2012

Fast, collaborative acquisition of multi-view face images using a camera network and its impact on real-time human identification

Rohith Bakkannagari
West Virginia University

Fast, collaborative acquisition of multi-view face images using a camera network and its impact on real-time human identification

by

Rohith Bakkannagari

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Electrical Engineering

Dr. Natalia A. Schmid, Ph.D.
Dr. Matthew C. Valenti, Ph.D.
Dr. Vinod Kulathumani, Ph.D., Chair

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2012

Keywords: Camera network, Multi-view face detection, Face recognition, Acquisition system, Fusion

Abstract

Fast, collaborative acquisition of multi-view face images using a camera network and its impact on real-time human identification

by

Rohith Bakkannagari
Master of Science in Electrical Engineering

West Virginia University

Dr. Vinod Kulathumani, Ph.D., Chair

Biometric systems have been typically designed to operate under controlled environments based on previously acquired photographs and videos. But recent terror attacks, security threats and intrusion attempts have necessitated a transition to modern biometric systems that can identify humans in real-time under unconstrained environments. Distributed camera networks are appropriate for unconstrained scenarios because they can provide multiple views of a scene, thus offering tolerance against variable pose of a human subject and possible occlusions. In dynamic environments, the face images are continually arriving at the base station with different quality, pose and resolution. Designing a fusion strategy poses significant challenges. Such a scenario demands that only the relevant information is processed and the verdict (match / no match) regarding a particular subject is quickly (yet accurately) released so that more number of subjects in the scene can be evaluated.

To address these, we designed a wireless data acquisition system that is capable of acquiring multi-view faces accurately and at a rapid rate. The idea of epipolar geometry is exploited to get high multi-view face detection rates. Face images are labeled to their corresponding poses and are transmitted to the base station. To evaluate the impact of face images acquired using our real-time face image acquisition system on the overall recognition accuracy, we interface it with a face matching subsystem and thus create a prototype real-time multi-view face recognition system. For front face matching, we use the commercial PittPatt software. For non-frontal matching, we use a Local binary Pattern based classifier. Matching scores obtained from both frontal and non-frontal face images are fused for final classification. Our results show significant improvement in recognition accuracy, especially when the front face images are of low resolution.

Acknowledgements

This thesis would not have been possible without the guiandance from several individuals who directly or indireclty supported me throughout this study.

I would like to express my deepest appreciation to my advisor, Dr. Vinod Kulathumani for granting me this opportunity. His supervision and guidance from the very early stage of my research is worth mentioning. His expertise, valuable suggestions and insights are truly commendable.

I am grateful to Dr. Natalia Schmid and Dr. Matthew Valenti for their unconditional support and time throughout the program.

Many thanks would go to my research group memebers Srikanth Parupati, Sriram Sankar, Rahul Kavi and Terry Ferrett for their encouragement and support. They played their part in making the work environment fun and able place.

I am also grateful to the Library Enhancement and Design group at The MathWorks for offering me an internship. I got to work with some of the modern tools and techniques used in the industry. This definitely would be a good first step in my professional career.

Finally, a special thanks to my parents, family members and friends who always stood by my side at all times. I am always indebted to them.

Contents

List of Figures

List of Tables

Notation

We use the following notation and symbols throughout this thesis.

F - Fundamental matrix

l - Epipolar line

Q - Message queue length

NTP - Network time protocol

fps - Frames processed per second

B_{ff} - Frontal face buffer

B_{sf} - Side face buffer

t_{nd} - Network delay in ms

t_{ff} - Average time to detect a face in a background subtracted image in ms

t_{sf} - Average time to detect a side face

t_s - Network clock synchronization error in ms

N_{ff} - Frontal face processing rate

N_{sf} - Side face processing rate in fps

$t(x)$ - Time stamp of a frame

$w(x)$ - Width of a detected face

s_k - Similarity score

d_k - Dissimilarity score score

χ^2 - Chi-square statistical estimate for dissimilarity measure

H_{ij} - Histogram frequencies

s_f - Matching score for frontal face image

s_{pp} - Matching score for partial profile face image

s_p - Matching score for profile face image

s_o - Overall score

w_f - Weight assigned to frontal face image

w_{pp} - Weight assigned to partial profile face image

w_p - Weight assigned to profile face image

FRR - False rejection rate

FAR - False acceptance rate

GAR - Genuine acceptance rate

Chapter 1

Introduction

1.1 Motivation

Face recognition systems have evolved into a reliable mechanism for establishing identity of individuals and countering fraud. They find applications in access control, surveillance, border security, smart cards etc. Face recognition systems have been traditionally designed to operate in unconstrained scenarios. Most of the data processing and computation is done offline that hinders the opportunity for real-time identification. However, there is a need to operate face recognition systems in unconstrained scenarios and in real time. Recent terror attacks, security threats, intrusion attempts and criminal activities have further stipulated the need for such biometric systems.

Realizing such a system using a single camera suffers from several limitations. The system is highly sensitive to small changes in pose and illumination. Further, the performance of the system is marred significantly if the subject is not cooperative. If a person is occluded then there is no way to detect that person and it provides coverage to a specific region. Multi-view camera network, where each camera is capable of processing locally are designed to meet the requirements. They can be deployed to provide coverage from different views of the scene, thus providing tolerance against variable pose, poor illumination and possible occlusions. But designing such a system poses several challenges. Video data is computationally intensive. In order to be scalable and operate in real-time there is a tradeoff between local processing and the computational burden at the central location. Care should be taken not to burden

the individual nodes and the central location. At the same time, individual nodes or units are required to sample as many frames as possible so as to not miss any information as the person is constantly moving. This entitles the requirement for a robust and reliable multi-view camera acquisition system to acquire as many frames as possible and a fusion scheme to effectively process the frames form multiple views. Images of different quality (pose, illumination and resolution) continually stream in from multiple cameras which need to be efficiently utilized.

Typical example would be to cover an area of interest such as monitoring a critical region using multiple cameras with overlapping field of views. This way, even if one or more cameras view is occluded or doesn't function there is a possibility to arrive at the result based on the evidences from the other cameras in the network. The limitations on pose variations, lightning conditions, facial expressions are somewhat minimized. Further, making a decision based on multiple evidences of the same object instills confidence. Multi-view camera network can potentially improve the accuracy of the human identification. Fusion scores across multiple views tend to enhance the recognition accuracy. Generally, frontal face images are the most suitable ones for reliable face recognition. However, in unconstrained environments it is not always possible to obtain enough high quality frontal face images required for accurate recognition. Under such circumstances, non-frontal face images acquired from a camera network can be used to improve the confidence of face recognition from frontal faces. Often profile (side view) face images contain moles and special markers that are useful in human identification [1]. Recent studies have shown that profile views and partial profile views can be used for reliable face recognition with high accuracy [2, 3, 4, 5]. That being said, acquiring multi-view face images is a challenging task in terms of computational overhead especially due to their diversity [6]. Typically, separate face detectors are trained for each pose that are then sequentially or hierarchically applied on each frame to detect a face [7, 8, 9, 10]. Alternatively, a pose classifier is first applied through a sliding window of different sizes (that could fit a face) and then the appropriate face detector for that pose is used to detect the presence of a face [7]. Both of these approaches involve significant image processing and unsuitable when we would like to maximize the number of faces acquired for recognition

1.2 Thesis contributions

We design a multi-view face acquisition system to acquire the faces accurately and at a rapid rate. We devise a fusion strategy to combine the acquired data effectively and to reach a decision quickly (yet accurately). Finally, we analyze the impact of multi-view faces on identification.

Our multi-view face acquisition system consists of 3 cameras, which can simultaneously take three views of a face at different angles. The data acquisition system exploits the geometry of multi-camera network to collaboratively acquire both frontal and non-frontal face images in real-time while maintaining a high sampling rate. An overview of our approach is as follows. We first train face detectors based on Haar-like features [11, 12] for each pose class that is required to be detected. We then run a frontal-face detector on each camera in the network. Whenever a frontal face has been detected on any camera, say C_f , it sends a notification to other cameras which then narrow down their search to the region surrounding the epipolar line corresponding to the point where the frontal face is detected in C_f. By applying a pose specific face detector on this much smaller region in the image, the cameras are able to quickly extract non-frontal face images and simultaneously index these faces into the corresponding face pose. Thus, we utilize the multi-view camera geometry and inter-camera communication to reduce the amount of image processing required for multi-view face detection. Using this we are able to process an image for detecting non-frontal faces at almost the same rate as for frontal faces. At the same time, by narrowing down the potential regions in an image for non-frontal face detection, we significantly improve the reliability of non-frontal face detection. Our system is easy to setup, does not require camera calibration and only depends on fundamental matrices of transformation between camera pairs.

To evaluate the impact of face images acquired using our real-time face image acquisition system on the overall recognition accuracy. We interface it with a face matching subsystem and thus create a prototype of real-time multi-view face recognition system. For front face matching, we use the commercial PittPatt software [13]. For non-frontal matching, we use a Local binary Pattern [14] based classifier. Matching scores obtained from both frontal and non-frontal face images are fused for final classification. We tested our prototype

face recognition system using an experiment with 30 human subjects, walking in isolation at different distances from the cameras. Our results show significant improvement in recognition accuracy, especially when the front face images are of low resolution. By improving recognition accuracy at larger stand-off distances and lower image quality, we expect the face recognition system to be applicable for real-time watch-list identification scenarios in unconstrained environments.

1.3 Thesis outline

The rest of the thesis consists of 3 main parts, namely, multi-view face acquisition system, fusion and face recognition system. In chapter 2, the background information and related work has been discussed. Chapter 3 discusses about the data acquisition system design and implementation. Chapter 4 describes about the fusion strategies and its applications. In chapter 5, the face recognition system design and experimental evaluation is presented. Finally, the conclusions and proposed future work are explained in chapter 6.

Chapter 2

Background information and Related work

Biometric system is defined as an automated method that helps recognize people based on physiological or behavioral characteristics. In today's world, these systems play a substantial part and cannot be overlooked.. Biometric systems target a wide-ranging applications rite from border security, monitoring a secure region to smart homes and biometric authentication for PDA's. A variety of identification techniques were developed exploiting the distinct and unique features of a person like face, fingerprint, iris, gait, etc. Each technique has its own merits and demerits. Some of these techniques are intrusive and some are not. For example, retina recognition is intrusive and capturing the retina sample may cause inconvenience to the user. On the other hand, face recognition is non-intrusive and passive and the image of the face can be captured from a distance without user intervention or not causing inconvenience to the subject. Face recognition has received a significant amount of attention over the years due to it being non-intrusive, non-contact process and reliable. Criminal identification, personnel screening and surveillance are some of the typical applications where face recognition is primarily employed.

Fig. 2.1 describes the face recognition system as a three stage process. The first stage is face detection which extracts faces from a scene. It is followed by feature extraction stage which involves extracting relevant features for further analysis. The last stage is face recognition where identification or verification is carried out. These three steps can be

merged and/or new stages may be added. A substantial amount of work has been done on both face detection and face recognition. Robust multi-view face recognition system relies on how well the face detection and face recognition are coupled.

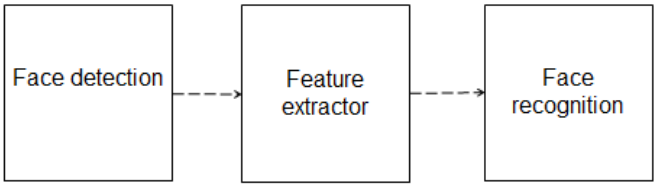

Figure 2.1: A generic face recognition system

2.1 Face detection

In [11], Viola and Jones have designed a face detector that is suitable for real-time frontal face detection. Their approach utilizes the AdaBoost algorithm to identify a sequence of Haar like features that indicate the presence of a face. Since then other frontal face detection algorithms have been developed, a survey of which is presented in [15]. Approaches for multiview face detection have generally been of two types. The first approach is to estimate the pose over each sliding window in an image (which may not necessarily have a face) and then applying the pose specific detector [7]. When involving multiple face poses this is a hard problem and moreover false estimates of a face pose will lead to incorrect detection of a face. In the second approach, different view-specific face detectors are applied sequentially or hierarchically to an image [7, 16, 17, 6]. In this thesis, we have used an OpenCV [18] implementation of the frontal face detector presented in [11] and trained pose-specific detectors using Haar like features for non-frontal faces as described in [12]. Then, we have used the information about a detected frontal view along with relative camera orientations and the subject location to detect non-frontal faces in other cameras and we observe that our approach decreases overall processing time.

2.2 Feature extraction

Feature extraction [19, 20] is defined as a process of extracting relevant information from the input image. To a large extent, input data is highly redundant and large. Processing such a data utilizes good amount resources and is time consuming. In the feature extraction stage, the high dimensional and redundant input data is transformed into a low dimensional and unique set of features (also called as feature vector). The *curse of dimensionality* problem is addressed with dimensionality reduction. Characteristics like localization of eyes, nose, mouth, texture etc in the face image are the features pertaining to the face image. All or subset of which form the feature vector. In the matcher, the feature vector is compared with the feature vectors extracted from the samples in the database to perform identification or authentication.

2.3 Face recognition

Considerable amount of work has been done over the years on face recognition. Traditional face recognition was done on 2D images, focusing on frontal views. 2D face recognition methods suffer from pose and illumination changes. 3D face recognition methods are invariant to changes in pose but are either slow or not accurate. Hence are not appropriate for real-time applications. A large number of these techniques work effectively with frontal views only. When these techniques or algorithms are used with non-frontal views they tend to fair badly. The performance drop with non-frontal views is due to the fact that non-frontal views have highly non- linear features which are hard to resolve.

Facial recognition can be image based or video based. Recognition on still images is termed as image based recognition and on video sequences is known as video based recognition. Many traditional methods were focused on still images. Later, face recognition using video sequences have become popular [21, 22]. We primarily focus on image based approaches. Image based face recognition techniques can be further classified(illustrated

in Fig. 2.2) into appearance based or hollistic methods and model based or feature based methods. Principal Component Analysis (PCA) [23], Linear Discriminant Analysis (LDA) [24], Local Binary Patterns (LBP) [25] and independent Component Analysis (ICA) [26] come under appearance based methods and Elastic Bunch Graph Model (EBGM) and 3D Morphable Model come under the category of model based face recognition.

Principal Component Analysis (PCA) is the one of the major developments in

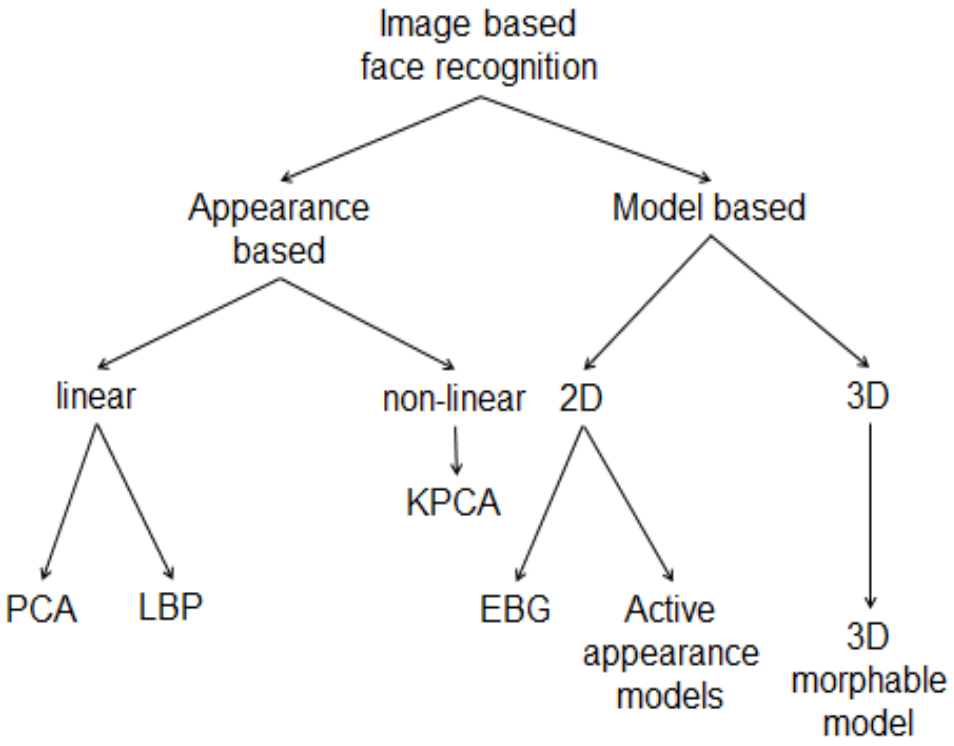

Figure 2.2: Classification of image based face recognition approaches

face recognition and is the first approach that is based on eigen faces. Later, a number of techniques were proposed based on PCA. PCA also known as *eigenface* method is used for image recognition and also for compression. The main idea here is to reduce the large dimensionality of the data space to low dimensionality feature space. This is possible when the data is correlated. The high dimensional data of the whole face image is projected onto a low dimensional subspace or feature space using a transformation. Linear Discriminant Analysis (LDA) and independent Component Analysis (ICA) can be somewhat called refinements to the PCA. LDA tries to find out a linear transformation that would best discriminate among

the classes or maximize the between class variance and minimize the in-class variance. ICA which provides a more powerful data representation differs from the PCA in that PCA considers image elements as random variables with Gaussian distribution and minimizes the second-order statistics, while ICA tends to look for components that are independent and non-gaussian.

A number of face recognition techniques like EBGM, LBP, KPCA and including those discussed above have a requirement that the image has to undergo pre-processing before it is used for recognition. They are sensitive to pose variations, illumination changes and image alignment. As a result, face recognition systems using these techniques have an additional overhead of implementing a pre-processing step, making these techniques unreliable and not so robust. Accordingly, they are not suited for automatic face recognition systems when used in isolation. Addressing these drawbacks companies started to invest in building complete recognition software that would do image pre-processing in addition to recognition. This resulted in development of commercial software like Faceit, PittPat [13], etc which compensate the above discussed drawbacks to a large extent.

Face recognition approaches can be split into two, single view based and multi-view based. In single view based approach we are matching the test image to the corresponding gallery of images with the same pose. On the other hand, in the multi-view based approach the training is done using multi-view face images and thus the test image is compared to the multi-view face gallery. Many face recognition systems with frontal view faces have been extensively studied [23, 25, 24]. Multi-view face recognition is a challenging task than the single view face recognition owing to the fact that multi-view face images have non-linear manifolds that exist in the data space. Multi-view faces have both frontal and non-frontal views. Multi-view face recognition is studied in [27, 28].

In general, face recognition algorithms developed over the years are more effective or tend to perform better when operated on frontal views. This is partly due to the fact that frontal view is likely to have more features than the non-frontal view and recognition using frontal views is less sensitive to pose variations and image alignment compared to the non-frontal views. Conventional face recognition systems carry out identification or authentication using single view (preferably frontal view) of a person's face. For these systems

to be reliable it is necessary to obtain a good quality face image. In real unconstrained scenarios, it is not always possible to have a good quality frontal image. As a result, these systems become less accurate and unreliable. These limitations can be overcome if we can establish the identity of an individual or perform authentication using multiple views of the person. By doing so, even if the frontal image is of low quality the available non-frontal views supplements the recognition accuracy. This way, the accuracy of the system can also be increased and is more robust.

2.4 Multi-view Data Acquisition System

To accomplish the task of obtaining multi-view face images, we design a wireless camera network that exploits the multi-view camera geometry between the cameras to acquire images at a rapid rate. Multi-view camera geometry has been exploited by several recent research efforts to effectively fuse information from different cameras and consequently improve the accuracy in the context of tasks such as object detection, behavior matching, action classification and reliable foreground extraction [29, 30]. By way of contrast, we have utilized multi-view geometry to improve the computational efficiency of the system by collaborating among the cameras in real-time and reducing the amount of image processing required.

In the context of face recognition, multiple cameras have been used for tracking in an active control mode by which one or more cameras are controlled to yield a dynamic coverage [31, 32]. An example of such a system is the combination of a fixed camera and PTZ camera that is used for close-up tracking of humans and subsequent identification. In our approach instead of continuously tracking an individual at close quarters to eventually get a good view that is suitable for recognition, we rely on redundancy offered by multiple camera views to opportunistically acquire a suitable face image for identification [33]. In order to reduce the amount of data transmitted to the base station, the approach taken in this thesis is to use the distributed cameras to perform collaborative face detection [15] and transmit only the region containing faces detected in each frame to the base station.

This is expected to save as much as 95% of the network bandwidth when compared to transmitting raw videos while also reducing the amount of processing required at the base station significantly [34, 33]. Our approach in this thesis is a balance between completely centralized camera network systems for surveillance [35, 36] that process all the data at the base station and completely local approaches (using video analytic cameras [37]) that do not utilize information from multiple views. Balancing centralized processing with local in network processing to reduce network and processing overload has been the focus of several sensor network based data acquisition projects over the past decade [38, 39, 40, 41]. However, achieving this balance for real-time identification with video data is significantly more challenging because of the computationally intensive nature of such data. In our work, we have exploited run-time collaboration between cameras to reduce this local processing time.

2.5 Fusion

Fusion is described as the assimilation of multiple sources of evidence (or information) to arrive at a comprehensive or unified decision (or result). Depending on what stage of the biometric system fusion is carried out, we have five fusion techniques : sensor, feature, score, rank and decision level fusion. Sensor module has the richest source of information and the amount of information is condensed as we move from sensor to decision module. Integrating match scores output from multiple biometric sources is termed as score level fusion. This is also known as measurement level or confidence level fusion. Because of its ease to access and consolidate, score level fusion strategy is widely used. Most existing work on biometric fusion [42, 43, 44] has assumed that data has been acquired a priori and prevailing fusion techniques operate in controlled scenarios. There remains a need for optimizing these techniques for operation in a dynamic mode where data is continually streaming in and each image varies in quality (ambient conditions, pose, resolution). Designing a fusion strategy to process this data and arrive at a decision in real-time poses significant challenges. It is not practical to process all the acquired data and we need to implement a strategy to fuse only the relavent

information. At the same time, it is required that the verdict is made reasonably quickly and is reliable. To address these, we investigate the initial steps towards real-time human identification using a wireless camera network.

2.5.1 Other related work

To facilitate human identification from low resolution videos, many image restoration techniques have been designed based on super-resolution of multiple frames [45, 46]. Also, algorithms have been designed for handling incomplete face data based on a recognition by parts approach [47] and for generating a composite face image based on multiple partial views of a face [48]. Such image restoration and fusion techniques are appropriate for use in conjunction with the distributed face image acquisition framework to enhance the recognition accuracy.

Chapter 3

Collaborative multi-view face acquisition system

This chapter discusses in depth about the data acquisition system that renders multi-view face images for recognition. The acquisition system is called *collaborative multi-view face acquisition system*, since the cameras collaborate to acquire multi-view face images across the network. Each camera in the network can act as frontal or non-frontal. This relinquishes the restriction that a person should always face a particular camera.

3.1 Outline

For any recognition system to be robust and operative in real time, the data acquisition system and the underlying recognition algorithm play a substantial part. To put together such a system the acquisition system and the recognition algorithm (or scheme) employed should satisfy certain constraints. The data acquisition system should transmit only the relevant information and at a rapid rate. This, safeguards against the high network bandwidth and possibility of missing important events. The underlying recognition algorithm (or scheme) should be fast and accurate enough to process the data acquired by the acquisition system. We realized a *collaborative multi-view face acquisition system* that is capable of acquiring multi-view face images across the network reliably and at a rapid rate. The

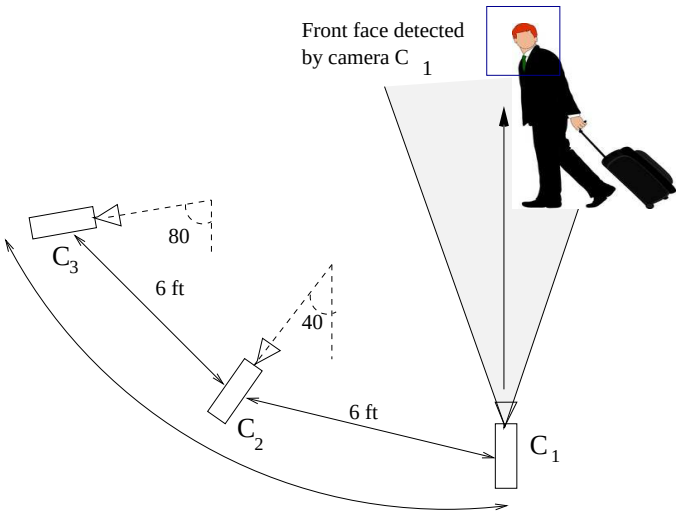

Figure 3.1: Our experimental deployment of 3 cameras. The cameras are deployed along an arc of radius 10 feet with a separation of 6 feet between the cameras along the arc as shown. The angles made by the principal axes of cameras C_2 and C_3 with that of camera C_1 are 40^o and 80^o respectively. The cameras are deployed on tripods at a height of 7 feet from the ground. All cameras run a frontal face detector. When a frontal face is detected on any camera, a notification is broadcast to other cameras.

system comprises of 3 cameras which are oriented in such a way so as to capture multi-view (frontal, partial profile (40^o), partial profile (80^o)) face images of the person. These face images are labeled and transmitted to the centralized location (or base station) which runs the recognition algorithm.

3.2 Acquisition system design

The collaborative face acquisition system entails a network of 3 cameras with overlapping field of views. These cameras are positioned to have the area of interest lie within the common region of the FOV's of all the 3 cameras. Our experimental setup consists of 3 cameras which are placed along the arc of radius 10 feet. The cameras are separated by a distance of 6 feet and are fixed on tripods at a height of 7 feet from the ground. The principal axes of the cameras is parallel to the horizontal plane as shown in the Fig. 3.1. The angle made by the cameras C_2 and C_3 with that of camera C_1 are 40^o and 80^o respectively. The cameras are connected wirelessly and have the same face acquisition software running on them.

Time synchronization between the nodes is established using NTP (Network Time Protocol). NTP is a protocol designed to synchronize the clocks of computers over a network. NTP is organized as an hierarchical client-server model. One of the node in the network is synchronized with the top level time servers available to the internet which in turn serves as the reference to the other nodes in the network. The clock on the other 2 nodes is synchronized with that of on the reference node. We note that clocks of any two nodes may not be in perfect synchronization at any time instant. Let t_s denote the maximum clock synchronization error between any pair of cameras in milliseconds.

Each camera in the network can act as frontal or non-frontal depending on which camera the person is facing. If the person is directly facing the camera C_1, then C_1 acts as frontal and C_2, C_3 act as non-frontal. Similarly, if the person is facing C_2, then C_1, C_3 act as non-frontal. If the person is facing C_3, then C_1, C_2 act as non-frontal. This behavior lets go the restriction of person having to face a specific camera. The multi-view faces collected across the network are transmitted to the base station for recognition.

Our experiment is carried out with the human subject facing the camera C_1. The cameras C_2 and C_3 act as non-frontal cameras for the subject. As a result, the camera C_1 captures the frontal-view, C_2 acquires the partial left (or right) profile and C_3 acquires the left (or right) profile of the subject. We use the yaw angle (that measures the rotation of the face image along the vertical axis) to define front, partial profile and profile faces(Fig. 3.2). If the yaw angle made by the subjects face image ranges from -30^o to $+30^o$ we define it as

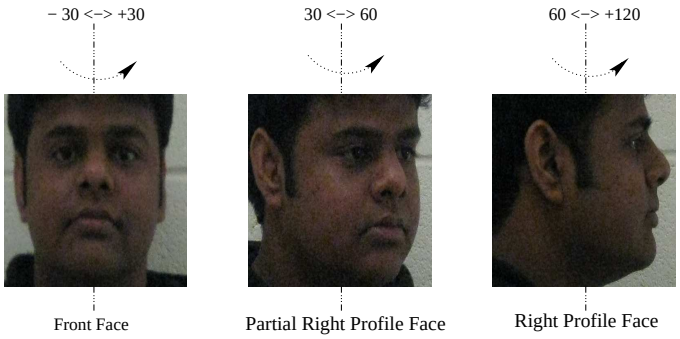

Figure 3.2: We classify faces into front, profile and partial profile based on the yaw angles

frontal. If it ranges from -30^o to -60^o (30^o to 60^o) we define it as partial left (or right)

profile face. If it ranges from -60^o to -120^o (60^o to 120^o) we define it as left (or right) profile face. The designed multi-view face acquisition system exploits camera geometry to acquire multi-view face images reliably and at a rapid rate.

3.3 System Operation

Our face acquisition system encompasses 3 cameras oriented to acquire frontal and non-frontal face images. All the cameras have the same software running on them. The tasks performed by a camera in the network can be categorized into 4 threads.

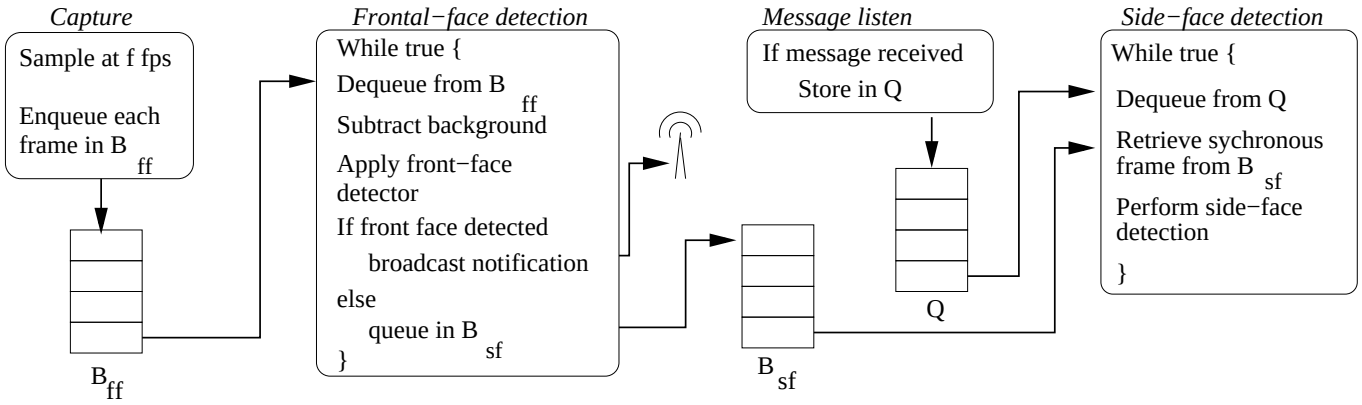

Figure 3.3: Pseudo-code for operations on each embedded camera. each node executes 4 threads: capture, frontal face detection, message listening and side-face detection. The capture thread samples images at F fps and queues them in B_{ff}. The frontal face detection thread dequeues frames from B_{ff} and applies frontal face detector on background subtracted images. If a face is detected, a notification is broadcast to other cameras, otherwise the background subtracted frame is stored in B_{sf}. The message listening thread queues any incoming message into Q. The side-face detection thread dequeues messages from Q, retrieves the synchronous frame corresponding to the message from B_{sf} and performs the side-face detection procedure.

3.3.1 Capture

The capture thread acquires images of the scene at F fps and queues them in the buffer B_{ff}. The timestamp of frame is defined as the time of capture of frame x and is denoted as

$t(x)$. Two frames are said to be synchronized if they have the same timestamp. Considering the fact that maximum clock synchronization t_s is around it is likely that two frames captured at the same time may not have the same time. In which case, we consider the frame having the closest timestamp. Let $|B_{ff}|$ denote the maximum number of frames in the buffer B_{ff}.

3.3.2 Frontal face detection

The frontal face detection thread dequeues the oldest frame the buffer B_{ff}. This frame, after background subtraction, is subjected to frontal face detector. We use the OpenCV implementation of the Haar Cascade based face detector [7]. If a frontal face is detected a notification message $M(c(x), t(x), w(x))$ is broadcast to all the other cameras in the network, where $t(x)$ is the timestamp of the frame x, $c(x)$ is the location of the center of the detected face and $w(x)$ is the width of the bounded square around the face detected. If a frontal face is not detected the frame is stored in the side face buffer B_{sf}. Let $|B_{sf}|$ denote the maximum number of frames stored in the side face buffer.

3.3.3 Message listening

The message listening thread listens for the notification messages $M(c(x), t(x), w(x))$ from the neighboring cameras. These messages are queued in the buffer Q. Let $|Q|$ denote the maximum number of messages in the buffer Q.

3.3.4 Side face detection

The side face detection thread dequeues a message from buffer Q one at a time. If the retrieved message is $M(c(x), t(x), w(x))$ then the corresponding frame y from buffer B_{sf} is dequeued such that $t(x) = t(y)$. Utilizing the concept of epipolar geometry and known information $w(x)$ and $c(x)$ the search space in the frame y is reduced to a square block of size $w \times w$ pixels. Based on the relative camera orientations, we determine the expected pose of a side face and apply side-face detector corresponding to the particular class on the extracted square block.

3.4 Epipolar geometry

Epipolar geometry describes the projective geometry between two views. Epipolar geometry reduces corresponding point search space from 2D image to 1D epipolar line since point x in one camera is constrained to lie on an epipolar line l' in the other image. Fundamental matrix, purely dependent on the internal parameters of the camera, is used to compute projective mapping between uncalibrated views and it is an algebraic representation of an epipolar geometry [49].

Properties of Fundamental matrix (F),

- Fundamental matrix is of rank 2 and has seven degrees of freedom.
- Point correspondence : If x and x' are two corresponding image points, then

$$x'^T F x = 0 \tag{3.1}$$

- Epipolar lines :

$$l' = Fx \tag{3.2}$$

  is the epipolar line corresponding to x.

$$l = F^T x' \tag{3.3}$$

  is the epipolar line corresponding to x'.

- Epipoles :

$$Fe = 0 \tag{3.4}$$

$$F^T e = 0 \tag{3.5}$$

Fundamental matrix computation:

Fundamental matrix is a 3×3 matrix of rank 2 and its computation is based on corresponding image points between images and independent of camera calibration and camera internal parameters. Several techniques have been proposed to compute the F, but normalized-8 point algorithm has shown superior performance since input data normalized before solving linear equations.

For our experimental setting, fundamental matrix computed between a pair of cameras be F_{12}. Using this fundamental matrix project the point $C(x)$ (center of the frontal face detected) in frame x onto a line (epilpolar line) in frame y. We then determine the intersection of line with the background subtracted image retrieved from B_{sf} and extract a square block of size $w \times w$ pixels.

3.5 Experiment setup

We implement our data acquisition system as a 3 node embedded camera network (schematics shown in Fig. 3.1). We assemble an embedded camera using a Logitech 9000 camera, a 1.6 GHz Intel Atom 230 processor based motherboard from Acer [Acer] and an IEEE 802.11 based wireless card. We consider one human subject in the scene at a time. Each subject stands at a distance of approximately 10 feet from the cameras (close to the center of the arc) facing any one of the 3 cameras. Note that, if the subject is facing camera C_1 as shown in Fig. 3.1, then the pose estimated by camera C_2 and C_3 are right partial profile and right profile respectively. We have tested the system with 10 different subjects with approximately 15 minutes of data collected for each subject.

(a) (b)

Figure 3.4: Example face images detected by our acquisition service. The white rectangles indicate the box enclosing the detected faces in each pose. Face images in each column are extracted from synchronous frames in the three cameras. (a) Images acquired with subjects facing C_2: (Top) Frontal face (Middle) Left partial profile face (Bottom) Right partial profile face. (b) Images acquired with subjects facing C_1: (Top) Frontal face (Middle) Right partial profile face (Bottom) Right profile face.

3.6 Results

We perform our experiments in two environments: one with a lot of clutter in the background and the other one with a relatively plain background. Images are sampled by each camera at 25 fps. Thus $t_f = 40$ms. In Table 3.1, we show the average execution times for the different processing modules in our system. In Table 3.2, we show the number of frames that are processed per second for detecting frontal faces and side faces. The frontal face detector is applied on background subtracted regions and sometimes applied even on spurious blobs detected as the foreground. The side-face detector on the other hand is applied only on a much smaller region that is corroborated by the frontal face detecting camera.

Operation	Time (ms) (clear)	Time (ms) (cluttered)
Image capture and storage	2	2
Background subtraction	2	3
Dilation	2	2
Frontal face detection	75	102
Total t_{ff}	81	109
Total t_{sf}	15	15

Table 3.1: Processing times: Multi-view face detection in clear and cluttered background

The actual number of frontal and side faces detected correspond to the output of the detector itself. The difference between frames processed and faces detected gives a measure of the false negatives for the respective detectors. In a clear background, the number of frontal faces detected per second are almost equal to the number of frames processed per second. All the frontal faces detected are notified to the other cameras and the number of side faces detected per second in each camera matches the frontal face detection rate. In a cluttered background, the number of missed detections for frontal faces are high and yields a frontal face detection rate of 6 faces per second and as seen in Table II, the side face detecting cameras are able to match this detection rate.

The maximum network delay is observed to be 50ms, but we note that this only affects the size of B_{sf} and not the overall face detection rate. We also note that the required buffering is very low (approximately 10 frames). By transmitting only the face images, that

Rates per second	Clear	Cluttered
Frontal face processed	11.1	8
Frontal face detected	10.2	6.05
Side-face processed	10	5.5
Side-face detected	9.7	5.2

Table 3.2: Detection rates for frontal and side faces

are on average 60×60 pixels in size, we are able to reduce communication bandwidth by 98% compared with transmitting the entire image (640×480 pixels) and by 80% when compared with transmitting the background subtracted image (100×200 pixels on average). By performing face detection and simultaneously estimating the pose, we also expect to reduce significant processing time at the fusion center for face recognition.

Chapter 4

Fusion

Fusion [42] can be defined as a process of integrating information from multiple sources to produce the most comprehensive unified data about an entity. The purpose of information fusion is to determine the optimum set of features (or scores) and devise a fitting function that can suitably combine the scores to arrive at a decision. Face recognition systems employ face modality for human identification. Each modality (or characteristic) on its own cannot always be reliably used to perform recognition. Hence fusion strategies are utilized to improve the overall accuracy of the recognition system.

Information fusion in biometric systems is studied in great detail and is continued to do so because of the wide range of applications biometric systems find itself in. Biometric systems using multiple biometric sources to perform recognition are termed as multibiometric systems. By fusing the data from multiple biometric sources, the multibiometric system is considered to outperform the traditional (or uni) biometric system which makes use of the single piece of evidence.

The face acquisition system simultaneously labels the pose of each acquired face image. In dynamic environments, the face images are continually arriving at the base station with different quality, pose and resolution. Fusing information obtained from multiple probe images poses significant challenges. Such a scenario demands that a verdict (match / no match) regarding a particular subject is quickly (yet accurately) released so that more number of subjects in the scene can be evaluated. The following questions then arise: in what order should probe images be matched ?, how to combine scores obtained from multi-

ple probe images ?, how soon can a verdict be confidently reached ?, what is the expected performance of such a fusion scheme ?.

Depending on what stage of the biometric system fusion is carried out, we have five fusion techniques : sensor, feature, score, rank and decision level fusion. Sensor module has the richest source of information and the amount of information is condensed as we move from sensor to decision module. Integrating match scores output from multiple biometric sources is termed as score level fusion. This is also known as measurement level or confidence level fusion. Because of its ease to access and consolidate, score level fusion strategy is widely used. Most existing work on biometric fusion [42, 43, 44] has assumed that data has been acquired a priori and prevailing fusion techniques operate in controlled scenarios. There remains a need for optimizing these techniques for operation in a dynamic mode where data is continually streaming in and each image varies in quality (ambient conditions, pose, resolution). Designing a fusion strategy to process this data and arrive at a decision in real-time poses significant challenges. It is not practical to process all the acquired data and we need to implement a strategy to fuse only the relavent information. At the same time, it is required that the verdict is made reasonably quickly and is reliable. To address these, we investigate the initial steps towards real-time human identification using a wireless camera network.

Multiple biometric sources can lead to multiple biometric traits or a single trait viewed in multiple ways. That is, we can have a biometric system equipped with a camera, fingerprint scanner and iris recording equipment. This system is capable of acquiring data from three different biometric traits namely face, fingerprint and iris. On the other hand, we can design a system having three cameras oriented in such a way as to acquire the frontal, partial profile and profile face image of the same subject. This system best describes the biometric system having the ability to acquire a single biometric trait (face in this case) in multiple ways.

4.1 Sources of multiple evidence

- **Multi-algorithm systems :** The systems where the same biometric information is processed using multiple algorithms are known as multi-algorithm systems. For instance, a frontal face image of a person can be processed using Principal Component Analysis (PCA), Local Binary Patterns (LBP) and Linear Discriminant Analysis (LDA). Later, the outcome of the individual classifier is fused to arrive at a unified decision. This technique is proven to increase the recognition rate.

- **Multi-sensor systems :** In these systems, the same biometric trait is pictured using multiple sensors. This system somewhat guarantees to have acquired diverse information of the biometric trait being imaged. For example, Marcialis and Roli, 2004a presented a strategy which would combine the fingerprint information obtained using an optical and a capacitive fingerprint sensor.

- **Multi-sample systems :** These systems acquire multiple samples of the same biometric trait. As a result, these systems account for the variations in the underlying biometric trait and are robust to slight discrepancies in the biometric information. A three camera system acquiring the frontal, partial profile and profile face image of a person is an example of such a system.

- **Multi-modal systems :** The systems that fuse the information from several biometric traits are known as multi-modal systems. The best example is a biometric system comprising of a fingerprint sensor and a voice analyzer.

- **Hybrid systems :** A combination of subset of any of the previous systems would result in an hybrid system. A multi-sample and multi-modal system can be integrated giving rise to hybrid system. These systems although improve the recognition accuracy suffers from the drawback of being costly.

Our fusion technique is an hybrid system. It is a combination of multi-sample and multi-algorithm systems. As an example, our system uses multiple algorithms, namely LBP based classifier and PittPatt commercial software and multiple samples of the same biometric

trait face, namley frontal, partial profile and profile face. PittPatt is used for frontal face recogntion and LBP based classifier is used for non-frontal face recogntion.

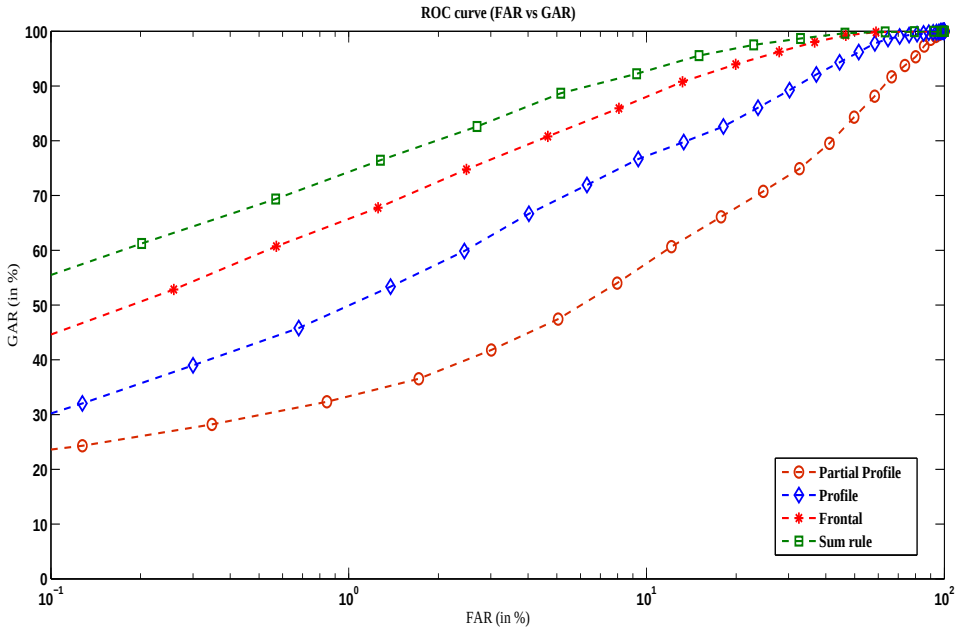

Figure 4.1: ROC curve (GAR vs FAR) based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

4.1.1 Illustrative example

The camera network system acquires multi-view face images, namely frontal, partial profile and profile faces. The experimental setup is discussed in 5.3. PittPatt software is used for frontal face recognition and LBP is used for partial profile and profile face recognition. Each recognition algorithm outputs a matching score or confidence. The database for our experiment consists of matching scores for frontal, partial and profile face images for 25 subjects. We have three score matrices, one for each view. ROC curve (illustrated in Fig. 4.1) GAR (Genuine Acceptance Rate) vs FAR (False Acceptance Rate) is plotted across each of the view using the score matrices. As expected, the frontal face results in higher accuracy than non-frontal face images because the the number of features offered in frontal view are higher than non-frontal view.

S_f, S_{pp} and S_p denote the matching score for frontal, partial profile and profile face

images. The scores across each view are integrated using a weighted sum rule resulting in a single score or confidence. The resulting score matrix is used to plot the ROC curve. We notice that by fusing the information from multiple sources the recognition accuracy increases. Fig. 4.2 compares GAR for fusion and frontal only at varying FAR. Fig. 4.3

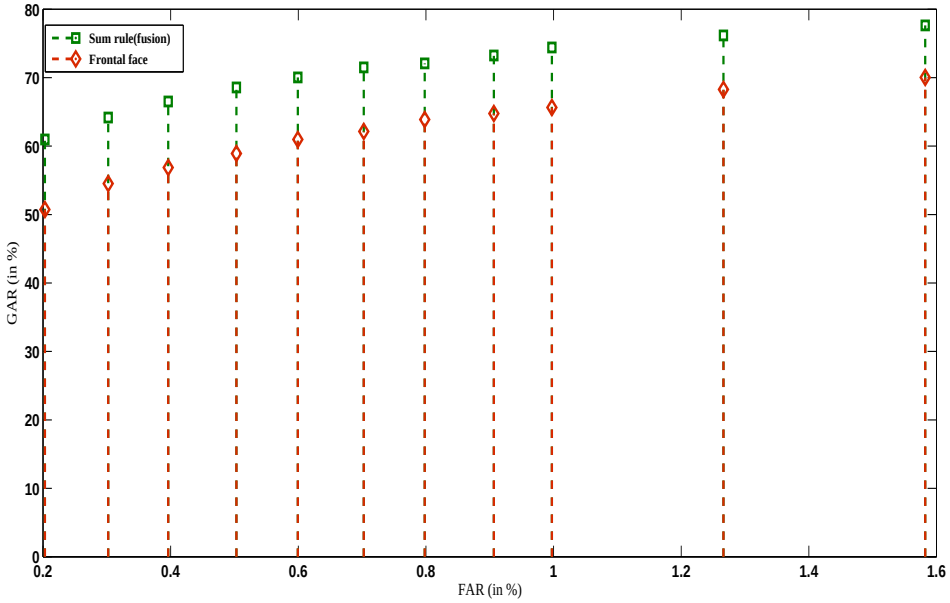

Figure 4.2: GAR vs FAR with:frontal only and fusion.

shows the schematic of a multi-view face recognition system having 3 cameras. The score output from each of the matcher is normalized and is subjected to a fusion rule (in this case, weighed sum rule).

Information fusion is categorized based on what stage of the multibiometric system fusion strategy is implemented. If the fusion is carried out on the raw data from the sensors it is referred as sensor level fusion. Combining features extracted from multiple biometric sources is called feature level fusion. Fusing the classifier output or score is termed as match score level or measurement level or confidence level fusion. Decision level fusion refers to integrating the decisions output by the biometric system independently. Of all the different levels of fusion discussed above, feature level fusion has the richest source of information followed by match score level fusion. Match score level fusion technique, where the match

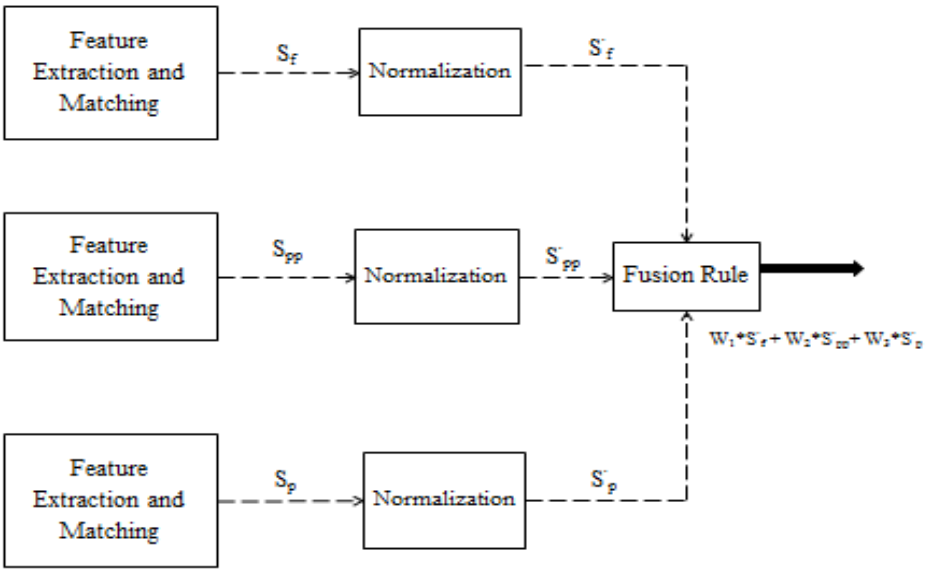

Figure 4.3: Score level fusion of frontal, partial profile and profile face images.

score outputs from the biometric system are integrated, is easy to implement and widely used method.

Since we have multiple samples (frontal, left partial profile (-40^o), left profile face (-80^o) face) of the same biometric trait (face biometric in this case) we use a multi-sample fusion system. These images are matched to the images in the database having the same pose to get the scores. The 3 scores are fused using a sum-level fusion strategy to get a unified score(illustrated in Fig. 4.3).

Chapter 5

Face Recognition System

Our face matching module consists of two components: front-face matching and non-frontal face-matching. In an ideal environment, the frontal-face matching alone would have been enough to get the desired results. This can be possible when the system is operated in controlled environments, no timing bounds and the captured images are of good quality. When the system is to be operated in unconstrained environments and we need to perform recognition in real-time the frontal-face matching alone may not suffice the purpose. This is when non-frontal faces come in handy. The additional information they offer will aid the frontal-face recognition, thus enhancing the overall accuracy. To reap the benefits of the available information (frontal and non-frontal scores) we need to have an information fusion strategy in place.

5.1 Setup

We use a network of 3 Firewire cameras located along an arc of radius 10 feet. The cameras are deployed on tripods at a height of about 7 feet from the ground. The angles made by the principal axes of the cameras C_2 and C_3 with that of camera C_1 are 40^o and 80^o respectively. The cameras are connected wirelessly. The multi-view face detection software (described earlier) is run on these cameras. As a result, multi-view face images (0^o, 40^0, 80^0) indexed by pose, are collected at the fusion (recognition) center. These images continuously

stream in, when a subject is present in the FOV of the camera network. The face images are time-stamped based on the reception time at the fusion center.

5.2 Method

For frontal face recognition, we use the PittPatt Face Recognition Software Development Kit from Pittsburgh Pattern Recognition [13]. This SDK provides recognition tools that extract templates from faces and compare templates to compute similarity scores. When using this software, there is no need to explicitly align the frontal images before feeding to the recognition algorithm. The underlying algorithm is also robust to slight variations in the pose and illumination changes. The PittPatt face recognition algorithm extracts the template of the test image and compares it with the templates of the face images in the database.

The PittPat software is designed to work with frontal images only. It supports variations in the yaw angle ranging from -20^o to $+20^o$. The underlying recognition algorithm outputs a similarity score that is subsequently normalized to a value in the range $[0, 1]$ using the min-max normalization technique, i.e., given matching scores s_k for $k = 1, ..n$, the normalized scores are:

$$s_{norm} = \frac{s - min\{s_k\}}{max\{s_k\} - min\{s_k\}} \tag{5.1}$$

Face recognition algorithm outputs a score which can be similarity or dissimilarity score. Similarity score is defined as the measure of how well the two images are alike. Dissimilarity score is defined as the measure of how different the two images are. In our case, PittPat outputs a similarity score and LBP outputs a dissimilarity score. For fusion, all the scores should either be similarity or dissimilarity scores.

For non-frontal face recognition, we use the Local Binary Patterns based classifier which considers both shape and texture information to represent the face images [50]. The LBP operator forms labels for the image pixels by thresholding a p pixel neighborhood around each pixel in comparison with the center pixel of the neighborhood, and considering the result as a binary number. This results in a p bit label for each neighborhood, with $2p$ possible values. A histogram of these $2p$ labels is then used as the image descriptor. Since its

introduction, the LBP operator has been extended to use different neighborhood sizes and shapes. In our system, we have specifically considered a circular $(8, 3)$ neighborhood, i.e., 8 sample points uniformly separated along a circle of radius 3 around each pixel. Furthermore, we use an extension of LBP, namely uniform LBP, in which only a subset of the $2p$ labels are used in forming the histogram feature. Specifically, only patterns in which there are at most 2 bitwise transitions from 0 to 1 or vice-versa are considered as uniform. A separate label is used for each of these uniform patterns, and one label is used for all the other patterns. In an $(8, 3)$ neighborhood this results in 58 uniform patterns, i.e. 59 labels. In forming the LBP feature vector, each face image is first divided into 5×5 equi-sized smaller sub-blocks (or cells). Division of a face image into smaller cells allows us to retain spatial information in the face image. Local 8 bit binary patterns are extracted and a separate histogram is obtained for each cell. Let R_j denote the jth cell where $1 < j < 25$. Let $f(x, y)$ denote the label of pixel (x, y), where $f(x, y)$ ranges from 0 to 59. Let $H_{i,j}$ denote the histogram frequency of the label i $(0 < i < 59)$ in region R_j . Let $I(A) = 1$ if predicate A is true and 0 otherwise. Thus we have:

$$H_{i,j} = \sum_{x,y} I[f(x, y) == i] * I[(x, y)\epsilon R_j] \tag{5.2}$$

The histogram frequencies H_{ij} for cell j are normalized as:

Figure 5.1: A face image divided into 5×5 windows.

$$H_{i,j\,norm} = \frac{H_{i,j}}{sum(H_{i,j)}} \tag{5.3}$$

The histogram frequencies across all cells are then concatenated into a single histogram that efficiently represents the face image. During matching, scores are calculated using the nearest neighbor classification technique with Chi-square (χ^2) statistical estimate for dissimilarity measure. Specifically, if O_i denotes the observed frequency of label i, M_i denotes the expected frequency of label i, the number of labels are denoted by L, then the χ^2 dissimilarity measure between the observed sequence S and the expected sequence M is given by:

$$\chi^2(S, M) = \sum_i \frac{O_i - M_i}{O_1 + M_1}, (i = 0, ..., L - 1) \tag{5.4}$$

The LBP scores are then normalized so that they are homogeneous and their range lies within $[0, 1]$ using the min-max normalization technique. Note however that the PittPatt front face matcher assigns a similarity score. In order to be consistent with the front face scores, the dissimilarity scores $(d_k)k = 1, ..n$ are converted to similarity scores as,

$$s_k = 1 - d_k, (k = 1, .., n) \tag{5.5}$$

5.3 Experiment

With the above face matching techniques in place, an experiment was carried out in a cluttered office background with 30 human subjects using a 3 camera network as shown in Fig. 3.1. The multi-view face acquisition software was installed on these cameras. The subjects walked facing one of the 3 cameras in the system at a speed of 2 to 3 feet per second and stayed within the FOV of the network for approximately 6 seconds in each trial. As each subject walked through the camera network system, approximately 50 to 60 images were acquired for each subject from each pose using our multi-view face acquisition framework, and these face images were simultaneously labeled into the appropriate pose. Probe images for each subject were indexed based on the timestamp associated with that image. Thus front, partial profile and profile images indexed by the same timestamp correspond to synchronous frames. To be able to quantify the recognition accuracy of the system at different image resolutions, we grouped the acquired subject probe images from each camera into the following resolution sets, low, medium and high, as indicated in **Table 5.1**.

The front, partial profile and profile images in each resolution set were compared against the

Table 5.1: Resolution sets for acquired face images

Distance from camera	Resolution	Average size of face images
8-10 feet	High	70x70
10-12 feet	Medium	55x55
12-14 feet	Low	48x48

respective gallery images for each subject using the PittPatt and LBP technqiues as described above. The probe images are only matched with gallery images of the corresponding pose. Let s_f, s_{pp} and s_p denote the matching scores for front, partial profile and profile face images of a subject respectively that are synchronous and indexed by the same timestamp. A score-based weighted linear fusion rule was applied to generate an overall score (s_o) using all types of face images corresponding to that timestamp,

$$s_o = w_f s_f + w_{pp} s_{pp} + w_p s_p \tag{5.6}$$

5.4 Results

For images in each resolution set, a graph of the False Acceptance Rate against the Genuine Acceptance Rate is obtained with only front face scores, only partial profile scores, only profile scores and with fused scores. Note that the threshold depending fraction of the falsely accepted images divided by the number of all impostor images is called False Acceptance Rate (FAR). The fraction of the number of rejected images divided by the total number of images is called False Rejection Rate (FRR) and $1 - FRR$ is called the Genuine Acceptance Rate (GAR). Recognition accuracy when $FAR = FRR$ is called the Equal Error Rate (EER) [51, 52]. In obtaining these graphs, the score-based fusion weights were determined using an iterative procedure: for different combinations of w_f , w_{pp} and w_p, the EER is determined for the fusion based classifier and the combination of weights that gives the highest EER is selected. The results are classified into 3 categories.

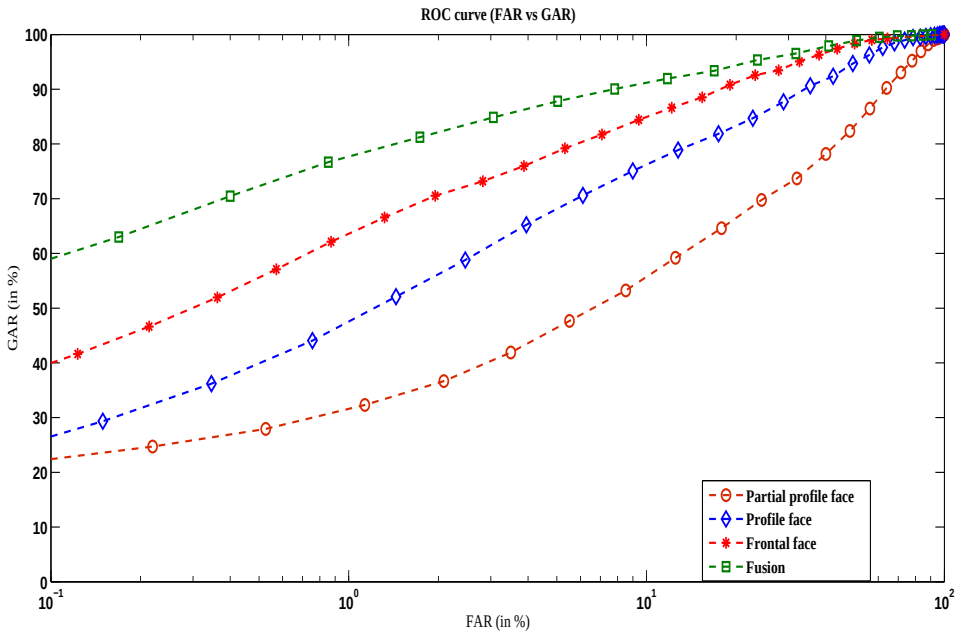

Figure 5.2: ROC curve (GAR vs FAR) for authentication based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

5.4.1 Multi-view fusion by treating each image independently

Each probe image is compared with all the gallery images (images in the database) and a similarity (or dissimilarity) score is generated for each comparison. For example, if we have n subjects and each subject has p probe and g gallery images, then it results in an $(n \times p \times g) \times (n \times p \times g)$ similarity matrix. The scores accross multiple views are combined using a weighted sum rule (5.6) and the fused scores are used to plot the ROC and CMC curve. Fig. 5.2 shows the ROC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.3 shows the ROC curve for medium resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.4 shows the ROC curve for high resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.6 shows the CMC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.7 shows the CMC curve for high resolution multi-view face (frontal, partial profile and profile) images.

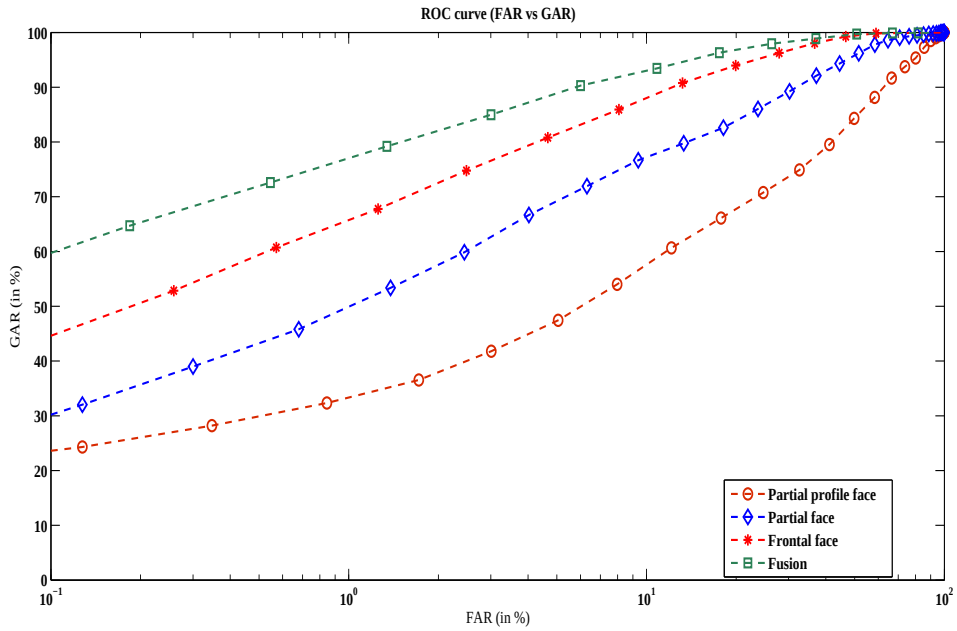

Figure 5.3: ROC curve (GAR vs FAR) for authentication based on medium resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

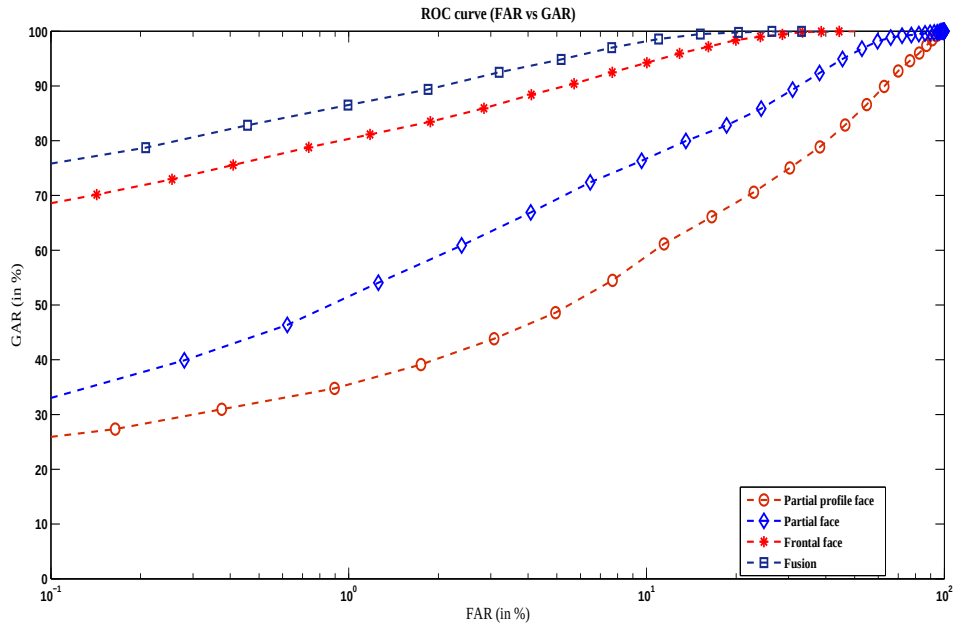

Figure 5.4: ROC curve (GAR vs FAR) for authentication based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

Fig. 5.5 shows the ROC curve when a score level fusion technique is applied to multi-view images obtained across the network with different image resolutions. The graphs clearly indicate that the impact of multi-view fusion is far more significant at lower resolutions.

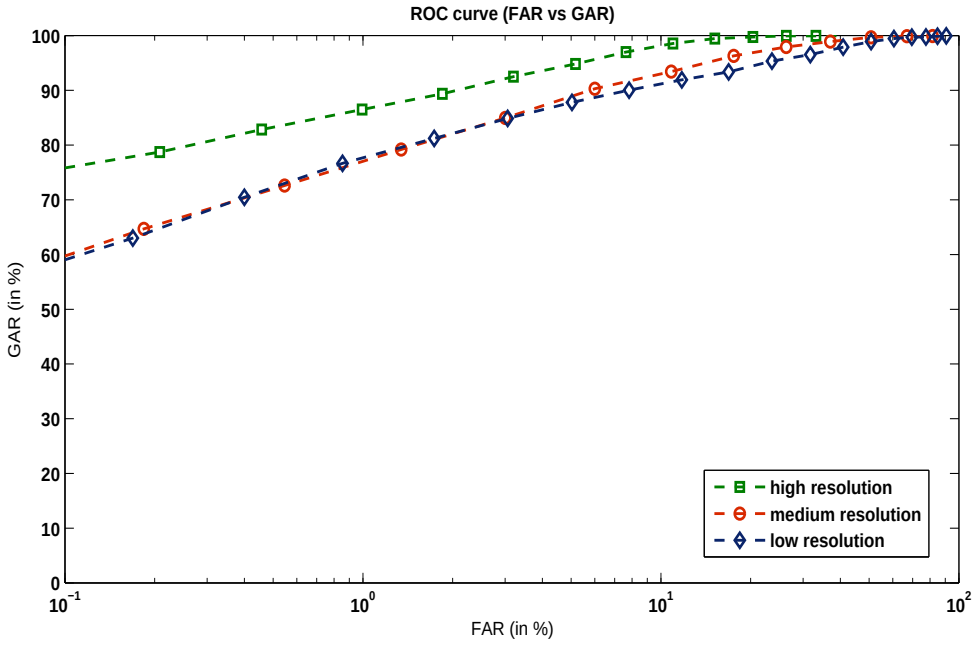

Figure 5.5: ROC curve (GAR vs FAR) by fusing multi-view images at low, medium and high resolution.

5.4.2 Multi-view, multi-sample fusion

Each probe image is compared with all the gallery images (images in the database) and a similarity (or dissimilarity) score is generated for each comparison. For example, if we have n subjects and each subject has p probe and g gallery images, then it results in an $(n \times p \times g) \times (n \times p \times g)$ similarity matrix. For each subject, the probe image resulting in the minimum score (we consider dissimilarity score) is retained. These scores accross multiple views are combined using a weighted sum rule (5.6) and the fused scores are used to plot the ROC curve. Fig. 5.8 shows the ROC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.9 shows the ROC curve for medium resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.10 shows the ROC curve for high resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.11 shows the

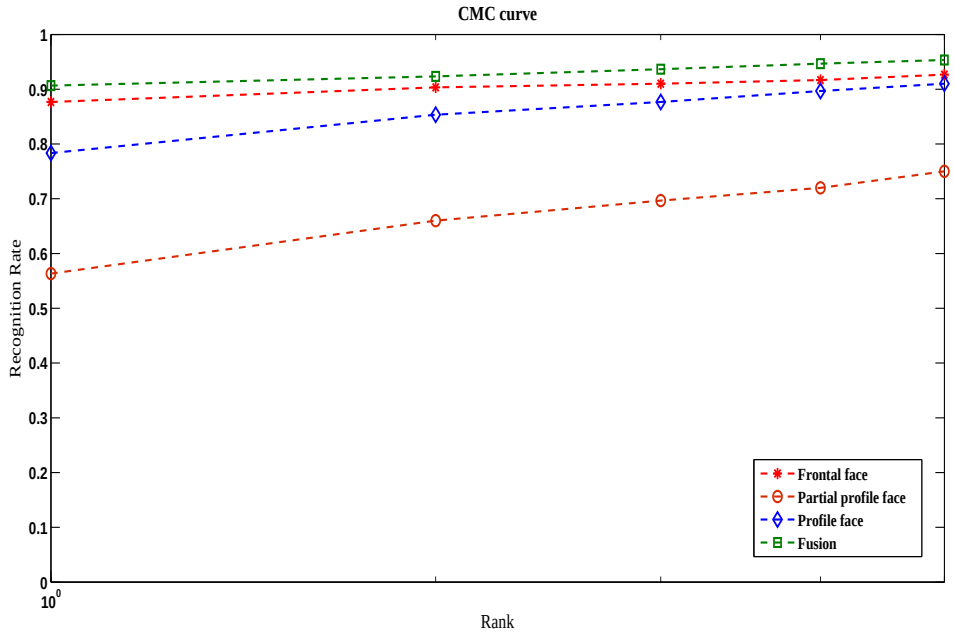

Figure 5.6: CMC curve (Recognition accuracy vs Rank) for identification based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

ROC curve when a score level fusion technique is applied to multi-view images obtained across the network with different image resolutions. Fig. 5.12 shows the CMC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.13 shows the CMC curve for high resolution multi-view face (frontal, partial profile and profile) images. The graphs clearly indicate that the impact of multi-view fusion is far more significant at lower resolutions.

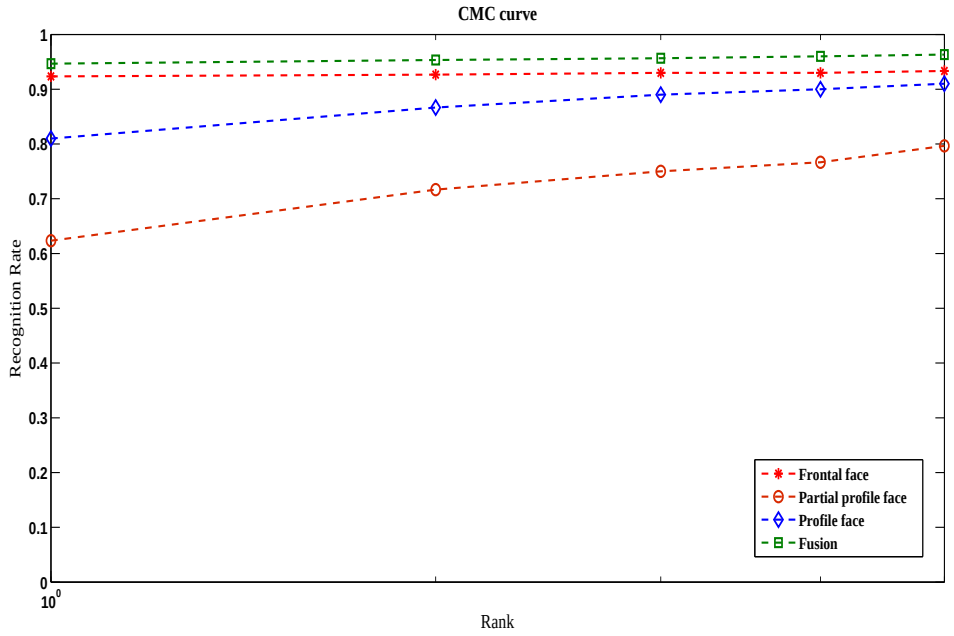

Figure 5.7: CMC curve (Recognition accuracy vs Rank) for identification based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

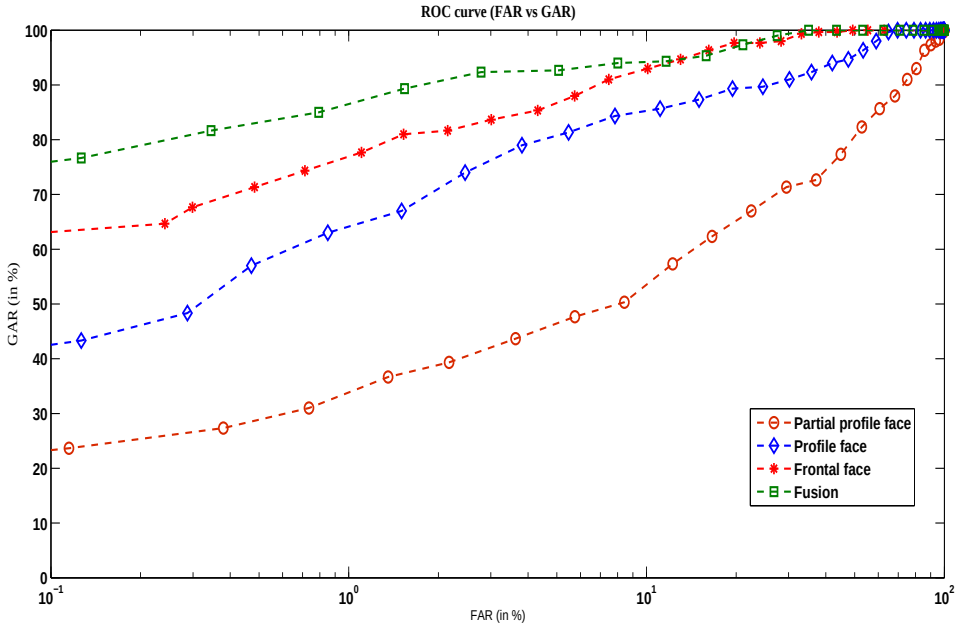

Figure 5.8: ROC curve (GAR vs FAR) for authentication based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

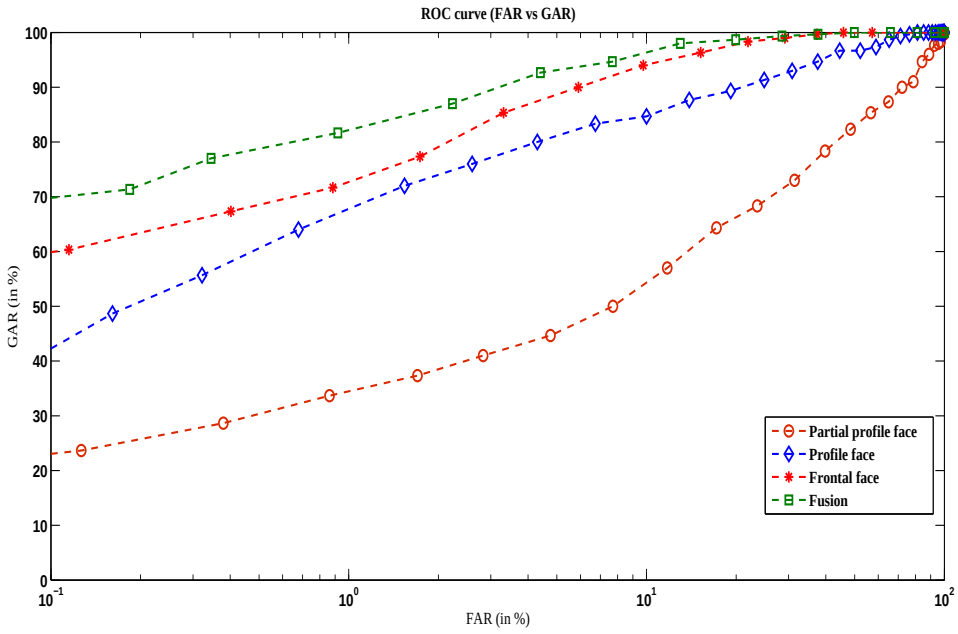

Figure 5.9: ROC curve (GAR vs FAR) for authentication based on medium resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

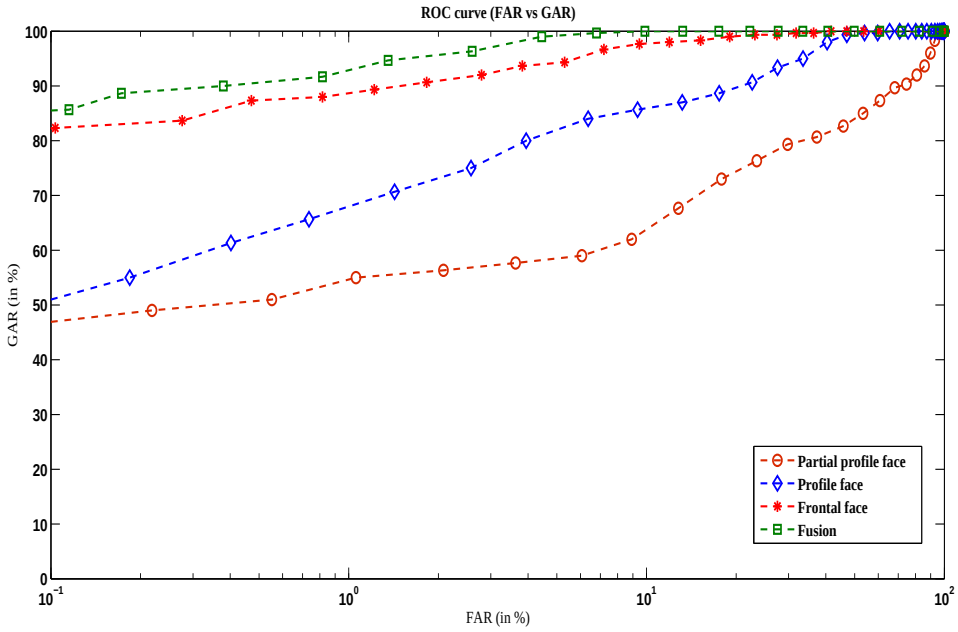

Figure 5.10: ROC curve (GAR vs FAR) for authentication based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

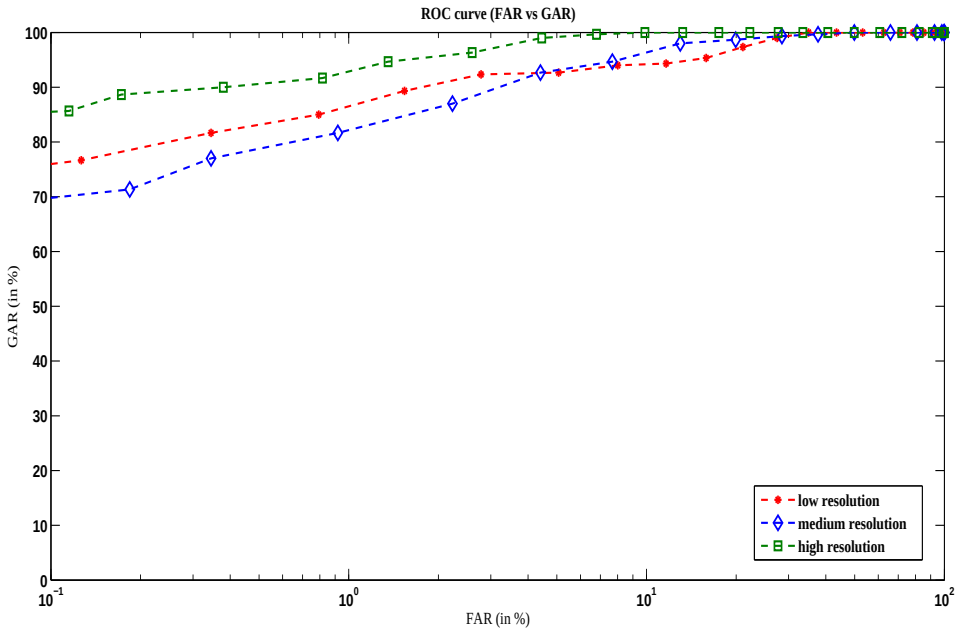

Figure 5.11: ROC curve (GAR vs FAR) by fusing multi-view images at low, medium and high resolution.

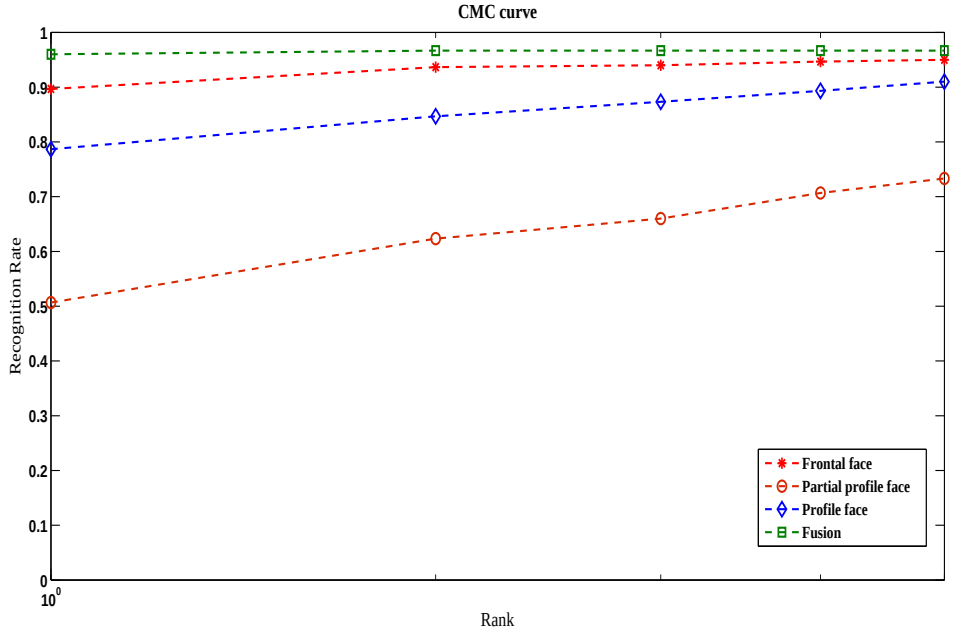

Figure 5.12: CMC curve (Recognition accuracy vs Rank) for identification based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

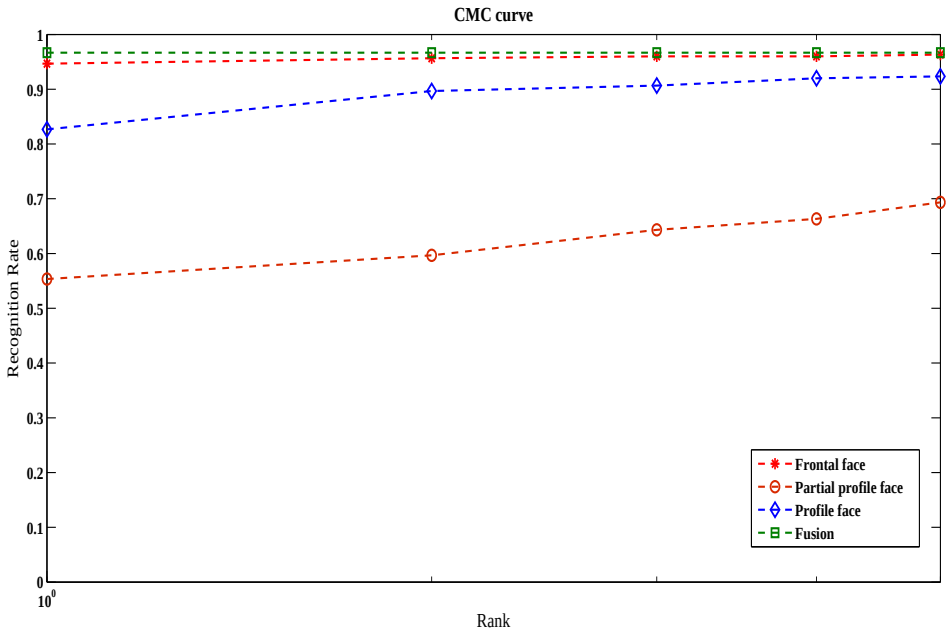

Figure 5.13: CMC curve (Recognition accuracy vs Rank) for identification based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

5.4.3 Robustness of weighted multi-view fusion

So far, the weights assigned to different views during fusion are determined by iterative procedure over all the subjects. In this experiment we determine the weights iteratively on a subset of the subject images and test it on the other images. This validates robustness of the weights determined. For each subject, the probe image resulting in the minimum score (we consider dissimilarity score) is retained. These scores accross multiple views are combined using a weighted sum rule (5.6) and the fused scores are used to plot the ROC curve. Fig. 5.14 shows the ROC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.15 shows the ROC curve for medium resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.16 shows the ROC curve for high resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.17 shows the ROC curve when a score level fusion technique is applied to multi-view images obtained across the network with different image resolutions. Fig. 5.18 shows the CMC curve for low resolution multi-view face (frontal, partial profile and profile) images. Fig. 5.19 shows the

CMC curve for high resolution multi-view face (frontal, partial profile and profile) images. The graphs clearly indicate that the impact of multi-view fusion is far more significant at lower resolutions.

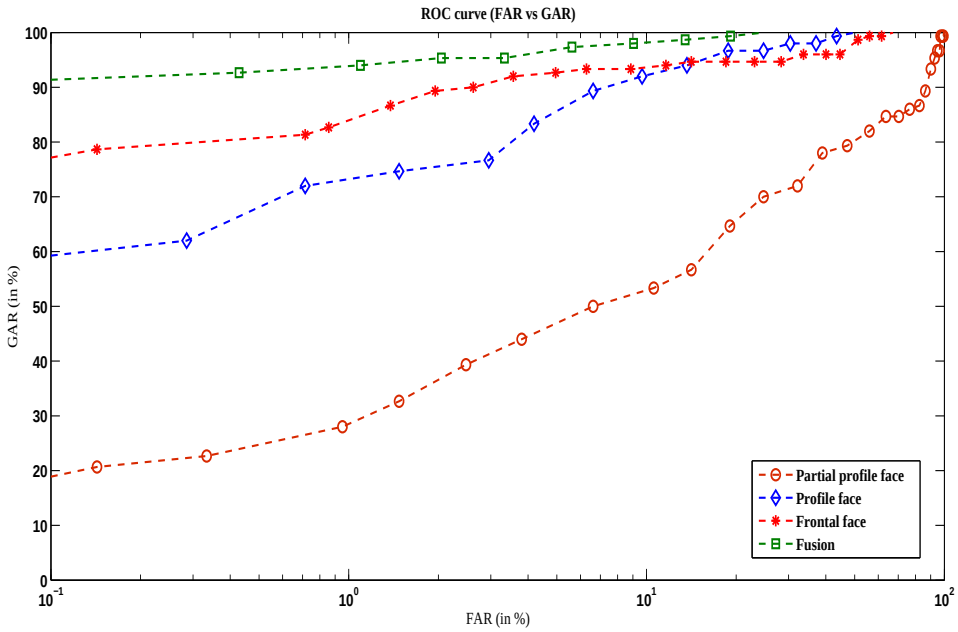

Figure 5.14: ROC curve (GAR vs FAR) for authentication based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

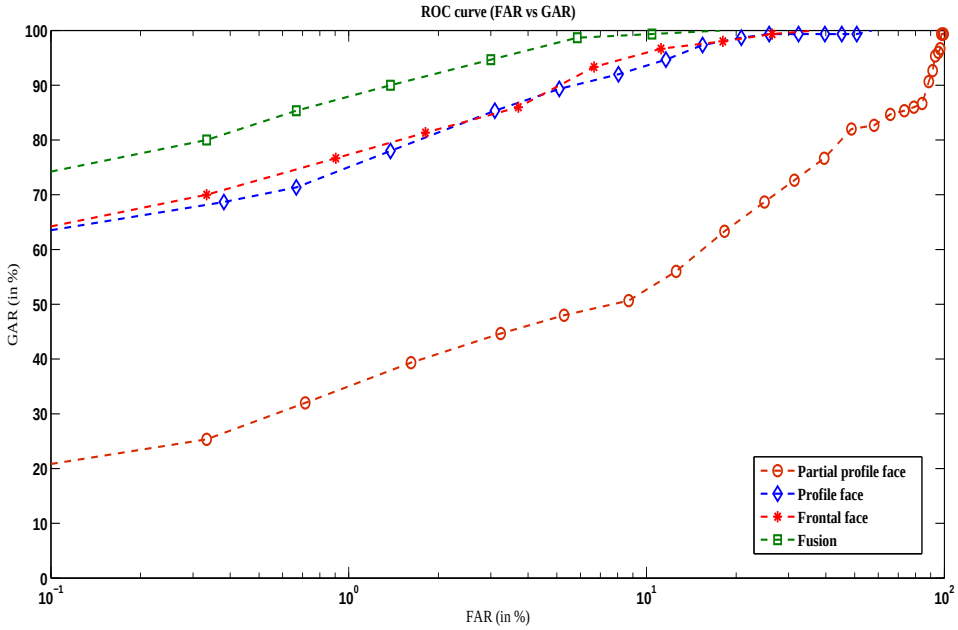

Figure 5.15: ROC curve (GAR vs FAR) for authentication based on medium resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

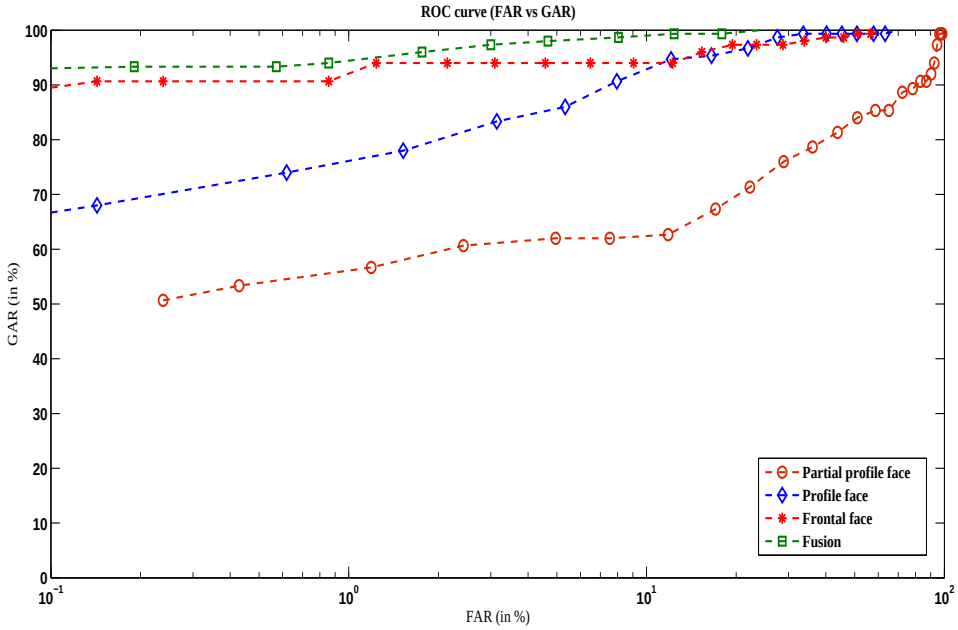

Figure 5.16: ROC curve (GAR vs FAR) for authentication based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

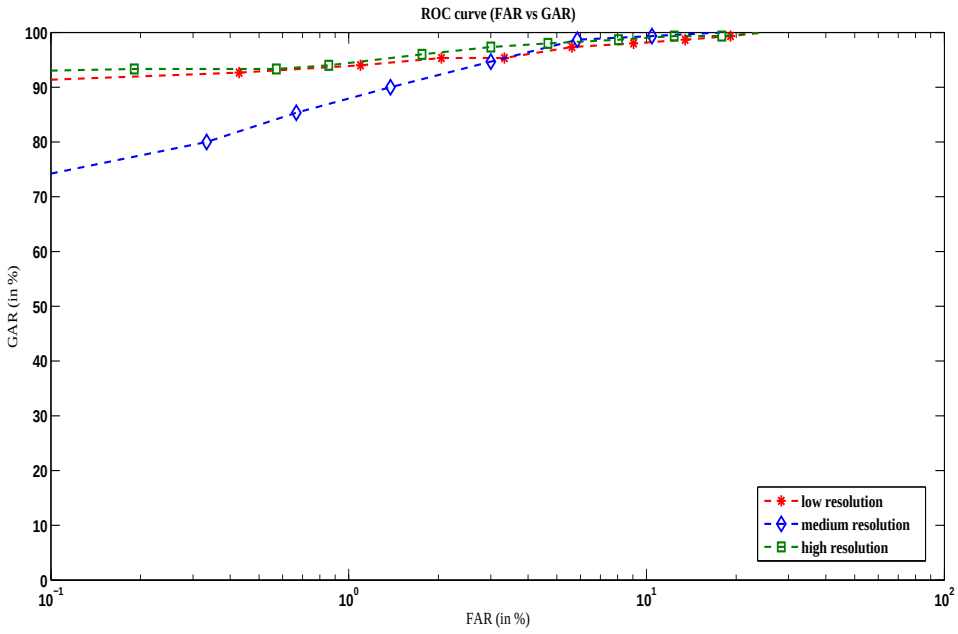

Figure 5.17: ROC curve (GAR vs FAR) by fusing multi-view images at low, medium and high resolution.

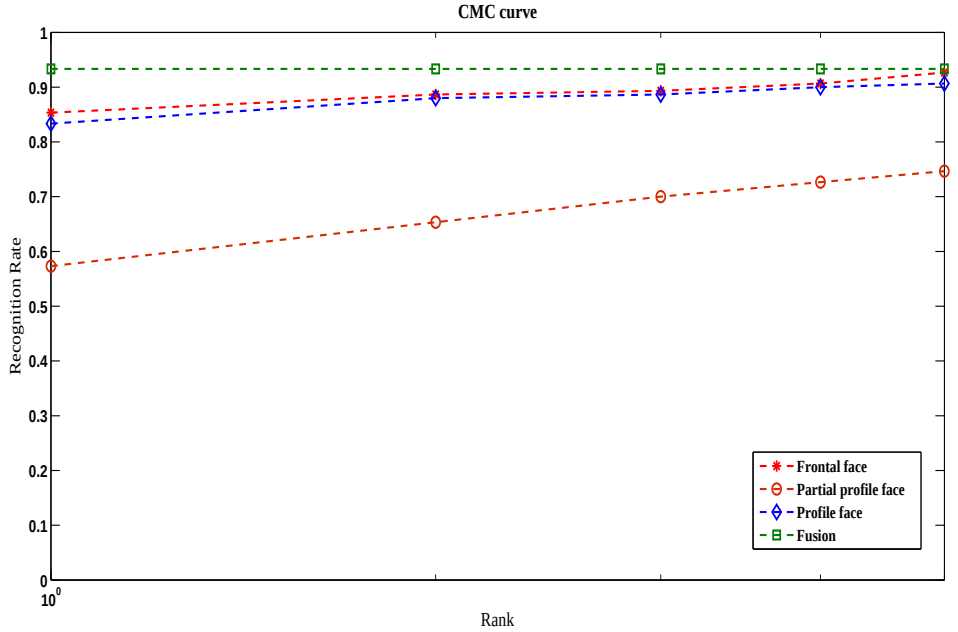

Figure 5.18: CMC curve (Recognition accuracy vs Rank) for identification based on low resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

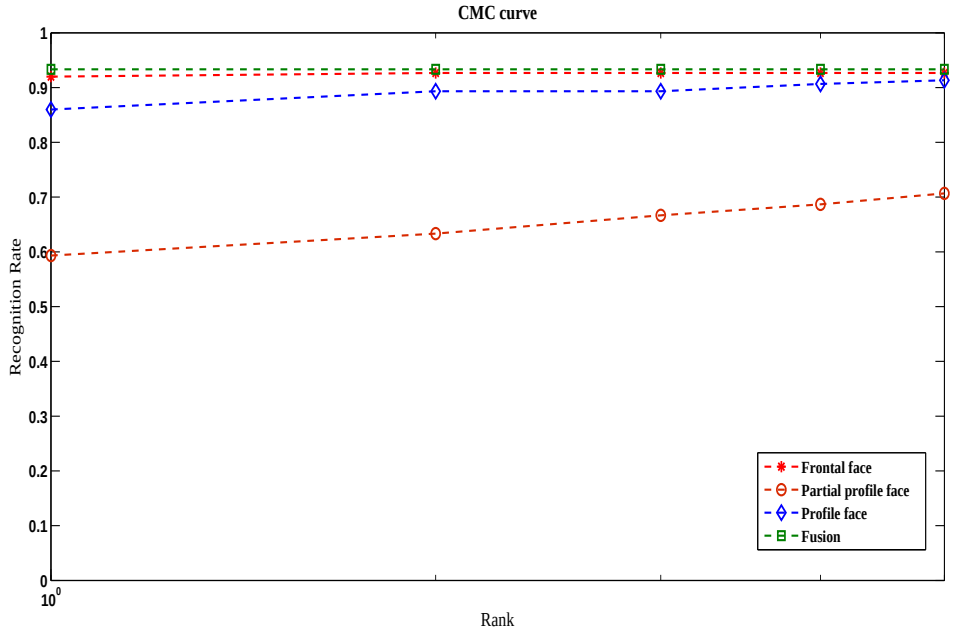

Figure 5.19: CMC curve (Recognition accuracy vs Rank) for identification based on high resolution images with: only front face, only partial profile (40^o), only profile (80^o) and multi-view face images.

Chapter 6

Conclusions and Future work

6.1 Conclusions

We presented a real-time face recognition system that is supported by a collaborative multi-view face acquisition service. Our service can detect and extract face images from different poses and simultaneously identify these poses while maintaining a high sampling rate. We avoid complex image processing and instead use multi-view camera geometry and inter-camera communication to reduce the processing time. We are able to achieve a non-frontal face detection rate that is almost equal to frontal face detection rate, thus highlighting the advantage over multi-view face detection schemes based on sequentially or hierarchically applying detectors for different poses. Our service is light-weight in terms of processing complexity, has low buffering requirements and is appropriate for implementation on different smart camera platforms [53] resulting in portable and even covert deployments for human recognition.

Our face image acquisition service was integrated with a multi-view face classification system using a combination of PittPatt SDK and LBP based classifiers. A score-based fusion technique was used for face recognition using a combination of front, partial profile and profile face images. Our results show significant improvement in recognition accuracy, especially when the front face images are of low resolution. By improving recognition accuracy at larger stand-off distances and lower image quality, we expect the face recognition system to be applicable for real-time watch-list identification scenarios in unconstrained en-

vironments.

We note that while we have used face detectors based on Haar-like features in our system, they could be replaced with other pose-specific face detectors as well. Also, while the specific performance numbers for processing rate are platform and algorithm specific, the key observation is that our system can be used to detect non-frontal faces at the same rate as frontal faces (where the rate of processing is determined by the algorithm and the platform). We have used the detection of frontal face images in a camera to guide the computation at run-time in other cameras. Alternatively, the detection of patterns or events other than frontal faces can also be used to trigger localized image processing operation in other cameras and improve the computational efficiency of the system. This gives rise to a more generalized use of our proposed framework for collaboration in a camera network.

In calculating the achievable recognition accuracy using a multi-view acquisition framework, we have used the individual probe image matching scores collected in each trial. However, it is possible to combine the matching scores obtained from multiple probe images of a given subject to improve the recognition accuracy.

6.2 Future work

Future work entails identification in a dynamic on- line mode, where data is continually streaming in and each image varies in quality (ambient conditions, pose, resolution). Such a scenario demands that a verdict (match / no match) regarding a particular subject is quickly (yet accurately) released so that more number of subjects in the scene can be evaluated. The following questions then arise: in what order should probe images be matched, how to combine scores obtained from multiple probe images, how soon can a verdict be confidently reached, and what is the expected performance of such a fusion scheme. Moreover, different features are likely to be better suited for classification of acquired images under different network parameters such as illumination, image resolution etc., and a systematic study will have to be completed to analyze the impact of image quality metrics on matching performance. Thus there is a need for adaptive face classification techniques as well as fusion algorithms that can intelligently combine the probe image inputs to determine a match while

simultaneously obtaining a confidence estimate for the match. Our future research will focus on these questions.

References

[1] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security (TIFS)*, 5(3):406–415, 2010.

[2] I. A. Kakadiaris, H. Abdelmunim, W. Yang, and T. Theoharis. Profile-based face recognition. In *8th IEEE international Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

[3] F.Yang, M.Paindavoine, H.Abdi, and D.Arnoult. Fast image mosaicing for panoramic face recognition. *Journal of Multimedia*, 1(2):14–20, 2006.

[4] K.W Cheung, J. Chen, and Y.S. Moon. Pose-tolerant non-frontal face recognition using EBGM. In *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2008.

[5] B. Bhanu and X. Zhou. Face recognition from face profile using dynamic time warping. In *International Conference on Pattern Recognition*, volume 4, pages 499–502, 2004.

[6] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):671 –686, April 2007.

[7] M. Jones and P. Viola. Fast multi-view face detection. Technical Report TR2003-96, Mitsubishi Electric Research Laboratories, 2003.

[8] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994.

[9] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(1):22–38, 1998.

[10] K. Sung and T. Poggio. Example-based learning of view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(1):39–51, 1998.

[11] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[12] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP*, 2002.

[13] Pittsburgh Pattern Recognition. PittPatt FTR Software Development Kit. http://www.pittpatt.com.

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

[15] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010.

[16] S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *European Conference on Computer Vision*, 2002.

[17] J. Feraud, O. Bernier, and M. Collobert. A fast and accurate face detector for indexation of face images. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[18] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[19] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399–458, December 2003.

[20] Athanasios Nikolaidis and Ioannis Pitas. Facial feature extraction and determination of pose. *Pattern Recognition*, 33:1783–1791, 1998.

[21] L. Torres, L. Lorente, and Josep Vila. Automatic face recognition of video sequences using self-eigenfaces. In *In International Symposium on Image/video Communication over Fixed and Mobile Networks, Rabat(Morocco*, 2000.

[22] Amit K. Roy-chowdhury and Yilei Xu. Pose and illumination invariant face recognition using video sequences, face biometrics for personal identification: Multi-sensory multi-modal systems, 2006.

[23] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, January 1991.

[24] W. Zhao, R. Chellappa, and P.J. Phillips. Subspace linear discriminant analysis for face recognition. Technical report, 1999.

[25] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[26] Chengjun Liu and Harry Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *International Conference on Audio and Video Based Biometric Person Authentication*, pages 22–24, 1999.

[27] Dakshina Ranjan Kisku, Hunny Mehrotra, Jamuna Kanta Sing, and Phalguni Gupta. Svm-based multiview face recognition by generalization of discriminant analysis. *CoRR*, abs/1001.4140, 2010.

[28] Zhi-Gang Fan and Bao-Liang Lu. Fast recognition of multi-view faces with feature selection. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 76–81, Washington, DC, USA, 2005. IEEE Computer Society.

[29] S. M. Khan, P. Yan, and M. Shah. A homographic framework for the fusion of multi-view silhouettes. In *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007.

[30] C. Wu, A. Khalili, and H. Aghajan. Multiview activity recognition in smart homes with spatio-temporal features. In *International Conference on Distributed Smart Cameras (ICDSC)*, 2010.

[31] N. Krahnstoever, Ser-Nam Lim T. Yu, K. Patwardhan, and P. Tu. Collaborative real-time control of active cameras in large scale surveillance systems. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[32] X. Zhou, R. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *ACM International Workshop on Video Surveillance*, 2003.

[33] V. Kulathumani, S. Parupati, A. Ross, and R. Jillela. Collaborative face recognition using a network of embedded cameras. In *Distributed Video Sensor Networks*, pages 373–389. Springer, 2011.

[34] S. Parupati, R. Bakkannagiri, S. Sankar, and V. Kulathumani. Collaborative acquisition of multi-view face images for real-time recognition using a wireless camera network. In *International Conference on Distributed Smart Cameras (ICDSC)*, 2011.

[35] Y. Yao, C. Chen, B. Abidi, D. Page, A. Koschan, and M. Abidi. Sensor planning for automated and persistent object tracking with multiple cameras. In *International conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[36] H. Jin and G. Qian. Robust multi-camera 3d people tracking with partial occlusion handling. In *International conference on Acoustics, Speech and Signal Processing*, 2007.

[37] Agent-vi. Video Analytic Systems. http://www.agent-vi.com.

[38] D. Chu, A. Deshpande, J. Hellerstein, and W. Hong. Approximate data collection in sensor networks using probabilistic models. In *IEEE International Conference on Data Engineering (ICDE)*, pages 3–7, 2006.

[39] R. Wagner, R. Baraniuk, S. Du, D.B. Johnson, and A. Cohen. An architecture for distributed wavelet analysis and processing in sensor networks. In *Information processing in sensor Networks (IPSN)*, 2006.

[40] J. gao, l. Guibas, N. Milosavljivec, and J. Hershberger. Sparse Data Aggregation in Sensor Networks. In *IPSN*, pages 430–439, 2007.

[41] S. Pattem, G. Shen, Y. Chen, B. Krishnamachari, and A. Ortega. Senzip: An architecture for distributed en-route compression in wireless sensor networks. In *Workshop on Sensor Networks for Earth and Space Science Applications (ESSA)*, 2009.

[42] Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24:2115–2125, 2003.

[43] U. Park, A. Jain, and A. Ross. Face recognition in video: Adaptive fusion of multiple matchers. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[44] A. Ross and R. Govindarajan. Feature level fusion using hand and face biometrics. In *Proc. of SPIE Conference on Biometric Technology for Human Identification II*, 2005.

[45] F. Wheeler and X. Liu and P. Tu. Multi-frame super-resolution for face recognition. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2007.

[46] F. Wheeler, X. Liu, P. Tu, and T. Hoctor. Multi-frame image restoration for face recognition. In *IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007.

[47] H. Wechsler. Linguistics and face recognition. *Journal of Vision, Language and Computing*, 20(3), 2009.

[48] Richa Singh, Mayank Vatsa, Arun Ross, and Afzel Noore. A mosaicing scheme for pose-invariant face recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(5):1212–1225, 2007.

[49] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[50] T. Ojala, M. Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 24(7):971–987, 2002.

[51] A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer Publishers, Berlin, Germany, first edition, 2006.

[52] A. K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[53] R. Kleihorst, B. Schueler, A. Danilin, and M. Heijligers. Smart camera mote with high performance vision system. In *ACM SenSys Workshop on Distributed Smart Cameras*, 2006.