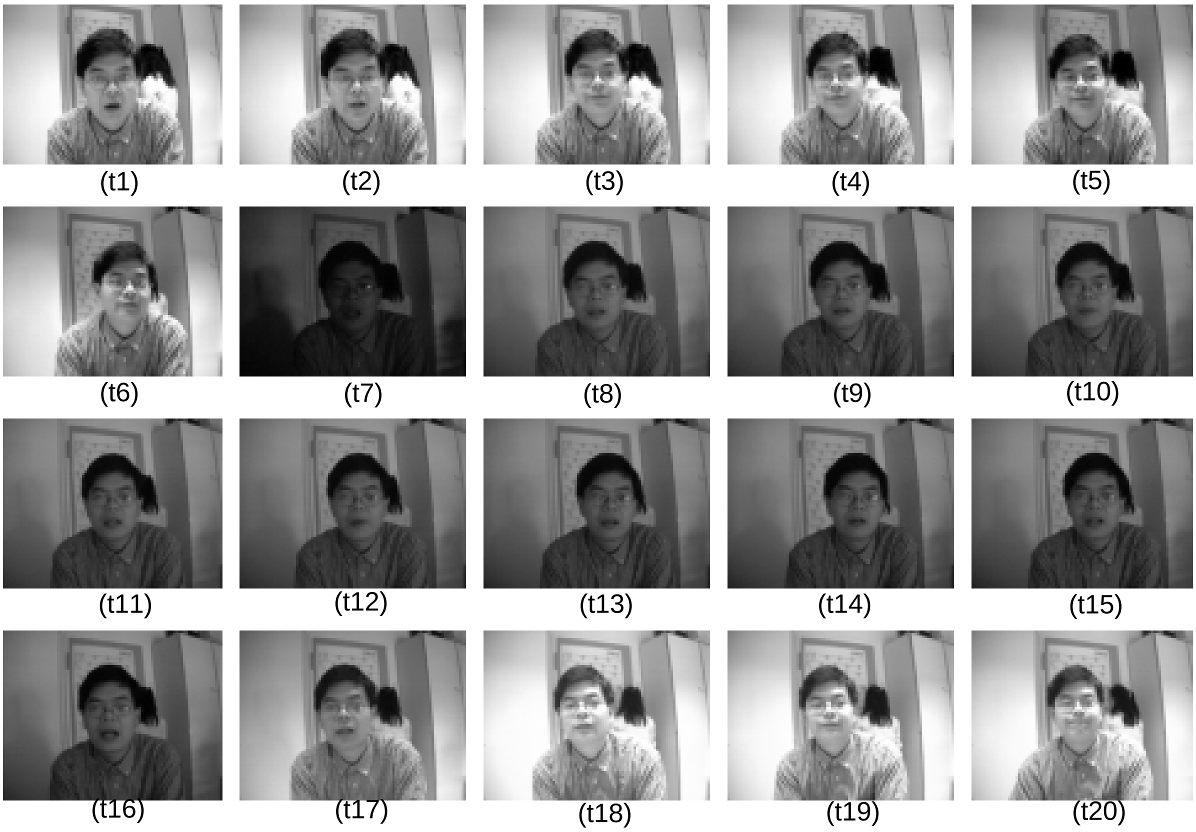

(t1) (t2) (t3) (t4) (t5)
(t6) (t7) (t8) (t9) (t10)
(t11) (t12) (t13) (t14) (t15)
(t16) (t17) (t18) (t19) (t20)

Figure 8.5 Sequence of one person with significant lighting changes

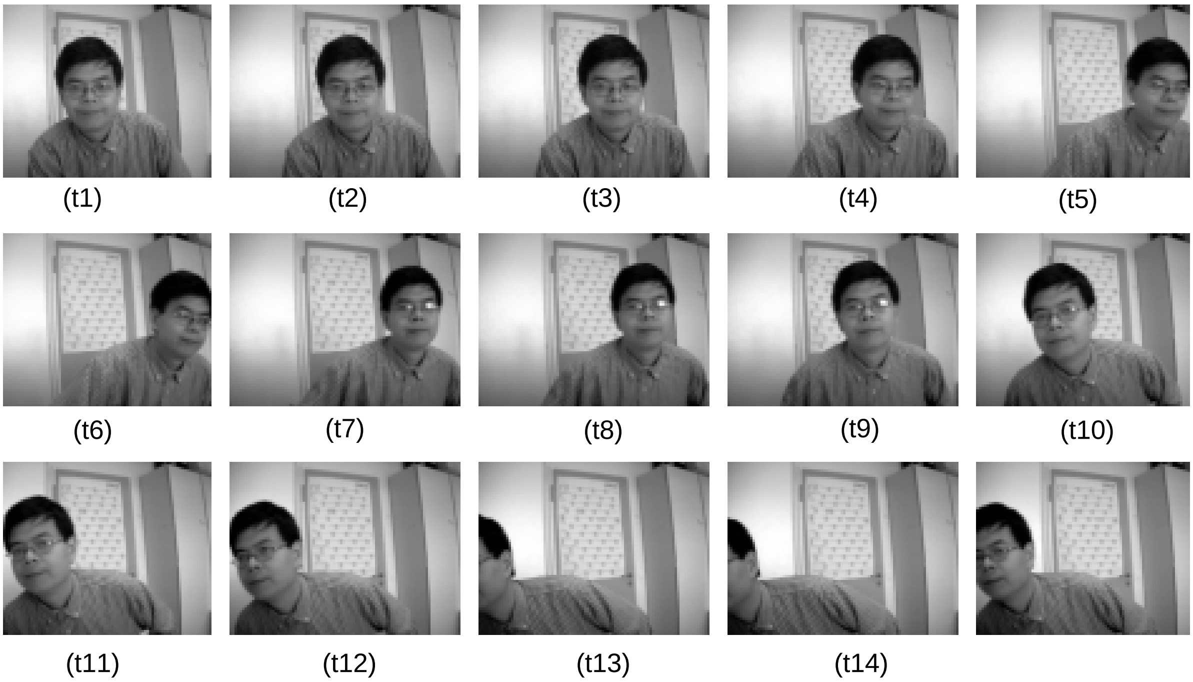

(t1) (t2) (t3) (t4) (t5)
(t6) (t7) (t8) (t9) (t10)
(t11) (t12) (t13) (t14)

Figure 8.6 Sequence of one person with fast motion

The final example sequence shown in Figure 8.6 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 the following table:

Table 8.1 Performance comparison between the IFD alone and our detection method

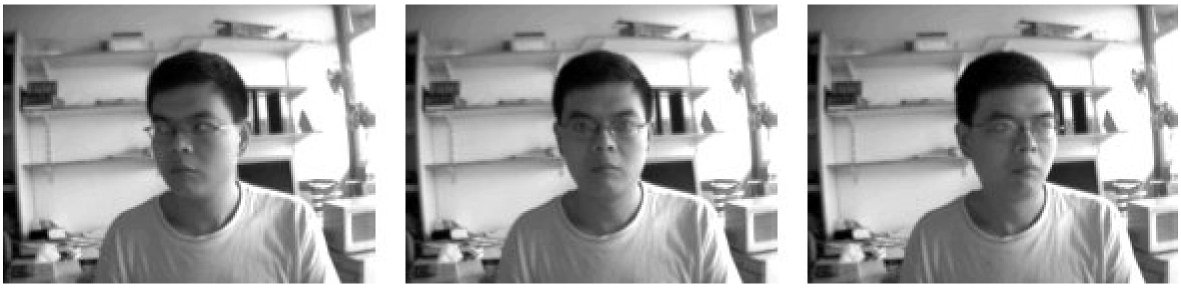
	Face detection rate under different parameters (in percentage of the whole sequence number)				
	Head pose (Figure 8.1)	Facial Expression (Figure 8.2)	Scale (Figure 8.4)	Luminance (Figure 8.5)	Motion Blurs and Occlusion (Figure 8.6)
IFD alone	45%	41.5%	58.2%	100%	69%
Our detector	93%	100%	98.2%	100%	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.

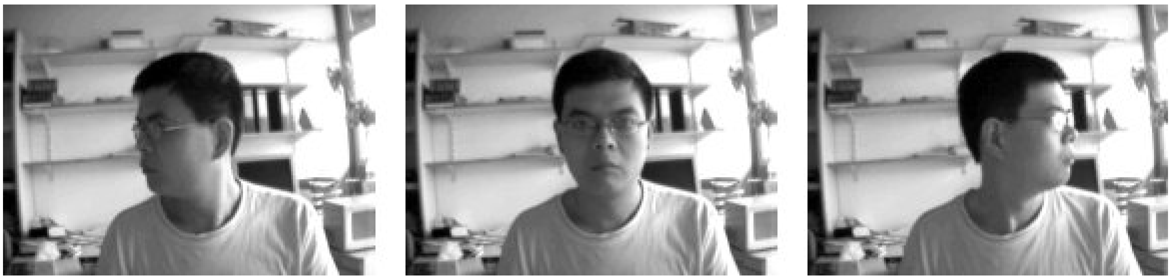
8.3 Performance of Face Recognition

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 similar ways, we can conclude that similar performance can be achieved. We have used the same sequences as for the face detection performance, as shown in chapter 8.2. The result supports our conclusion very well. In the following, we only take one typical case of head pose variations to demonstrate our recognition performance. Most recognition systems have problems to deal with this critical change.

We are demonstrating the performance with three kinds of head pose variations: yaw, pitch and roll. We have used the FaceVACS recognition classifier alone for comparison with our combined recognition algorithms.



(a) Classifier only recognition result

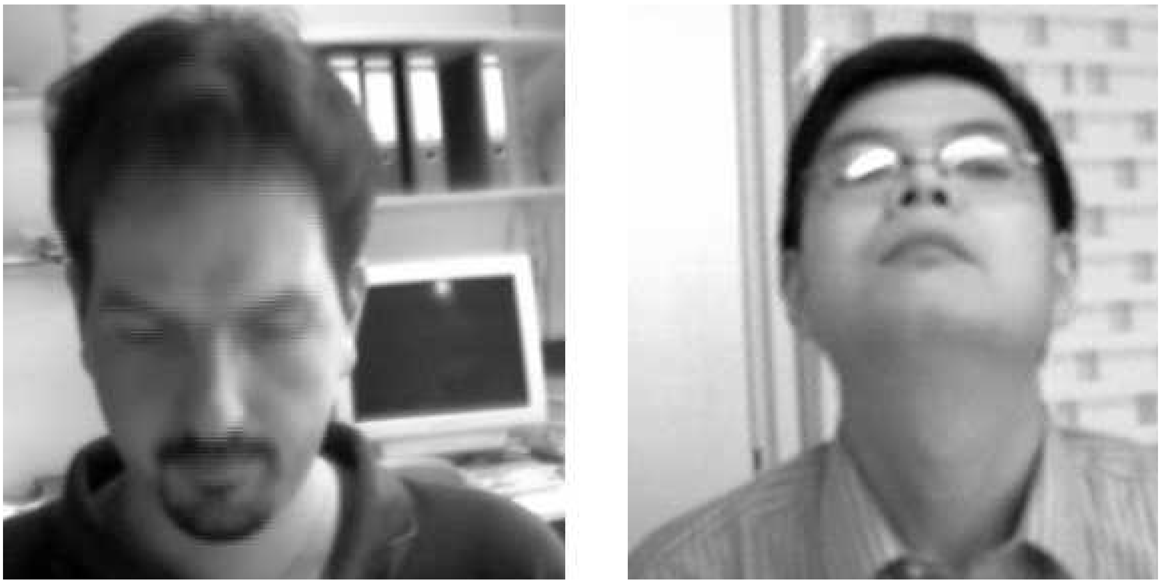


(b) Combined same face decision result

Figure 8.7 Face recognition performance test--yaw angle variance

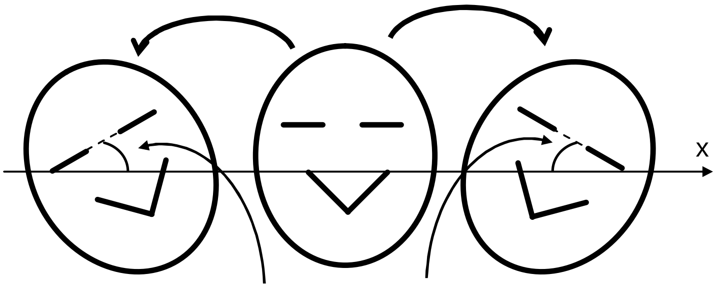
- **Influence of yaw angles**: The classifier is quite sensitive to this kind of head rotations. Since 2D image analysis is applied and this pose change is actually 3D. It is difficult to show the precise yaw angle limit of the classifier. We are hence using a set of sequences to visually demonstrate it. Figure 8.7(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 degrees 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 Figure 8.8(a). The artifact is subject to be produced due to the poor performance of cheap cameras or cheap video capture card. 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.



(a) (b)

Figure 8.8 Examples of artifacts around the eyes which cause recognition errors



Roll angle *q*

Figure 8.9 Definition of roll angle

- **Influence of roll angles**: Since roll is the in-plane head movement, it can be precisely measured in the 2D image. Figure 8.9 illustrates the definition of roll angle ?. 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[4]. 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

---

[4] A freely available graphic tool, available at http://www.irfanview.com/

using the classifier alone. Figure 8.10 shows such a comparison with two face images. From the figure, we draw the conclusion that the classifier cannot recognize the face with *?* bigger than 15° when the similarity threshold is set to be 80%. For faces with 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.

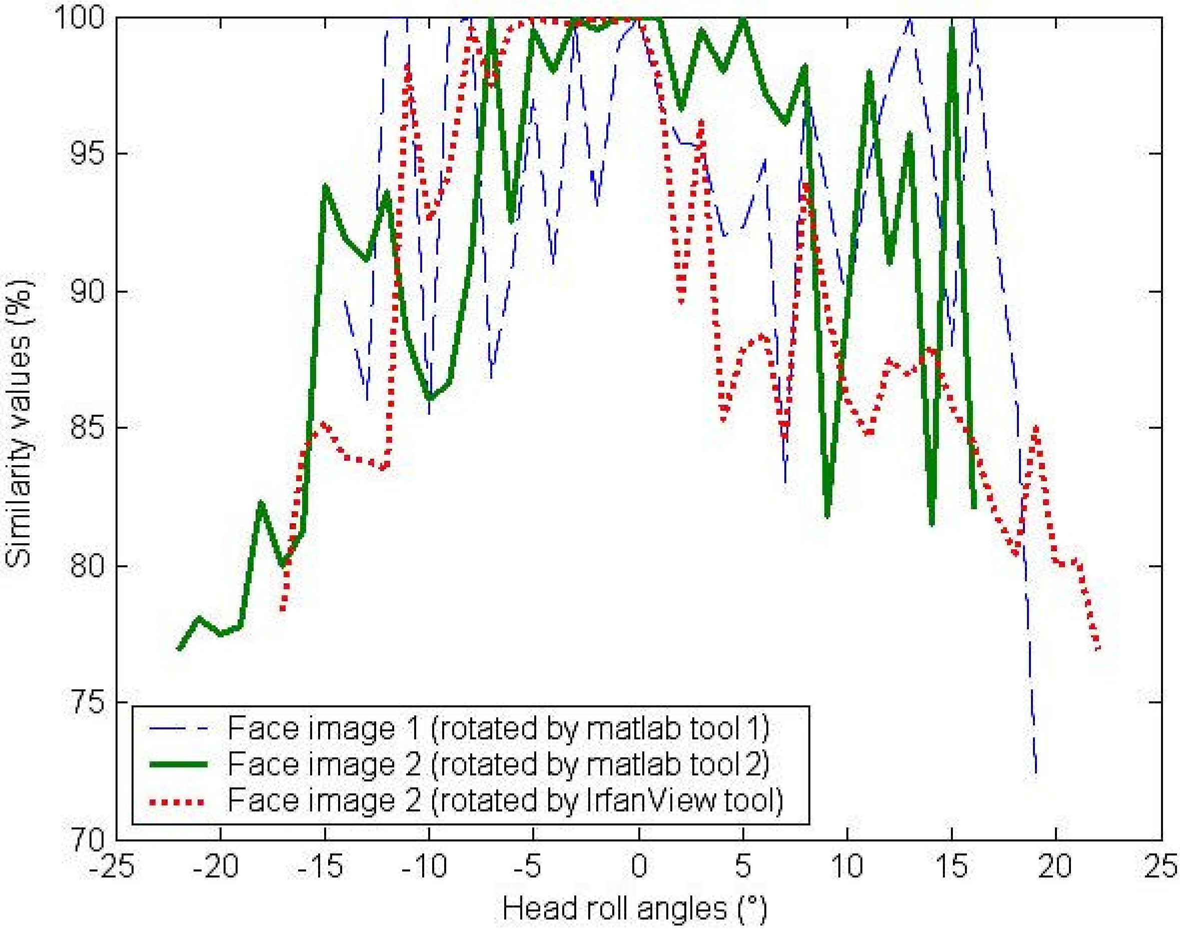
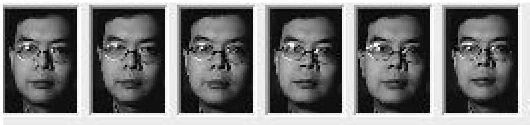


Figure 8.10 Roll angle influence on the recognition classifier

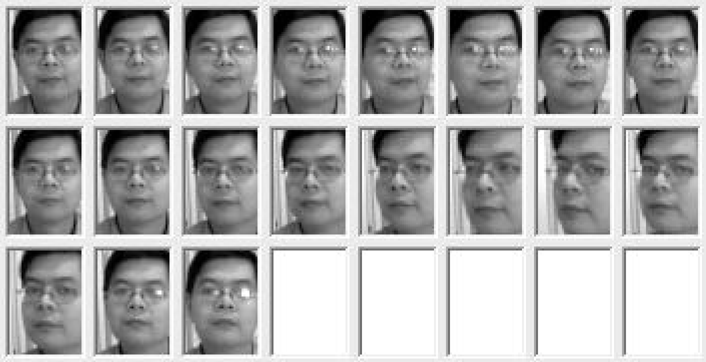
8.4 Performance of Database Construction Algorithms

We have introduced the features for an optimum face database in chapter 5.4. In this section, we are evaluating these characteristics.

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. Figure 8.11 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 figure (a), while 19 out 100 face shots have been selected in (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

Figure 8.11 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 Figure 8.11 (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 listed in Figure 8.12. There are quite a lot differences in the source sequences. But the same person is always successfully recognized, demonstrating a satisfactory uniqueness.

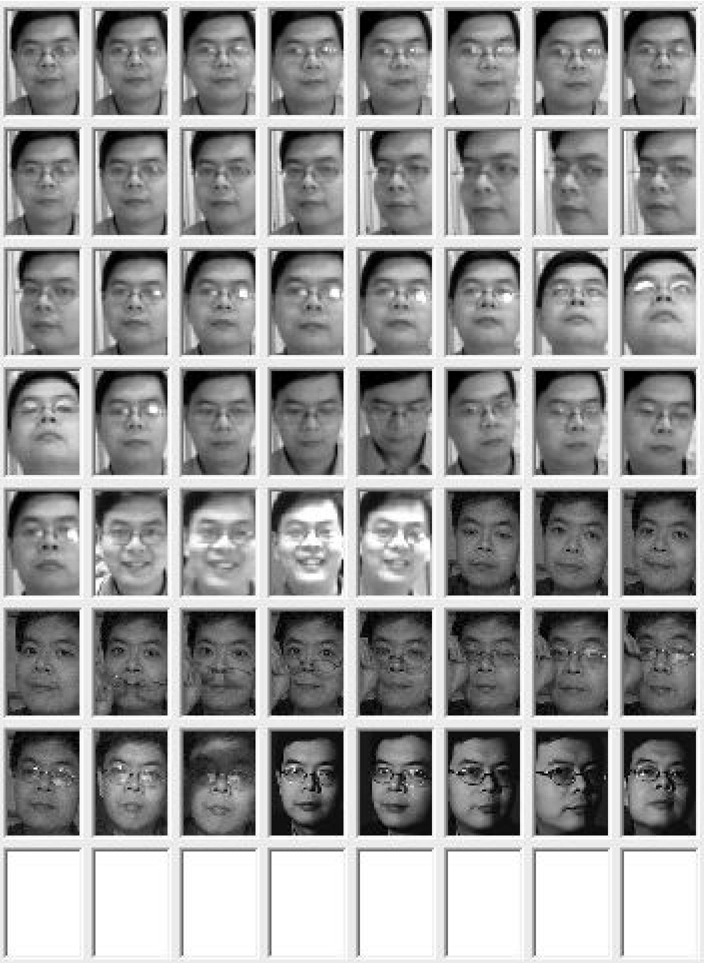


Figure 8.12 Updatability and uniqueness of database construction

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

8.5 Overall Performance of the Whole System

Although there are some standard databases available ([86], [87], [88]) 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 performance and make quantitative comparisons with existing methods. Therefore, we have made our own sequences for the overall system evaluation.

8.5.1 Online version

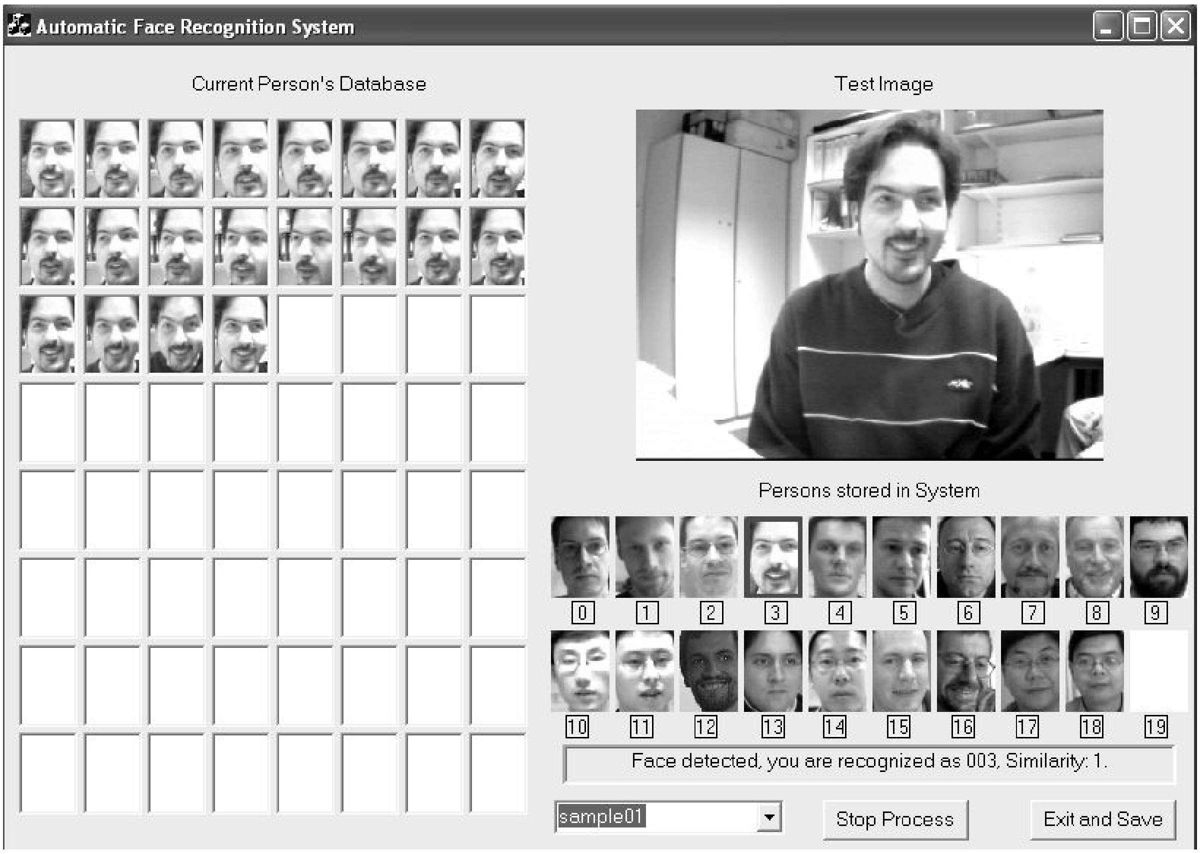
As the first group of the experiments, we have fed the online TV news and online video from cameras with different resolutions.

Figure 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 a 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 19 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 (FRR) is high. But the harmful FAR keeps zero, which is quite satisfying.

8.5.2 Offline version

For the offline version, we have asked 20 subjects for recording sequences in a span of two years. During the recording, we did not require any cooperation from the subjects and did not provide any manual operation. 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. But there is one person Peron 000 and Person 002 have been falsely recognized as two persons. That is due to the same reason as described in chapter 8.4. With further experiments, we have observed that when the starting sequence of a certain person provides more than 30 face shots, FRR can be kept reasonably low. This requirement is not highly restricted. It corresponds to around 1 minute period when a person is talking to someone.



Face detected, you are recognized as 003, Similarity: 1.

Figure 8.13 Offline performance of the whole system

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°.

Chapter 9 Conclusions and Future Directions

9.1 Conclusion

In this dissertation, we have proposed a promising video-based face recognition system for consumer electronics, which features self-learning and intelligent. It 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. Experiments show that the system can deal with two major challenges in the state-of-the-art face recognition research: the aging problem and the difficulty of no better recognition quality in videos than in images.

The main contributions of the algorithms for such an autonomous recognition system are:

- Novel combined face detection methods for improving both face detection and recognition rate from video
- Combined same face recognition algorithms for face recognition from video
- Adaptive face database construction algorithms and database structures for high quality face recognition
- State machines contributed to the whole automatic procedure.

9.2 Future Directions

There are several proposals for the future exploration.

Our system has no problem to deal with the case that multiple people show up at the same time. But it can only recognize one salient face from a single frame. It will be quite interesting to include the function of recognizing multiple people at the same time. In chapter 3.5, we have briefly mentioned a possible method to detect and extract multiple faces in one frame. Since it isn’t general enough, we haven’t implemented it in the current system. A more careful research and related implementation are to be considered. Another possibility is to apply an image-based multiple face detector directly, which is in principle a better way.

In chapter 5.4, we have mentioned the way of merging falsely separated databases that are from the same person. Since it does not frequently occur when 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 explorations are hence required in this direction.

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 dissertation, we are concentrating on face recognition. But the proposed rules and algorithms can be more widely applied for any biometric recognition which requires automatic and self-learning characteristics. A wider range of application is to be further explored.

References

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

[2] 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)*, pp. 191-198, 2003.

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

[4] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillipsl, “Face Recognition: A Literature Survey”, *Technical Report (CS-TR-4167R), University of Maryland*, August 2002. Available at <ftp://ftp.cfar.umd.edu>.

[5] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, J. M. Bone, “FRVT 2002 Evaluation Report”, *Technical Report*, March, 2003, available at <http://www.frvt.org>.

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

[7] D. Mou, R. Schweer, A. Rothermel, “Automatic Databases for Unsupervised Face Recognition”, *IEEE International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’04)*, Vol. 5, p.90, June 27th-July 2nd, 2004.

[8] D. Mou, R. Schweer, A. Rothermel, “Face Recognition System and Method”, Patent application, European Patent Office, EP05/001988, filed on Feb. 25th, 2005.

[9] D. Mou, R. Schweer, A. Rothermel, “A Self-learning Video-based Face Recognition System”, paper accepted in *IEEE International Conference on Consumer Electronics 2006*, Jan. 7-11, 2006.

[10] B. Cumbers, “Passive Biometric Customer Identification and Tracking System”, *U.S. Patent*, patent No. 6554705, Apr. 29, 2003.

[11] Y.T. Lin, “Adaptive Facial Recognition System and Method”, *U.S. Patent application publication*, Pub. No.: US2002/0136433, Pub. Date: Sep. 26, 2002.

[12] 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)*, pp.1202-1207, Vol.2, 2002.

[13] E. Hjelmas and B.K. Low, “Face Detection: A Survey,” *Computer Vision and Image Understanding*, pp. 236-274, 2001.

[14] Ming-Hsuan Yang; D.J. Kriegman, N Ahuja, “Detecting faces in images: a survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.34-58, Vol.24, No. 1, Jan. 2002.

[15] M. Yang; “Recent Advances in Face Detection”, *IEEE ICIP 2003 Tutorial*, Barcelona, Spain, September 14, 2003, available at http://vision.ai.uiuc.edu/mhyang/face-detection-survey.html.

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

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

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

[19] G. Yang and T. Huang, “Human Face Dtection in Complex Background”, *Pattern Recognition*, Vol. 27, No.1, pp. 53-63, 1994.

[20] 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, pp.2537-2540, 1997.

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

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

[23] R, Féraud, O. Bernier, J. Villet, and M. Collobert, “A Fast and Accurate Face Detector Based on Neural Networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 42-53, Vol. 22, No. 1, January, 2001.

[24] 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*, pp. 130-136, 1997.

[25] 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, pp. 746-751, 2000.

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

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

[28] M. Jones, P. Viola, “Fast Multi-view Face Detection”, *Mitsubishi Electric Research Laboratories Technical Reports*, TR2003-96, 2003, available at http://www.merl.com/reports/docs/TR2003-96.pdf.

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

[30] M. Turk and A. Pentland, “Eigenfaces for Recognition”, *Journal of Cognitive Neuroscience*, Vol.3, pp.72-86, 1991.

[31] *S. Palanive, B.S. Venkatesh, B. Yegnanarayana*, “Real time face recognition system using autoassociative neural network models”, *IEEE Conference Proceedings on Acoustics, Speech, and Signal Processing (ICASSP '03)*, pp.833-836, vol.2, 6-10 April 2003.

[32] T. kim, S. Lee, J. Lee, S. Kee, and S. Kim, “Integrated approach of multiple face detection for video surveillance”, *Proceedings, IEEE 16th Conference on Pattern Recognition*, pp. 394-397, vol.2, 2002.

[33] D. Butler, C. McCool, M. McKay, S. Lowther, V. Chandran, and S. Sridharan, “Robust Face Localisation Using Motion”, Colour & Fusion *Proceedings, Digital Image Computing: Techniques and Applications (DICTA 2003)*, pp.899-908, 2003.

[34] C. Wren, A. Azerbayejani, T. Darrell, and A. Pentland, “Pfinder: A Real-Time Tracking of Human Body”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.

[35] 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*, pp. 228-233, 1998.

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

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

[38] *R.C. Verma, C. Schmid, K. Mikolajczyk*, “Face detection and tracking in a video by propagating detection probabilities”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1215-1228, Vol.25, Issue10 ,Oct. 2003.

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

[40] K. Bowyer, K. Chang, and P. Flynn, “A Survey Of 3D and Multi- Modal 3D+ 2D Face Recognition”, *Notre Dame Department of Computer Science and Engineering Technical Report*, January 2004.

[41] 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.

[42] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, “The FERET Evaluation Method for Face Recognition Algorithms”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090-1104, 2000.

[43] D. M. Blackburn, M. Bone and P.J. Phillips, “FRVT 2000 Evaluation Report”, *Technical Report*, Feb. 16th, 2001, available at http://www.frvt.org.

[44] 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*, pp. 202-207, 2002.

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

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

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

[48] M.S, Bartlett, J.R. Movellan and T.J. Sejnowski, “Face Recognition by Independent Component Analysis”, *IEEE Transactions on Neural Networks*, 13(6), pp.1450-1464, 2002.

[49] 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, pp. 831-836, Aug. 1996.

[50] P. Belhumeur, J.P. Hespanha, D.J. Kriegman , “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[51] 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, pp. 2721-2724, May 1998.

[52] 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.

[53] 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)*, pp.305-310, 2004

[54] 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)*, pp. 52-56, 2003.

[55] L. Wiskott, J. Fellous, N. Krüger and C. v.d. Malsburg, “Face Recognition by Elastic Bunch Graph Matching”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775-779, 1997.

[56] C. Liu and H. Wechsler. “A Gabor Feature Classifier for Face Recognition”, *Proceedings of Eighth IEEE International Conference on Computer Vision*, vol. 2, pp. 270 –275, 2001.

[57] B.A. Draper, K. Baek, M.S. Bartlett and J.R. Beveridge, “Recognizing faces with PCA and ICA”, *Computer Vision and Image Understanding*, vol. 91, pp.115-137, 2003.

[58] A.M. Martinez and A.C. Kak, “PCA versus LDA”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), pp. 228-233, 2001.

[59] 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)*, pp. 215-220, May 2002.

[60] 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, pp. 117-126, Jan. 2003.

[61] Y. Zhang, L. Lang and O. Hamsici, “Subspace Analysis for Facial Image Recognition: A Comparative Study”, available at http://www.stat.ohio-state.edu/~goel/STATLEARN/.

[62] 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*, 19(9-10):631–638, 2001.

[63] 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*, pp.688-694, 2001.

[64] J. Huang, V. Blanz, and B. Heisele, “Face Recognition Using Component-Based SVM Classification and Morphable Models”, *SVM 2002*, LNCS 2388, pp. 334– 341, 2002.

[65] S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back, “Face Recognition: A Convolutional Neural Network Approach”, *IEEE Transactions on Neural Networks*, Vol. 8, Nr. 1, pp. 98–113, 1997.

[66] T. Kurita, M. Pic, 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)*, pp.53-58, 2003.

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

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

[69] 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)*, pp.813-818, 2004.

[70] 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)*, pp.345-349, 2004.

[71] 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, p.70, June 27th-July 2nd, 2004.

[72] 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*, pp. 384-389, 1998.

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

[74] H. Wechsel, V. Kakkad, J. Huang, S. Gutta, and V.Chen, “Automatic Video-Based Person Authentication Using the RBF Network”, *Proceedings of 1st International Conference on Audio And Video-based Biometric Person Authentication*, pp. 85-92, 1997.

[75] C. Lambert, “Autonomous Face Recognition Machine”, *U.S. Patent*, patent number: 5012522, date of patent: Apr. 30, 1991.

[76] J.L. Center JR., “Real-time Facial Recognition and Verification System”, *U.S. Patent application publication*, publication number: US2003/0059124, Pub. Date: Mar. 27, 2003.

[77] 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*, pp.516-521, 1998.

[78] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, J.M. Buhmann, “Topology Free Hidden Markov Models: Application to Background Modeling”, *IEEE 8th Conference on Computer Vision (ICCV 2001)*, pp. 294-301, vol.1, 2001.

[79] http://www.cognitec-systems.de/brochures/FaceVACSalgorithms.pdf.

[80] http://www.portrait-artist.org/face/structure4.html.

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

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

[83] T. Kondo and H. Yan, “Automatic human face detection and recognition under nonuniform illumination”, *Pattern Recognition* 32, pp.1707–1718, 1999.

[84] 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, pp.322–336, 2000.

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

[86] BioID Face database, available at http://www.humanscan.de/support/downloads/facedb.php.

[87] CMU/VASC image database, available at http://vasc.ri.cmu.edu/idb/html/face/index.html.

[88] AT&T face database, (formerly “The ORL Database of Faces”) available at http://www.uk.research.att.com/facedatabase.html.

[89] Christophe Garcia , “A Survey of Face Detection and Recognition Techniques”, *France Telecom R&D, Séminaire CNRT TIM – Vision par ordinateur pour les Télécommunications*, May 18th, 2004.

[90] G. Potaminanos, C. Neti, G. Gravier, A. Garg, A. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, *Proceedings of the IEEE*, Vol. 91, Issue 9, Sep. 2003.

[91] E. Acosta, L. Torres, A. Albiol, Ed Delp, “An Automatic Face Detection and Recognition System for Video Indexing Applications”, *IEEE Conference on Acoustics, Speech and Signal Processing*, volume 4, Orlando, Fl, pages 3644-3647, 2002.

[92] L. Torres, J. Vilà, “Automatic Face Recognition for Video Indexing Applications”, *Pattern Recognition*, Vol 35/3, pp. 615-625, December 2001.

[93] Yee Whye Teh, Geoffrey E. Hinton, “Rate-coded Restricted Boltzmann Machines for Face Recognition”, *Advances in Neural Information Processing Systems 13*, Papers from Neural Information Processing Systems (NIPS) 2000, pp. 908-914, Denver, CO, USA. MIT Press 2001.

[94] M.Castrillón, O. Déniz, D. Hernández and A. Domínguez, “Identity and Gender Recognition Using the ENCARA Real-Time Face Detector”. *X Conferencia de la Asociación Española para la Inteligencia Artificial*, CAEPIA, San Sebastián, 12-14 Nov., 2003.

[95] A.R. Chowdhury, R. Chellappa, S. Krishnamurthy and T. Vo., “3D Face Reconstruction from Video Using A Generic Model”, *Proceedings of International Conference on Multimedia and Expo*, pp. 1007-1012, Aug. 26-29, 2002.

[96] K. Lee, J. Ho, M. Yang, and D. Kriegman, “Video-based face recognition using probabilistic appearance manifolds”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, Vol.1, pp.313-320, June 18-20, 2003.

[97] A.J, Howell and H. Buxton, “Active vision techniques for visually mediated interaction”, *16th Proceedings of IEEE Conference on Pattern Recognition*, Vol. 2, pp. 296-299, 2002.

[98] S. Zhou and R. Chellappa, “Probabilistic Identity Characterization for Face Recognition”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, Vol.2, pp.805-812, June 27-July 2, 2004.

[99] http://home.t-online.de/home/Robert.Frischholz/index.html.

[100] K. Baek, B.A. Draper, J. R. Beveridge, K. She, “PCA vs. ICA: A comparison on the FERET data set”, *6th Joint Conference on Information Sciences*, March 8-14, pp.824-827, 2002.

[101] C. Garcia, G. Zikos, and G. Tziritas, “A wavelet-based framework for face recognition”, *Proceedings of the Workshop on Advances in Facial Image Analysis and Recognition Technology, 5th European Conference on Computer Vision (ECCV'98)*, pp. 84-92, June 1998.

Curriculum Vitae

Personal Data:

Name:	Dengpan Mou
Gender:	Male
Date of Birth:	Jan. 19th, 1976
Place of Birth:	Shandong, China
Nationality:	Chinese

Education:

Sep. 2000 ~ Sep. 2004, PhD candidate,
Microelectronics Department, University of Ulm. Title: Autonomous Face Recognition.

Jul. 1998 ~ Jun. 2000, M.S.,
University of Ulm. Title: Implementation of an Improved Video Sync Signal Processing.

Sep. 1993 ~ Jun. 1998, B.S.,
Major in Microelectronics Technology, minor in Computer & Its Application, Dept. of Electronic Engineering, Beijing Polytechnic University, Beijing, China.

Sep. 1990~Jun. 1993, high school,
High School Attached to Tsinghua University, Beijing, China.

Professional experience:

Oct. 2004 ~ present,
System engineer, Harman/Becker Automotive GmbH XSYS Division, Villingen, Germany

Sept. 2000 ~ Sept. 2004
Research Assistant, Microelectronics Department, University of Ulm

Major Honors & Awards:

2003,	Kooperationspreis Wissenschaft/Wirtschaft 2003 der Universität Ulm
2002,	DAAD-Preis 2002 für hervorragende Leistungen ausländischer Studierender
1995~1996,	Second Prize in Beijing, Chinese National University Students Modeling Contest of Mathematics

Social Activity:

2000~2003,	Chairman, Association of Chinese Students and Scholars in Ulm (ACSSU)
2002~2003,	"Ambassador" to facilitate the academic cooperation between University of Ulm, Germany and Shandong University, China
2001-2003,	Vice president, Chinesischer Verein für Süddeutschland .e.V.

Publications:

[1] D. Mou, R. Lares, W. Yan, F. Rominger, A. Rothermel, "Design and Implementation of a Novel Sync Processing System for Composite Video Signals", IEEE Transaction on Consumer Electronics, pp.1286-1291, Vol. 49, Num. 4, Nov. 2003.

[2] D. Mou, R. Schweer, A. Rothermel, "Automatisches System zur Erkennung und Verwaltung von Gesichtern bzw. Personen", 21. Jahrestagung der FKTG (Fernseh-und Kinotechnische Gesellschaft e.V.), May. 24-27, Koblenz.

[3] D. Mou, R. Schweer, A. Rothermel, "Automatic Databases for Unsupervised Face Recognition", submitted to 1st IEEE Workshop on Face Processing in Video (FPIV'04), in conjunction with IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'04), June, 2004.

[4] D. Mou, R. Schweer, A. Rothermel, "A Self-learning Video-based Face Recognition System", paper accepted in IEEE International Conference on Consumer Electronics 2006, Jan. 7-11, 2006

Patents:

D. Mou, R. Schweer, A. Rothermel, "Face Recognition System and Method", Patent application, European Patent Office, EP05/001988, filed on Feb. 25th, 2005.