



133
246
THS



This is to certify that the

thesis entitled

DETECTION AND RECOGNITION OF FACES IN IMAGES

presented by

Umar Farooq

has been accepted towards fulfillment
of the requirements for

Master's degree in Computer Science
& Engineering

Major professor

Date July 31, 1998

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
<hr/>	<hr/>	<hr/>
<hr/>	<hr/>	<hr/>
<hr/>	<hr/>	<hr/>
<hr/>	<hr/>	<hr/>
<hr/>	<hr/>	<hr/>

DETECTION AND RECOGNITION OF FACES IN IMAGES

By

Umar Farooq

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Computer Science and Engineering

1998

ABSTRACT

DETECTION AND RECOGNITION OF FACES IN IMAGES

By

Umar Farooq

The issues of face location and recognition have been studied with great interest over the past three decades by researchers working in the areas of pattern recognition and computer vision. Numerous algorithms and reports on their performance can be found in the literature on various image databases acquired under different imaging environments. In this thesis we propose an automatic face recognition system which can recognize faces appearing in images acquired in relatively uncontrolled environments. The system can be visualized as a computer controlled TV channel selector, which can restrict the channel selection depending upon the viewers watching the TV. A camera placed on the TV set grabs images of the audience, which are processed by an algorithm controlling the channels. We use gray scale images of 640x480 resolution for audience identification. The first step is to locate faces appearing in a test image. A face location algorithm searches for faces at all possible locations with all possible sizes to find the exact location and size of each face in the test image. Complex background and relatively unconstrained imaging environments make this task very complex. The second stage obtains the size and face

location information from the first stage and matches each face with a database of known faces created at the time of training. Once a face has been recognized, the algorithm can initiate some pre-selected actions like restricting access of some TV channels, etc. Image database used for training and evaluation comprised of 300 images with two to four subjects appearing in each image. System was trained on the images of 10 subjects and tested on images of 20 subjects (including the 10 subjects used in training). The results show the effectiveness of our method to recognize faces with large variations in scale, orientation, illumination, expressions, and background. Still, better methods for face location and recognition need to be developed to improve the recognition accuracy. Processing of each image takes a considerable amount of time (~ 90 seconds on a Sun Ultra-1 workstation) due to the computational complexity of our algorithm. This time varies between 45 to 90 seconds depending upon the number of faces appearing in the image and the complexity of the background.

Acknowledgements

I would like to acknowledge all my teachers who have been a source of guidance and inspiration throughout my academic career. I am highly indebted to my advisor, Professor Anil K. Jain, for his continuous encouragement, guidance, and his valuable time, which he spent on this research. It was my privilege to be his student. Without his efforts, commitment, and kindness it would not have been possible for me to complete this work in the given time frame. I would also like to thank Professors George Stockman and John Weng for serving on my Master's thesis examination committee.

I am also grateful to all my colleagues in the PRIP research laboratory who were always willing to render useful advice and provided assistance whenever it was requested. I want to extend special gratitude to Nicolae Duta for his valuable suggestions, time and effort, which he devoted in editing this thesis.

I would like to dedicate this thesis to my family who constantly encouraged me and they were the most important source of motivation for me. Without their love, patience and support, I would not have completed this thesis. I owe my accomplishment to them.

Finally, I want to thank those friends who helped me in collecting the data required for the research and all the members of PRIP lab who are so friendly and collectively maintain a very conducive research environment. It was a nice experience to be a member of this wonderful team.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1	Introduction	1
1.1	Historical Overview	1
1.2	Problem Definition	3
1.3	Recent Work	4
1.3.1	Face Location	4
1.3.2	Face Recognition	6
1.3.3	Performance Evaluation	7
1.4	Applications	8
1.5	Thesis Outline	9
2	Face Detection	10
2.1	Correlation Template	11
2.1.1	Using Edge Images	12
2.1.2	Eigenface Method	13
2.2	Deformable Templates	15
2.3	Image Invariance	16
2.3.1	View-based Learning Algorithms	16
2.3.2	Neural Net Approach	17
2.4	Face Detection from Image Sequences	19
2.5	Discussion	20
3	Face Recognition	22
3.1	Hand-crafted Shape Rule Method	23
3.1.1	Geometric Feature Based Matching	24
3.2	Elastic Matching	25
3.3	Bayesian Similarity Measure for Recognition	26
3.4	Eigenfaces	28
3.4	Neural Network Approach	29
3.5	Linear Subspaces and Discriminant Analysis	31
3.6	SHOSLIF	33
3.7	Discussion	33

4	Automatic Face Detection and Recognition System	35
4.1	Implementation Approach	35
4.2	Face Detection	36
4.2.1	Feature Extraction	36
4.2.2	Learning Phase	38
4.2.3	Detection Phase	39
4.3	Preprocessing Before Recognition	41
4.3.1	Image Cropping and Scaling	41
4.4	Principal Component Analysis	43
4.5	Training of the System	45
4.6	Testing Phase	48
4.7	Summary	49
5	Experimental Results and Analysis	50
5.1	Image Database	50
5.1.1	Imaging Environment	50
5.1.2	Training Set	51
5.1.3	Test Set	53
5.2	Experimental Setup and Decision Strategy	54
5.3	System Performance	60
5.3.1	Recognition Performance for Various Sizes of Training Set	60
5.4	Analysis of the Results	62
5.4.1	Rejection Rate	62
5.4.2	Effects of Background Masking	65
5.5	Summary	67
6	Conclusions	68
6.1	Summary	68
6.2	Future Research	69

LIST OF TABLES

5.1	Recognition results for 10 subjects based upon the first match.	61
5.2	Recognition results for 15 training images per subject with reject option.	63
5.3	Recognition results for 20 training images per subject with reject option.	63
5.4	Recognition results for 25 training images per subject with reject option.	63
5.5	Recognition results for 25 training images per subject after masking the background.	67

LIST OF FIGURES

1.1	Block diagram of an automatic face recognition system.	3
1.2	Examples of faces with cluttered background.	5
1.3	Some examples of face images showing variation of expressions, scale, viewpoint, and direction of light source.	7
2.1	Projection of images in the high dimensional feature-space. (a) Face images for which the algorithm was trained and their projections. (b) Faces and their projections for which the algorithm was not trained. (c) Non-face images and their corresponding projections.	14
2.2	Face detection from complex backgrounds showing detection failures, false alarms, and inaccurate segmentation.	18
3.1	Feature extraction using geometric parameterization.	23
3.2	Object recognition using elastic matching.	27
3.3	An auto-association and classification neural network.	30
3.4	Projections of the same data for Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).	32
4.1	Image cropping and feature extraction.	37
4.2	Learning of face concept.	39
4.3	Process of face detection.	40
4.4	Cropping and preprocessing of face images for PCA algorithm.	42
4.5	The mean face representation.	44
4.6	First five eigenfaces.	45
4.7	Cropping inconsistencies resulting from changes in lighting, orientation, and distance between the camera and the subjects.	47
5.1	Variations of light, orientation, and background in the images collected for the experiments.	52
5.2	Examples of non-face images accepted by the face location algorithm.	55
5.3	Decision methodology based on acceptance threshold.	56
5.4	Histograms and Receiver Operating Characteristic (ROC) curves for various sizes of training set depicting recognition behavior.	59

5.5	Examples of the images processed by the system. Single boxes depict the faces detected by the face detection algorithm and double boxes show recognized faces.	64
5.6	Examples of test images of known faces rejected by the recognition algorithm.	65
5.7	Masking scheme to reduce unwanted effect of cluttered background. . .	66

Chapter 1

Introduction

1.1 Historical Overview

The foundation of pattern recognition can be traced to Plato [1], which was later extended by Aristotle [2], who drew a dividing line between an *essential property* and *accidental property* of patterns. Today, these terms are more comprehensively covered by definitions of *inter-class* and *intra-class* scatter, respectively. Pattern recognition can be viewed as finding those properties of a category, knowledge of which could lead to an automatic decision mechanism in order to differentiate it from other categories. The literature on decision theory and pattern recognition continues to grow rapidly. Some related disciplines like statistics, machine learning and neural networks expand the foundation of pattern recognition. Other disciplines such as computer vision and speech recognition rely on it heavily. Perceptual psychology, cognitive science, psychobiology, and neuroscience investigate how humans and animals perform pattern recognition. The knowledge acquired from these fields helps us to strengthen the foundation of pattern recognition [3].

Human face detection and recognition has emerged as an important and challenging problem in the field of pattern recognition and computer vision [4,5,16,24]. Humans often seem to perform these difficult tasks effortlessly and routinely, and with surprisingly good accuracy under wide environmental variations such as ambient light, distance to the subject, and orientation of the subject. As a result, automatic face recognition has become an important application domain for pattern recognition researchers. Extensive research has been conducted in the specific areas of face detection and face recognition over the past three decades. Various aspects have been explored by engineers, psychophysicists, and neuroscientists. These efforts have been directed at understanding the human cognitive process and how it could be implemented on machines. A reasonable progress has been made and algorithms are available which claim to perform this task with over 90% accuracy. Despite this progress, none of these algorithms can be thought of as touching the boundaries of human cognitive capabilities.

During the early to mid 1970s, most of the work on face recognition was done using classical pattern classification techniques, involving feature extraction using measured attributes of facial features or face profiles. Research on face recognition technology fell dormant in the 1980s [4]. During this decade, artificial intelligence and symbolic programming were extensively studied. Since the early 1990s, research in face recognition has seen a very significant growth. This change can be attributed to many factors, including (i) increased surveillance needs due to drug trafficking and terrorist activities, (ii) applications involving human computer interfaces, (iii) reemergence of neural networks classifiers with emphasis on real-time processing of data,

(iv) tremendous growth in the processing power of desk top computers, and (v) availability of inexpensive, high capacity storage devices.

1.2 Problem Definition

The problem being addressed here can be summarized as follows: Given still or video images of a scene, detect and identify one or more persons in the image, using a known database of faces. A solution to this problem requires solving the following three sub-problems: (1) detection and segmentation of faces from a cluttered scene, (2) relevant feature extraction, and (3) identification and matching. Figure 1.1 shows a block diagram of one such recognition system.

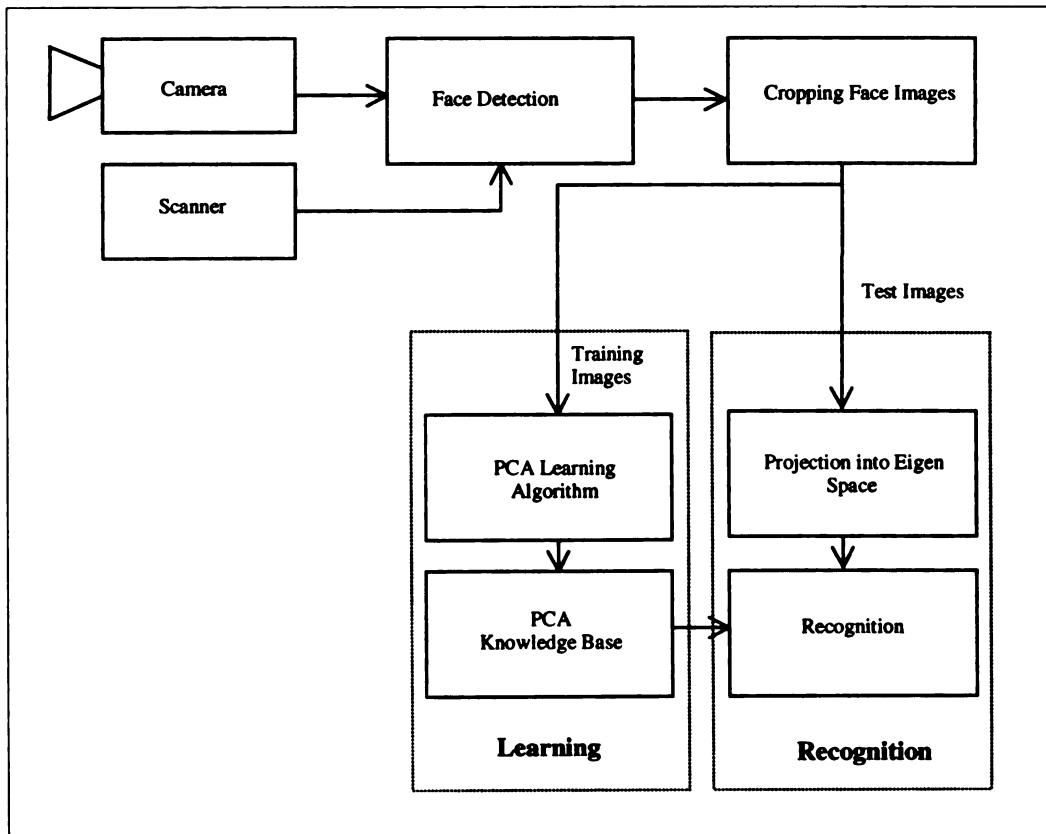


Figure 1.1: Block diagram of an automatic face recognition system.

1.3 Recent Work

The following two subsections introduce the research work done in the relevant areas. More details can be found in Chapters 2 and 3.

1.3.1 Face Location

As mentioned earlier, there has been significant progress in face recognition technology in recent years. Methods have been developed using neural networks, which can locate human faces from a cluttered scene with good identification rates [5]. One of the earliest approaches in face detection technology was by Kelly [6]. This is essentially based upon face detection using edge maps extracted from an input image. Some researchers have based their algorithm on the outline of the head [7], using a segmented template for right-side line, left-side line, and hairline to determine the presence of head. Another approach uses a hierarchical or multi-scale representation of the face image [8]. Burt [9] used the coarse-to-fine approach whereas Shepherd [10] used a coarse-to-fine hierarchical search to locate a face in the given image.

Sirohey [11] segments the face from a cluttered scene using both intensity and edge images. The edge image is generated using the Canny edge detector, and the human face is approximated using an ellipse as an analytical tool. Eigenfaces have also been used to determine the presence or absence of a face in an image [12]. Complexities resulting from factors like a cluttered background, orientation, occlusion and scale variation are obvious from the images shown in Figure 1.2.

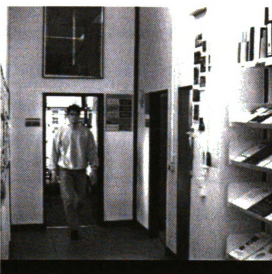
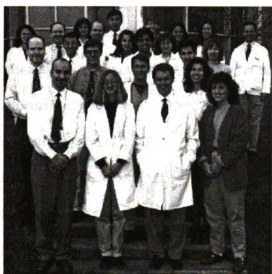
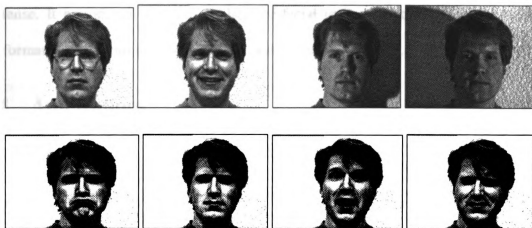


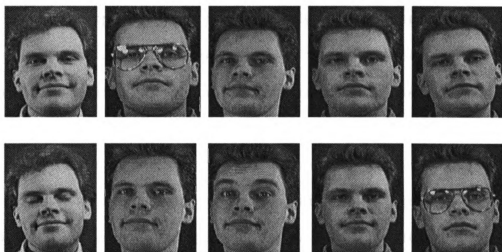
Figure 1.2: Examples of faces with cluttered background.

1.3.2 Face Recognition

Recently, the Karhunen-Loeve (KL) transform for representation and recognition of faces has regenerated interest in face recognition. Sirovich and Kirby [13] approached the problem of face image representation using KL transforms. Once eigenvectors (referred to as “eigenfaces”) are obtained, any image in the ensemble can be approximately reconstructed using a weighted combination of eigenfaces. This approximation improves as the number of eigenvectors used in the image reconstruction is increased. Swets and Weng [14] introduced the concept of Most Discriminating Features (MDF) using the classical discriminant analysis. Yuille *et al.* [7] suggest extraction of facial features using deformable templates. In order to get the best representation, these templates are allowed to rotate, translate, and deform. Figure 1.3 shows some images to illustrate the difficulties involved in the recognition process due to changing expressions, orientations, and illumination conditions. These examples have been taken from two different face databases. Images in Figure 1.3(a) are from the Yale database, which have more variations in illumination and expressions. These images are well framed with a fixed viewpoint. Images in Figure 1.3(b) have been taken from the Olivetti face database, which incorporates orientation variations with changes in expressions and scale. There is, however, no significant change in the intensity and the direction of light illuminating the subject. No particular effort has been made to normalize the images in the Olivetti database by aligning some common features like eyes of the subjects. Effect of background is more significant in Figure 1.3(a), but still it is not as complex as expected in the application being considered in this thesis. A rough estimate of the complexity can be made by comparing the images in Figures 1.2 and 1.3.



(a)



(b)

Figure 1.3: Some examples of face images showing variation of expressions, scale, viewpoint, and direction of light source.

1.3.3 Performance Evaluation

A rapid growth in face recognition technology and need for consistent performance evaluations have prompted the creation of standard test databases. One such example is Face Recognition Technology (FERET), a program sponsored by the US Department of

Defense. It provides a standard database of facial images, in order to benchmark the performance of various algorithms developed by researchers [15].

1.4 Applications

Technological advancements have resulted in improved methods for law enforcement agencies, but criminals also have access to hi-tech equipment and knowledge of surveillance devices so crimes have also become more sophisticated. The major application of face recognition technology is law enforcement and security, however other commercial applications exist as well. The advantages of using computers for face recognition can be summarized as follows: (1) the large size of databases which can be handled by computers, (2) the speed at which information can be processed, and (3) there is no such phenomenon as forgetting or blurring of information with a passage of time as there is in the case of human memory. These obvious advantages are creating more and more interest in the field of automatic face recognition. A few popular applications are processing driver's license and passport applications, crowd surveillance, witness identification, bank and store security, authentication for credit cards and computerized banking, surveillance of sensitive installations like airports, entry control at restricted areas, searching through large face databases, etc. These applications can be grouped under the following two categories that highlight their implementation differences: (1) static matching, and (2) dynamic matching [4]. Static matching refers to the process of matching still images, which are generally acquired under controlled environments to be used for identification purposes. A typical example is mug shot photographs. Similarly, photographs used in documents like passports and driving licenses are categorized as

static images. A typical example of the second category is a sequence of video images. These images could be from a surveillance camera or a clip of a movie, but in both these cases the subject's orientation, illumination direction, and image quality may not be ideal for processing by a recognition algorithm. On the other hand, it is relatively easy to control scale, orientation, lighting and camera characteristics, in the case of static matching. These constraints, which may not be difficult to implement in all application domains, certainly make the task of subsequent recognition much easier as compared to an unconstrained situation. The advantage of dynamic matching is that we have more than one image available to look for a face. Using motion detection and background subtraction techniques, a significant reduction in the search space is also possible. The additional requirement of near real time processing also exists in many such applications.

1.5 Thesis Outline

The organization of this thesis is as follows. Chapter 2 briefly reviews the relevant literature on face detection and image segmentation. Chapter 3 covers issues in face recognition, various techniques for recognition, and their comparative performance. The architecture and implementation details of the proposed system are covered in Chapter 4. Experimental results are presented in Chapter 5, and Chapter 6 presents conclusions and some ideas for future research.

Chapter 2

Face Detection

Any recognition algorithm requires information about the subject's location in an image under test. Therefore, recognition of a subject is always preceded by its detection. In applications involving face recognition, where imaging conditions are not controlled and faces may appear in any part of the image, face detection is the first important step of a fully automatic human face recognizer. Besides its commercial applications, face detection is interesting from an academic viewpoint because faces make up a challenging class of naturally structured objects with fairly complex pattern variations [19]. Face detection is difficult because face patterns can have significantly varying appearances due to different expressions, facial hair, glasses, hairstyles, and orientation. Therefore, classical pattern recognition methods which are good for rigid objects with small intra-class variations tend to perform poorly for face detection. This can only be avoided under well-constrained situations, which can ensure that the face of a subject always appears at a given location, orientation, and scale. Most practical applications, however, require less restrictive implementations. Therefore, a more general method of face detection becomes

an essential prerequisite to recognition for such systems. This situation is further complicated when background conditions are also unconstrained and multiple faces may appear in the test image. Face location becomes a real challenging problem when no constraints are exercised on the scale, orientation, camera characteristics, and subject illumination. Suggested solutions can be broadly categorized in the following two classes: (1) learning, and (2) non-learning methods. Another sub-division is: (1) correlation template matching, (2) deformable template matching, and (3) image invariants [19]. Additional techniques like detection using motion and background subtraction are useful only for a sequence of images obtained from video cameras. In the following sections we briefly review these methods and draw some comparisons on their relative performance.

2.1 Correlation Template Matching

This technique is based upon a difference measurement between a candidate and a fixed reference pattern. A decision is made as to whether the candidate is a match or not based on whether its distance to the reference pattern is below or exceeds a pre-selected threshold value. Due to variations of face patterns, it is difficult to capture all possible representations in a single reference model. The use of multiple correlation templates is one solution to such a problem. Two important implementation strategies are described in the following subsections.

2.1.1 Using Edge Images

This is the earliest approach reported in the literature, proposed by Sakai *et al.* [17]. An edge image extracted from a gray level image is used to locate an oval shaped outline resembling a human head. The template must be matched at all possible positions with all possible sizes over the entire image. Positions where potential matches are reported are searched using a detailed feature match at the expected locations of the eyes, nose, and lips. A system proposed by Kelly [6] was the first to perform the task of segmentation automatically. The approach was based on a top down analysis of the image. In the first step, the body outline is located by subtracting the background from the image. The outline of the head is located using smoothed versions of the original image. These extracted features are projected back on to the original image and then a detailed search for essential features is carried out at those locations. This search involves many heuristics for confidence measurements of potential candidates. The technique is useful for images with a known background.

Craw *et al.* [8] describe a method to extract the head area from an image. The image is presented to the search algorithm at multiple resolutions, starting with 8x8 pixels. This resolution is doubled for each hierarchical step up to 128x128. The head outline template is constructed at the lowest resolution. A Sobel mask is used for calculating the edge magnitude from a gray scale image. The head outline is constructed using a line following algorithm which is followed by a search for more detailed features such as eyes, eyebrows, and lips based upon knowledge of the head outline. Performance of the edge detection technique suffers when the background is cluttered, or when the changes

in the intensity or the direction of illumination are significant. Therefore, such methods are useful only for relatively controlled imaging environments.

2.1.2 Eigenface Method

Turk and Pentland [12] suggest using eigenfaces for face detection. Their approach is based upon the idea of Principal Component Analysis (PCA), which is used for face recognition as well. The key idea is that the projections of different faces in a high dimensional space, do not change as radically as do other objects. This fact has been illustrated in Figure 2.1, where the projections of those face images for which the algorithm was trained, have been compared with projections of faces which were not part of the training set and few “non-face” objects. The images of size 64x88 were projected in the eigenspace computed for 64 training images retaining the first 30 eigenfaces, which account for 95% of the total variance. In view of this observation, the distance between a test image and a face template can give a good measure of “faceness”. This distance is calculated over the entire image space by sampling it with a stepping window over the image. The area captured by the window is then projected into the eigen-space and compared with a prototype face projection. Windows closer to the prototypical face will be expected to contain a face image. Since the size of the face appearing in an image may not be known, this sampling process has to be repeated several times for various window sizes. Clearly, a direct application of this technique is computationally expensive, so the authors suggest an alternate method to reduce the search space by detecting motion and then applying this strategy more efficiently. This technique, which is useful for detection

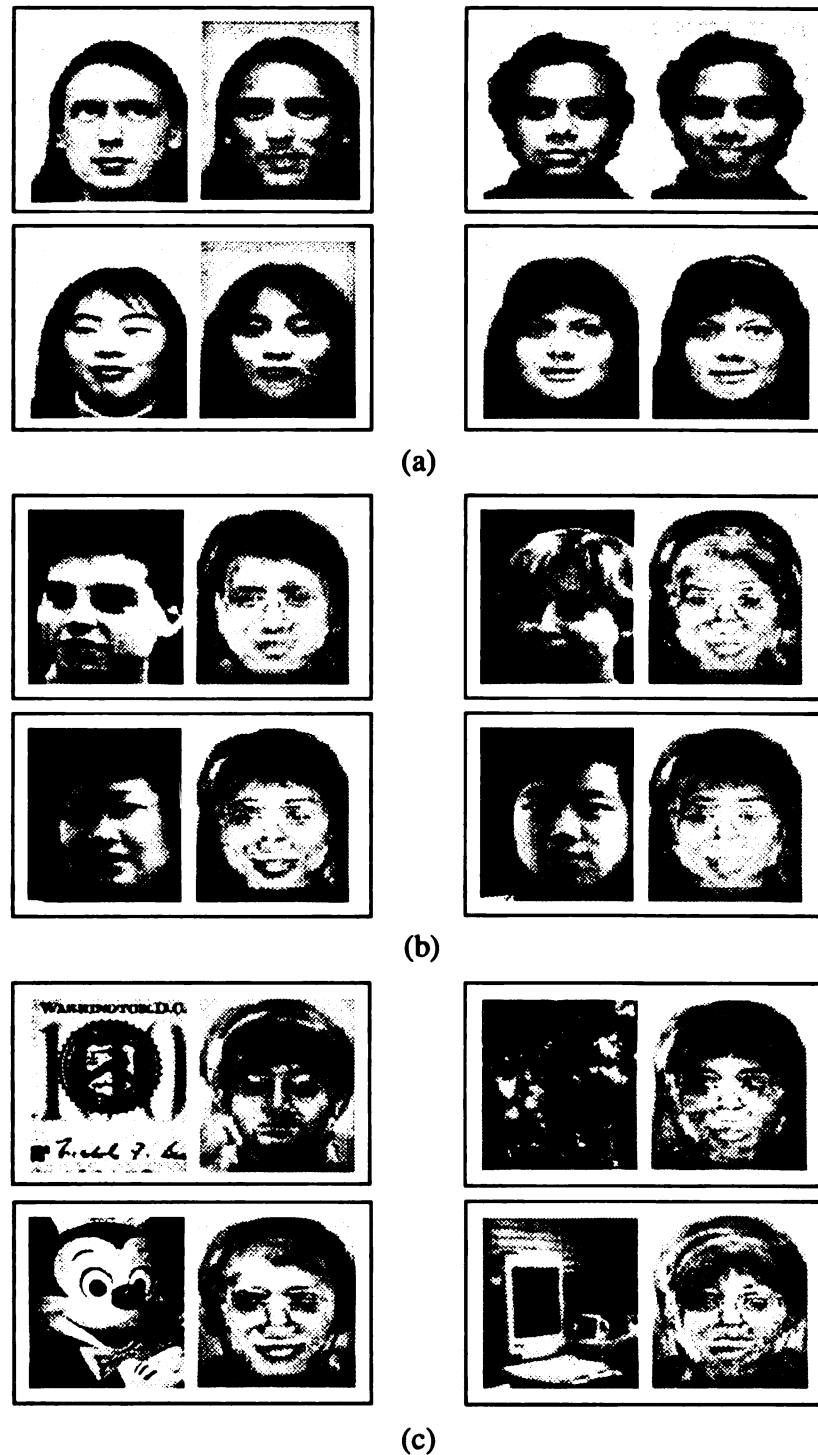


Figure 2.1: Projection of images in the high dimensional feature-space. (a) Face images for which the algorithm was trained and their projections. (b) Faces and their projections for which the algorithm was not trained. (c) Non-face images and their corresponding projections.

from video sequences, has been explained in Section 2.4. An important point to note is that such a measure is useful only for detecting those faces from the images for which the recognition algorithm has been trained. Our results show that projections of faces, for which the system has not been trained, are very similar to “non-faces” and are rejected at about the same rate.

2.2 Deformable Templates

Attempts have been made to improve template matching and make it more rugged when working with cluttered backgrounds. These methods are based on a search using deformable templates. This approach is closely related to the correlation templates described above. The correlation templates used for this purpose have a built-in non-rigidity component. The aim is to make the template just flexible enough to accommodate variations caused by non-rigid features like eyes and lips. Yuille *et al.* [7] proposed a method of locating faces using deformable templates. A slightly modified approach was later adopted by Govindaraju *et al.* [21] who defined their template based on the head outline. The template consists of three segments that represent the curvature discontinuities of the human head, i.e., right-side-line, left-side-line, and the hairline of the head. For these three curves, four features are extracted from each and used for confidence measurements. These features are the length of the curve, the chord in vector form, the area enclosed between the curves, and centroid of this area. The presence of a head is ascertained by finding all these features in an image at any location with a particular orientation. The templates are allowed to translate, scale and rotate according to a spring-based model. The center of these features represents the potential location of

the center of the face. The authors make the claim that their system never failed in finding a face in their test images, but they do not report the false alarm rate.

2.3 Image Invariance

These schemes [19] are based on the assumption that all face images have common spatial image relationships, which are possibly unique to all face patterns even under different imaging conditions. These invariants can be face templates, relative positions of face features, complexion, and texture. Such features when extracted and suitably grouped can be used to detect a face in an image. The algorithm looks for portions of an image containing such invariants.

2.3.1 View-based Learning Algorithms

Sung and Poggio [19] base their solution on one such algorithm, which calculates representative clusters of feature vectors in a high dimensional space. Faces are treated as a class of spatially local target patterns, which can be represented by compact clusters in the high dimensional space. With this assumption, clusters of “face-like” images should be closer to a face test image as compared to “non-face” samples. Training images are normalized and scaled before clustering. Both positive and negative examples are clustered in the high dimensional space using a modified k-means algorithm. Their algorithm results in a total of 12 clusters, six each for positive and negative examples. For each test image, two types of distances are computed, one Euclidean and the other Mahalanobis distance from each cluster. These pairs of distances are used as a feature vector for classification. The final decision stage is a multilayer perceptron, whose output

depends upon these distance measurements. System performance has been evaluated for various classifier architectures like nearest-neighbor classifier and single perceptron unit. A multilayer perceptron proved to be the best option for their algorithm.

A scheme of learning the face concept from gray scale images, principally based upon texture analysis has been described by Duta and Jain [16]. Thresholded output, from three cascaded classifiers, is used to determine the existence of a face-like pattern in the test image. A detailed description of this system can be found in Chapter 4.

2.3.2 Neural Net Approach

Problems like face recognition, gender classification, and classification of facial expression have been addressed using neural networks [4]. As such, researchers working on face location systems have also been attracted to use neural networks [5]. From an academic viewpoint, most of the implementations involving neural nets are hybrid in nature, where neural nets are the primary, but not the only classifier in a decision fusion. The reason for the popularity of neural networks can be attributed to the fact that the cognitive processes, which we (humans) use in face recognition and detection, are still very little understood. Adaptive systems like neural nets, which can be trained by presenting the system with examples, are good approaches for solving such problems. Despite some limitations such as heuristics involved in the convergence process during training, long training periods, and the requirement of large sized training set to achieve acceptable performance, neural networks remain an attractive option.

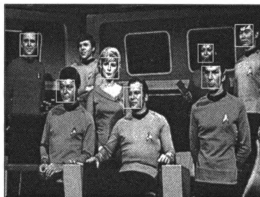


Figure 2.2: Face detection from complex backgrounds showing detection failures, false alarms, and inaccurate segmentation (Source: CMU face detection demo on the web).

Lin *et al.* [18] describe an approach for locating deformable objects using a probabilistic decision-based neural network. The study focuses on detection of human faces from images for surveillance and video browsing. Sub-nets of the system are designed to model the log-likelihood functions of the object classes. Underlying distributions have been assumed to be Gaussian. The output of the network serves as a measure of confidence for the presence or absence of a face in an image.

Rowley *et al.* [5] describe a neural network-based face detection system. Their system operates in two stages: the first stage applies a set of neural network based filters to each location in the image at several scales, searching for potential candidates. The second stage is an arbitrator, which detects and eliminates overlapping. The architecture of the neural net is based on retinal connections to their input layers. In order to detect faces of various sizes, a pyramid of test images is generated. This strategy ensures that one design of neural network can process the test image for all possible face sizes. Authors report a detection rate of 91.5% to 100% for various orientation angles. Figure 2.2 shows example images tested using their algorithm. The training set included negative examples as well, which were selected in a progressive training process.

2.4 Face Detection from Image Sequences

Segmentation of a moving object, from a video sequence, is the most important area of image sequence analysis with direct applications to face recognition [4]. This segmentation process is based on the fact that people are constantly moving. Even when we are sitting, we keep changing the position of our head, adjusting our body position and blinking our eyes. Simple motion detection and tracking algorithms can estimate the

location of the head from a given image sequence. Knowledge of the face space can be used to more precisely locate the head in conjunction with a tracking algorithm. A template-based strategy can also be used to track a face. One such approach has been described by Bichsel and Pentland [20]. It utilizes a minimum number of different face templates, determined by an analysis of geometric transformations such as scaling, rotation, and translation. A coarse-to-fine approach is used to zoom in for a possible face in the image. A rough estimate is refined with successive finer scales by tightening the acceptance threshold. As mentioned earlier, thresholding the difference between consecutive frames is one of the simplest methods of detecting moving objects. An analysis of image difference becomes difficult in the case of illumination changes, occlusion, or when camera is moving. More sophisticated segmentation techniques rely on analyzing the optical flow field. Accurate computation of optical flow is an unresolved problem, which has led researchers to make use of other flow fields like image flow [4]. These types of analysis also become more complicated when either the subject or camera is moving.

2.5 Discussion

From the preceding analysis, it can be concluded that all the detection systems rely on scanning the image space at various scales to search for the presence of a desired pattern. Face detection is therefore inherently computation intensive. Search space can be reduced substantially by employing the techniques of background subtraction and movement tracking. The system described in [5] promises a high detection rate, but the system

architecture and training is, however, complex. The scheme presented in [16] is slower, but offers good detection performance with a low false alarm rate.

Chapter 3

Face Recognition

Faces are complex, multidimensional, and meaningful visual stimuli, which play a dominant role in our daily social interaction. Besides being the most important source of our identity, faces also represent our emotions and reactions to events happening around us. Face images are used most widely in the driver licenses, passports and other such identification documents. Therefore, it has always been desirable to develop an automatic face recognition system which can be used to automate the processing of such documents and for a variety of tasks such as automatic access control, banking, ATM machines, crowd surveillance, etc.

Faces are semi-rigid objects and, therefore, developing a computational model for automatic face recognition is quite difficult [12]. Many techniques have been developed and tested over the last three decades to find a suitable solution to this important problem in computer vision and results have been reported with varying degrees of accuracy. All techniques suggested so far, for automatic face recognition, are essentially based on minimum distance classification. Suitable features are extracted from the training images and used for classification of a test image by calculating its distance from all the training

patterns (or a representative subset of the training patterns) in the feature space. A test image is assigned to the class, which is at a minimum distance from the test pattern in the high dimensional space. Feature extraction techniques primarily distinguish one method from the other. Notable techniques are described in the rest of this chapter, with specific reference to their implementation. The last section comprises a comparative analysis of these methods, within the framework of the particular application being addressed in this thesis.

3.1 Handcrafted Shape Rule Method

Initial research emphasis had been on the design of efficient matching algorithms, from a manually designed feature set, with handcrafted shape rules. Inherently, these techniques are computationally expensive and difficult to implement. The primary problem is the need for accurate and efficient location and segmentation of face features, before the matching algorithm could be invoked for recognition. Images shown in Figure 3.1 depict processing involved in one such feature extraction technique.

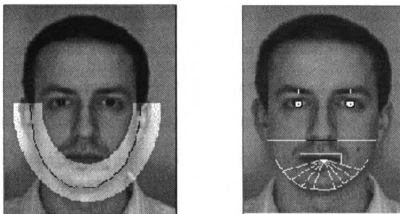


Figure 3.1: Feature extraction using geometric parameterization (Source: ref. [24]).

3.1.1 Geometric Feature Based Matching

This is the earliest scheme suggested for face recognition by computer [4]. Classification of a test image is done by comparing the relative position of facial features with those of training images. Features used for classification can either be extracted manually or automatically by the machine. Locations of the eyes, nose, mouth and chin are useful features, which have been most commonly used to form a feature vector. The distance between these features is measured and then normalized before using them for classification. Normalization is desired to make the algorithm scale invariant. Translation invariance can be achieved if we can locate one common point, say the tip of the nose in the test and training images. Such a point can serve as the origin of the coordinate system in which the image is represented. If we are able to locate just another reference point, rotation invariance shall also be possible. If features are measured in terms of relative angles instead of relative distances between facial features, scale invariance is automatically built in to the system. When all this preprocessing has been done, correlation between the test and training images can be found by using any standard classification technique. This correlation can be made robust against illumination variations by intensity normalization, which is done by normalizing the test image pixels over a suitably large neighborhood. Accurate extraction of features without human intervention is the most challenging task in this strategy. The issue becomes even more difficult when lighting, expressions, background, and viewing angle change significantly.

Kelly [6] presented an automatic approach to feature extraction in his doctoral thesis on face recognition. It was the first attempt to fully automate the feature extraction and

recognition process. His approach was based on a top-down methodology of feature extraction, locating the body and head of the subject first and taking various measurements like height and width of head, neck, and shoulders. Distances between eyes, nose and corners of the lips were also measured at a finer level. A nearest-neighbor classifier was used for the final recognition task. An improvement to this geometric parameterization was accomplished by Kanade [22]. He worked with distances as well as angles between facial points such as eye corners, mouth extremities, chin top, and nose. Feature extraction was done in two stages. Initially, a coarse search was done to identify four sections of interest from a low-resolution image. These sections which contained the left eye, right eye, nose, and mouth were again processed at a high resolution for accurate location. A total of 16 features were extracted in the form of angles, distance ratios, and areas spanned by facial features. These parameters were normalized to make the system scale invariant. The author reported recognition rates of 45% to 75%, depending upon the number of features used. A better identification rate was reported when some ineffective features were disregarded.

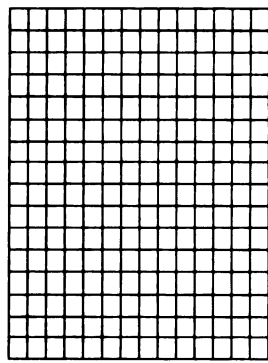
3.2 Elastic Matching

Unlike rigid objects, the appearance of faces can change significantly due to changes in expressions and viewing angles. As described earlier, such changes affect the performance of methods based upon geometric correlation. A solution to this problem, known as elastic matching is described by Zhang *et al.* [23]. Their algorithm defines a face template based on a lattice, which is of much lower resolution than the original image. Features are extracted by a convolution of the image and a two-dimensional

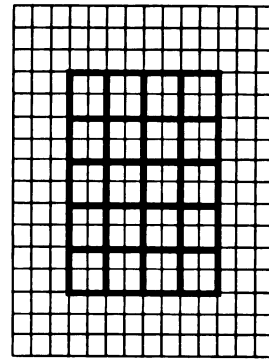
Gabor filter at each intersection of the lattice or grid. When an image is presented for classification, a similar grid is superimposed on the test image. The best match for each intersection point is found by allowing the test image to deform elastically, within specified limits. Selection of the test image class is made using the template with the minimum overall energy difference. Results reported show that this method is robust to geometric variations caused by the aforementioned factors. Successful recognition rates of 80% to 100% have been reported over four different image databases. Feature extraction and matching techniques are elaborated in Figure 3.2.

3.3 Bayesian Similarity Measure for Recognition

Moghaddam *et al.* [28] suggest an object recognition scheme based upon Bayesian analysis of image deformations. They model two types of variation in object appearance. One type represents within class variation (intra-class) and the other encodes variations between the classes (inter-class). Training data is used to estimate probability density functions for each class from the feature vectors which are the warping coefficients computed for both the classes. From these probability density estimates, *a posteriori* probabilities for each class are computed. These probabilities serve as a measure of similarity for the two classes in contrast to other methods, which use the distance between the probe and a training sample as a measure. They calculate an optimal non-linear decision boundary for recognition, by equating the two *a posteriori* probabilities. The authors also introduce a novel approach for computing image deformations in three-dimensional XYI space. The spatial variations are represented by XY and the intensity variations by the I axis. Deformation cost is computed by image warping in this XYI space. Test results have been reported on a selected subset of the FERET face database

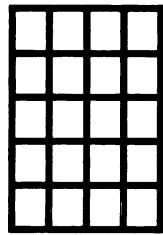


Sample Image

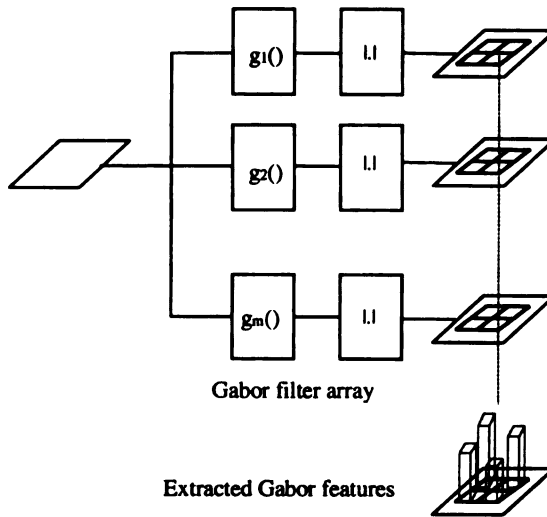


Low resolution grid
superimposed upon image

(a)

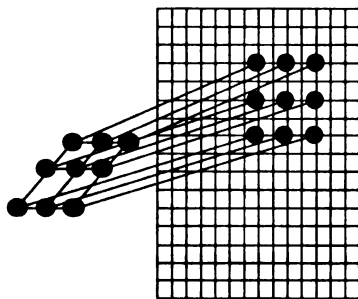


Sampled image

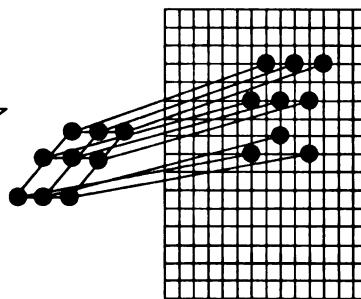


Extracted Gabor features

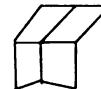
(b) Feature extraction using Gabor filters



Matching without deformation



Matching with deformation



(c)

Figure 3.2: Object recognition using elastic matching (Source: ref. [23]).

which show performance superiority of their approach over similar techniques based upon simpler representations like intensity differences and image flow. Due to computationally expensive XYI warping, the authors use PCA as a preprocessing stage. The top 10 matches from PCA stage are processed to compute the XYI warping.

3.4 Eigenfaces

As described earlier, the classifiers used for recognition are typically minimum distance classifiers. Correlation techniques, whether based on pixel to pixel matching or geometric parameter matching, are computationally expensive. In order to condense the large image space, while preserving the details required for recognition and discrimination, principal component analysis (PCA) has proved to be effective. Turk and Pentland [12] describe the details of this scheme. This method can be summarized as follows: each face in the database can be represented as a vector of weights; the weights are obtained by projecting the face in the so called, Most Expressive Feature (MEF) space. These features are extracted from the two-dimensional training images by computing the eigenvectors of their covariance matrix. Only those vectors are retained which correspond to the largest eigenvalues. The number of these vectors, to be used in projection, is selected such that most of the variance in the training set is retained. When a test image, whose identification is required, is presented, its projection in the MEF space is calculated. A match is found by locating the face in the database, which is at the minimum Euclidean distance in the feature space from the test image. The method is fast, accurate and space efficient, however, it is sensitive to scale, ambient light, orientation and background changes. Implementation details of the algorithm are given in Chapter 4. Comparisons of

eigenface method with other recognition techniques can be found in [23,24,27]. Under controlled environmental conditions, this method is quite accurate and efficient both in time and space.

3.5 Neural Network Approach

Face recognition using neural nets can be accomplished using various learning mechanisms. It can be realized with the classical back propagation learning, using the concept of associative memory, or exploiting the properties of radial basis functions. This approach can lead to classifiers that exhibit good tolerance to noise and are reasonably immune to environmental changes. Multilayer feed forward networks are useful tools which have been used for classification of images by many researchers [4]. Despite problems such as long training periods and heuristics involved in the parameter selection, neural nets are popular tools for performing classification tasks under difficult operating environments. The most prohibitive problem in face recognition is the dimensionality of the image, which has a direct relationship to the complexity of the network. Various techniques have been suggested in the literature to reduce the image dimensionality to manageable limits. Noteworthy methods are auto-association networks, local image sampling, and self-organizing maps [25].

An auto-association network is composed of three layers, forming a fully connected feed forward network. The input and output layers contain as many nodes as the dimensionality of the input images. Nodes in the hidden layer are much fewer than the other two layers. A schematic diagram of an auto-associative network and a classification perceptron is shown in Figure 3.3.

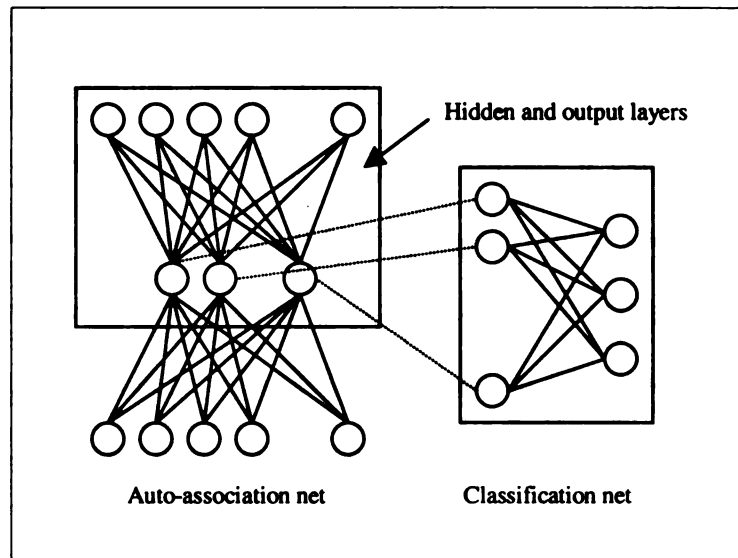


Figure 3.3: An auto-association and classification neural network.

The network is trained using the classical error back propagation learning. The input takes an image vector X and is trained to produce an output vector Y . This output is the best approximation of the input face. As the network converges, the hidden layer constitutes a compressed version of the input. This compressed vector, which is a representation of the input face image with reduced dimensions, is used as a feature vector for the classification of a test image. Bourlard and Kamp [26] have investigated the nature of this compressed vector. The authors showed that under the best circumstances, using linear transfer functions at nodes, the hidden layer produces a feature vector which is the same as KL basis, or projection of input in the eigenspace. On the other hand, if a nonlinear transfer function is used, the optimal performance cannot be guaranteed, as the vector could deviate from the best approximation.

The technique of self-organizing map is an unsupervised learning process, which learns the distribution of training patterns without any category information. Training patterns

are projected from the input space to the corresponding positions in the information map. Unlike clustering algorithms, this method encodes the mapping information in a topological ordering of the classes. This mapping technique is like the one found in the human nervous system for image mapping in the visual cortex. A test image is projected in the mapped space and is classified as its closest topological neighbor. Neural network based recognition systems have been used in conjunction with other techniques to form hybrid systems [4]. Other techniques involve dimensionality reduction methodology based on scaling, low pass filtering and projections to lower dimension spaces. In such arrangements, a neural network based classifier is mostly used as the final classification stage.

3.6 Linear Subspaces and Discriminant Analysis

Face recognition techniques based on correlation and image projection in a low dimensional space, like the eigenface approach, tend to perform poorly when the subject illumination, and orientation are changed. Moreover, the projection algorithm does not take into account the inter-class and intra-class variability while transforming to eigenspace. Fisher Linear Discriminant Projection, on the other hand, is calculated so as to maximize the compactness of and separation between various classes [3]. Figure 3.4 illustrates both the projections. Any performance gain of LDA over PCA in terms of better class separation in the low dimensional space is hard to predict as the results depend on the data distribution in the high dimensional space. In order to make the recognition system insensitive to lighting variations, an extension to this method is suggested by Belhumeur *et al.* [27].

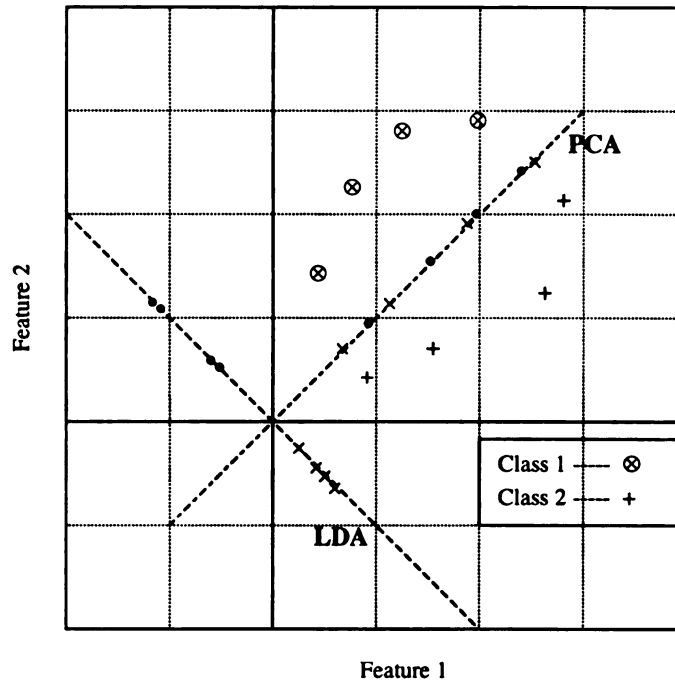


Figure 3.4: Projections of the same data for Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

The technique is based upon the concept that for a Lambertian surface, images of a particular face lie in a 3D linear subspace if there is no shadowing. Given three images of each subject, from one view point, with three linearly independent known light sources, we can extract the albedo and surface orientation using the well known method of photometric stereo. Conversely, given three images we can reconstruct a face image under arbitrary lighting conditions using a linear combination of the three training images. Therefore, if faces are considered as Lambertian surfaces, the above stated fact leads to a classification method which is robust to illumination variations over a wide range. Experimental results report a high degree of robustness using this method for large lighting variations. The most important constraint in the above stated technique is that of viewpoint, which may not be possible to exercise in most practical applications.

3.7 SHOSLIF

The importance of comprehensive learning and its applications in the field of computer vision has motivated the researchers to develop systems capable of operating in complex real-world environments. SHOSLIF is one such system developed in PRIP research laboratory of MSU [29]. SHOSLIF stands for the Self-organizing Hierarchical Optimal Subspace Learning and Interface Framework. The concept is based upon automatic selection of most useful features and organizing this visual information using a coarse-to-fine space partition tree to achieve a logarithmic time complexity for retrieval of relevant information from a large visual database. SHOSLIF uses a core-shell model to make the learning and control part task-independent. The core is common to all the applications whereas the shell which deals with the interfaces to the external world, changes according to a specific application. The author proposes an approach to compute discriminant projections from KL projection which is named as discriminant Kahunen-Loeve (DKL) projections. The new features so extracted are called the Most Discriminant Features (MDF). A subsystem of SHOSLIF called SHOSLIF-O has been designed specifically for object recognition. Performance results have been reported on an image database consisting of both face and non-face images.

3.8 Discussion

Many algorithms have been developed and tested to recognize faces automatically in the course of research conducted over the last three decades. A careful comparison of their performance can help in the selection of a suitable method in a particular application. Among all the techniques introduced in this chapter, PCA is the most attractive solution

for the problem being handled in this thesis. The reasons for this choice are speed, robustness and simplicity of the algorithm. Limitations like sensitivity to scale, orientation and illumination variations, and effect of background can be tackled by suitably processing the test image before classification and selecting a large training set to capture these variations.

Chapter 4

Automatic Face Detection and Recognition System

4.1 Implementation Approach

The problem of interest in this thesis is to integrate face location and recognition algorithms in order to build a working system for relatively uncontrolled environments. A bottom up implementation approach was adopted in the course of developing the system for automatic face location and recognition. The algorithm used for face location was adopted from [16] and the recognition method is based upon Principal Component Analysis (PCA) [12]. These algorithms have been modified to make them suitable for our application. Implementation details are covered in the remaining sections of this chapter. System performance and analysis of the results are provided in Chapter 5.

4.2 Face Detection

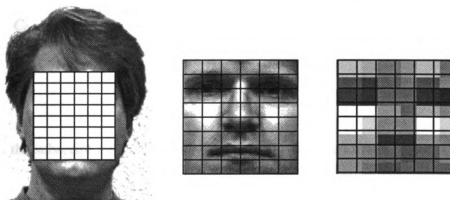
Our approach to face detection is based upon the technique of example-based learning which has been adopted from Duta and Jain [16]. The algorithm learns the human face concept using features extracted from the training data and classifies a test image as a face if some predefined criterion is fulfilled. Three different feature vectors are used to derive the human face concept, which are extracted from the central part of the face. These features are used to train three stages of a classification algorithm. Detection is done by scanning the entire image at all possible scales and locations with a sampling window. The area captured by the window at each step is tested to find the presence of a face. All the candidate locations are tested in three steps for their corresponding feature vectors.

4.2.1 Feature Extraction

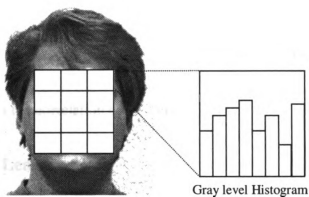
A training data set consisting of 1,200 face images was used to compute the representative features in order to define the face concept. The training set was manually processed to crop and scale the required area of each training face. Resulting images with an aspect ratio of 4x3 covered the central portion of the face. Each test image was processed to extract three feature vectors as shown in Figure 4.1. The first feature vector was extracted by subdividing the sampling window into 48 equal sized windows, 6 of them along the width and 8 along the height of the image. Each sub-window was equalized to one of 32 allowed gray levels. This resulted in a 48-dimensional feature vector with each feature quantized to one of the 32 gray values that could be represented by five bits.



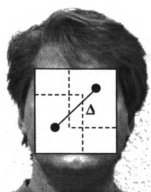
(a) Cropping of a training image to extract the central window



(b) Extraction of first feature vector



(c)



(d)

Second and third feature vectors, based on texture analysis

Figure 4.1: Image cropping and feature extraction.

The second and the third feature sets were based upon texture analysis. In order to capture the second feature vector, the training face was sub-divided into 12 equal sized regions. A gray level histogram was computed for each region with 8 gray levels, thereby resulting in a feature vector of 96 (12x8) dimensions. The third feature set was based upon correlation coefficients computed for 40 pairs of points (in 8 directions and 5 displacements) in the image plane. The correlation coefficients are computed as follows:

$$C_{\Delta} = \frac{1}{K} \sum_{i=1}^K \frac{(X_i - \text{mean}(X_i))(X_{i+\Delta} - \text{mean}(X_{i+\Delta}))}{\text{var}(X_i) \times \text{var}(X_{i+\Delta})}, \quad (4.1)$$

where

- i is the subscript indicating the location (ix, iy) of a pixel in the image.
- X_i , and $X_{i+\Delta}$ are the gray levels at pixels i and $i+\Delta$, respectively.
- Δ is the translation $(\Delta x, \Delta y)$ of the plane.
- K is the number of point pairs in the image considered for calculating the correlation.
- C_{Δ} is the correlation coefficient.

4.2.2 Learning Phase

In order to derive face clusters in the feature space, Ward's clustering algorithm was used. Resulting dendrograms were cut at level 4, where the covariance matrices were non-singular and the clusters formed had some semantic meanings in terms of direction of light illuminating the subjects. For each cluster found by the clustering algorithm, the

following parameters were computed: (1) the centroid of the cluster, (2) inverse of the covariance matrix, and (3) the minimum Mahalanobis radius, such that it contains at least 95% of the population in the cluster. These parameters were computed for each feature vector and stored as the knowledge base for the classification algorithm. A block diagram of the entire learning process is shown in Figure 4.2.

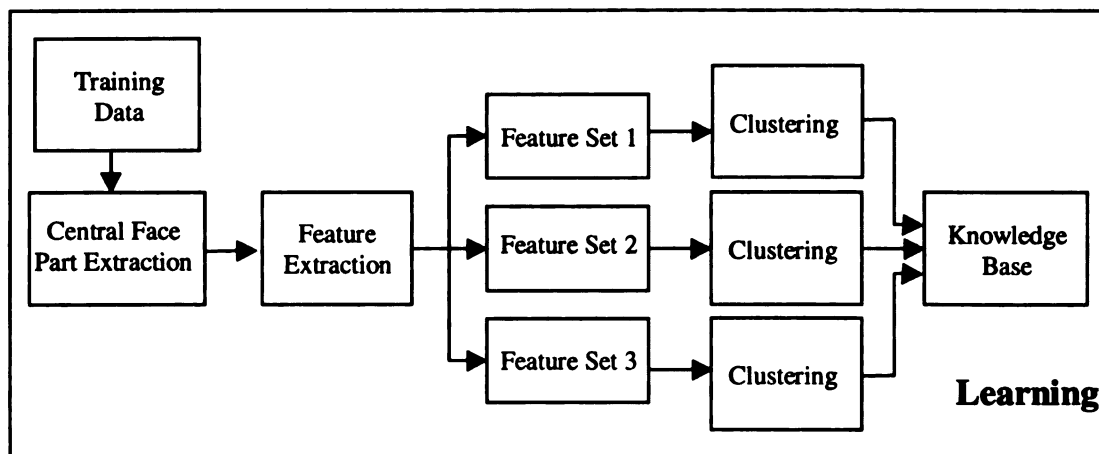


Figure 4.2: Learning of face concept.

4.2.3 Detection Phase

The detection algorithm works as a three-staged classifier, with each feature vector captured and classified at the respective stage. The test image is sampled by a scanning window stepped over the image at each location. Features are extracted from the image area captured by the window. The process is repeated for various sizes of sampling window to look for all possible sizes of a face appearing in the test image. If the distance of the extracted feature vector from the centroid of any cluster is found to be less than or equal to the learned Mahalanobis radii, computed during learning, then the window is accepted to contain a face by the classification stage. In order to reduce the computations

involved, only the first feature vector is extracted and tested for all the windows generated by the sampling process. The second stage only processes those windows that meet the required criteria at the first stage. A similar strategy is followed while processing candidate windows at the third stage as well. So, in the overall perspective, the classifiers operate in a cascaded arrangement rather than in parallel, which avoids computationally expensive feature extraction and classification involved in the last two stages by filtering out a huge number of windows generated during the sampling process. A block diagram of the face detection process is shown in Figure 4.3.

Finally, all locations, identified to contain face images are further processed to eliminate multiple detections of one face at different scales and overlapping positions. The sizes and center of the detected face are estimated by weighted averaging over various detections and, finally, coordinates of the window containing the face are computed.

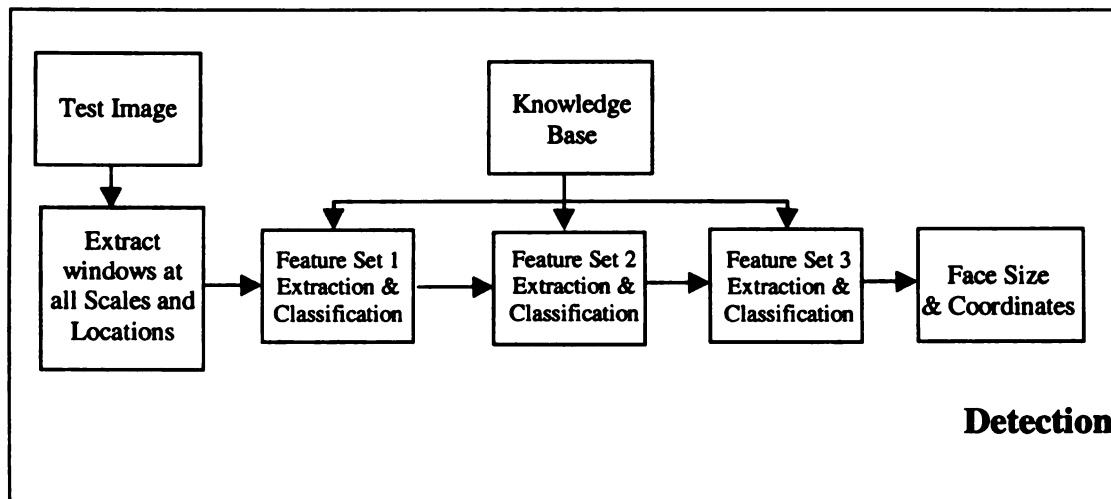


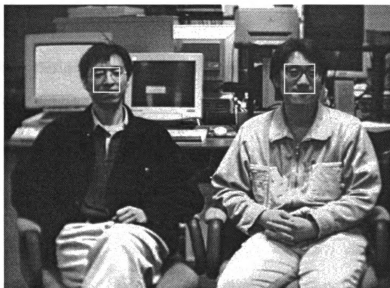
Figure 4.3: Process of face detection.

4.3 Preprocessing Before Recognition

Portions of the test image that contain faces are passed on to the recognition algorithm. Since no particular constraint has been exercised on the distance between the subject and the camera, background, lighting and orientation, except that the faces should be upright, the faces detected in a test image may not be of the same size. Besides this inconsistency of scale, effect of the background is also significant in test and training data. In order to compensate for these variations, some kind of pre-processing is required before the face images could be used by the recognition algorithm. The face size normalization is illustrated in Figure 4.4.

4.3.1 Image Cropping and Scaling

The first step, after face location, is to obtain the portion of the test image where a face has been located. This process of image cropping is based upon the information generated by the face detection algorithm. It is evident from Figure 4.4(a) that the actual coordinates generated in the detection process may not cover the entire face area useful for recognition. Therefore, the face-cropping algorithm modifies the coordinates by doubling the detected width and setting the height of the cropped image so as to maintain an aspect ratio of 4x5. After cropping the face image, it is scaled down/up to a standard size of 40x50 pixels. This step ensures that the recognition algorithm requires only a single database of face images.



(a) Detection of faces from a given image



(b) Cropped out face images according to the detected size



(c) Face images scaled to the standard size of 40x50 pixels

Figure 4.4: Cropping and preprocessing of face images for PCA algorithm.

An alternative approach would have been to create multiple databases for various scales using multiple sized training images and selecting the most appropriate database at the time of recognition. Besides being inefficient in terms of space and learning time, this technique can result in higher error rates due to quantizing error introduced in cropping the images.

4.4 Principal Component Analysis

The classifiers used for face recognition are mostly minimum distance classifiers. Correlation techniques, whether based on pixel to pixel matching or geometric parameter matching, are computationally expensive. In order to condense large image space while preserving details required for recognition and discrimination, Karhunen-Loeve expansion, also known as PCA, has proven its effectiveness. This method can be summarized as follows: each face in the database can be represented as a vector of weights; the weights are obtained by projecting the images in the so called Most Expressive Feature (MEF) space or eigenspace. These features are calculated from two-dimensional face images by computing eigenvectors of covariance matrix resulting from high dimensional representation of training images. Each image is represented in the form of a single high dimensional vector formed by concatenating the gray values of all its columns. Only those m eigenvectors are considered for projection which correspond to the m largest eigenvalues. The value of m is selected such that we retain most of the variance present in the training set. When a face image is presented whose identification is required, its projection in the lower-dimensional eigenspace is calculated. A match is found by locating a face in the database, which is at the minimum Euclidean distance in

the feature space from the test image. The method is fast, accurate and space efficient. It is, however, sensitive to scale, ambient light, orientation and background changes, effects of which are minimized by the processing as described in Section 4.3. Details of the PCA algorithm are as follows:

- All the pixels of each training image are concatenated to form one large vector of dimension $d \times 1$, where $d = \{\text{No. of columns}\} \times \{\text{No. of rows}\}$ in an image, so each image is converted to a vector whose features are gray levels of each pixel in the image. In our implementation, the face images were scaled to 40×50 , resulting in a vector of 2,000 dimensions.
- The mean (average) image is computed from all the vectors in the training image set;

$$\bar{Z} = \frac{1}{N} \sum_{i=1}^N Z_i, \quad (4.2)$$

where $\{Z_1, Z_2, \dots, Z_N\}$ are the column vectors obtained from the N training images. Figure 4.5 shows the mean image.



Figure 4.5: The mean face representation.

- The covariance matrix C of the ensemble is calculated by using the following expression;

$$C = \frac{1}{N} \sum_{i=1}^N (Z_i - \bar{Z})(Z_i - \bar{Z})^T. \quad (4.3)$$

- Compute the eigenvectors of the $d \times d$ covariance matrix C . The eigenvectors are a set of orthonormal vectors. The d eigenvalues are sorted in decreasing order such that

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d.$$

- Calculate the number of eigenvectors, m , that should be retained to preserve most of the variance in the training images. As a guideline, 95% of the variance is retained.

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{j=1}^d \lambda_j} = 0.95 \quad (4.4)$$

The value of m which satisfies Eq. (4.4) gives the number of eigenvectors, corresponding to the m most significant eigenvalues which should be retained. The value of m depends upon the number of images used for training. For the purpose of this implementation, first 85 eigenfaces were retained for a training set of 250 images. As a result of the above process, we get a matrix M with dimensions $d \times m$. The d -dimensional column vectors in matrix M are known as eigenfaces. Figure 4.6 shows the images corresponding to the first five eigenfaces.

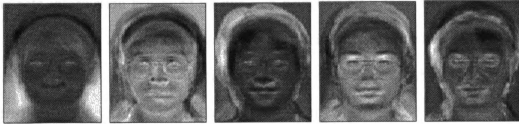


Figure 4.6: First five eigenfaces.

- Project each training image in the eigenspace using Eq. (4.5). Here Y_i represents the projection of the i th training image in the eigenspace, M is the $m \times n$ matrix containing the first m eigenvectors, and X_i is the vector obtained after subtracting the average face from the i th training sample.

$$Y_i = M^T X_i, \quad i=1, 2, \dots, N \quad (4.5)$$

- Project the test image to the eigenspace in the same way as the training image. The mean vector is subtracted from the test vector, which is projected in the eigenspace by computing its product with matrix M^T . The resulting vector has a dimensionality m . Classification is done by computing the Euclidean distance between the projection of the test image and each of the training images. Test image is assigned to the class that is at the minimum distance.

4.5 Training of the System

The training process of our system can be sub-grouped in the following two stages; (i) preprocessing stage, and (ii) PCA algorithm. The manual approach to crop out training face images and feeding them to the learning module was prone to recognition errors. This is because in the test phase, cropping of face images for recognition is based on the coordinate information generated by our face detection algorithm, which may locate a face in a different way than the manual approach. This fact is demonstrated in Figure 4.7, where the face images of a few subjects have been cropped out with different sizes and different face centers under varying ambient light, orientation, and distance from camera to the subject. This behavior is due to the fact that texture information changes with skin complexion and subject illumination. The best way to reduce the effect of such a

behavior was to prepare a training face database by applying the detection algorithm on the training images for image cropping. After preparing the training database, the system was trained using the standard PCA algorithm. Our training set consisted of 10 subjects with 25 images per subject for a total of 250 images. Details of the image database can be found in Chapter 5.



Figure 4.7: Cropping inconsistencies resulting from changes in lighting, orientation, and distance between the camera and the subjects.

4.6 Testing Phase

Steps involved in the testing phase can be summarized as follows:

- Test image is searched for faces by the face detection algorithm.
- Coordinate information regarding the position and size of faces found during the detection process is passed onto the image cropping procedure.
- Face images are cropped out after re-computing coordinates such that the entire face is captured in the cropped portion.
- Retrieved images are scaled to a standard size of 40x50 pixels, a size that matches the face database stored during the learning phase.
- Projection of the test face image is computed in the eigenspace and its minimum Euclidean distance from a training class is found.
- If the distance is below a threshold, the image is identified as known; otherwise it is regarded as an unknown image.
- The above process is repeated until all the faces found by the detection algorithm have been classified.

4.7 Summary

The proposed automatic face detection and recognition system has been implemented in two separate modules. One module prepares the face image database for training of the PCA algorithm using face detection and preprocessing like cropping and scaling. The second module uses the training images for learning and prepares a PCA database of known faces. During recognition, the test image is processed to locate the faces and then after cropping and scaling these images are used by the PCA algorithm for classification.

Chapter 5

Experimental Results and Analysis

5.1 Image Database

The image database used for the purpose of performance evaluation was collected in the PRIP lab of MSU. The requirement was to capture the images of the subjects in small groups of 2 to 4 persons, as they may appear in front of a camera under normal operating situations. A total of 300 group images of 20 different subjects were collected. An effort was made to include subjects from all age groups in the data set. Our data set comprised of gray scale images of size 640x480 pixels. The subjects included persons with facial hair and glasses and of different races.

5.1.1 Imaging Environment

Efforts were made to acquire the images under as realistic an environment as possible. The only constraints exercised were to have the images of subjects with upright faces and some limitations on the distance of the subjects from the camera. The constraint of upright faces was aimed at simplifying the complex recognition task and to ensure the best performance of the face-location algorithm, which was trained to find upright faces.

The distance between the camera and the subjects was maintained within limits of 5 to 10 feet, which is generally the normal viewing distance of a viewer from TV. Imaging background was not controlled and images were collected with different backgrounds to evaluate the performance. Similarly, illuminating conditions were also varied. Orientation of faces was not restricted to only the frontal views and our image set contained a wide viewpoint changes. To allow even more variations, images of most of the subjects were collected in two sessions with a time gap of up to two weeks. Images shown in Figure 5.1 illustrate some of the images collected for this experiment.

5.1.2 Training Set

Group images were processed to prepare an image database covering only the face area of those subjects for whom we intended to train our recognition algorithm. One way to prepare such training set was to crop out faces from the group images manually selecting the image bounds. Such a strategy, which appears to be reasonable, could lead to serious problems during the recognition phase. The reason is that the face location algorithm, which is used to find faces appearing in a test image during the test phase, does not locate the faces as we do. Factors like face orientation, complexion, direction and intensity of the light illuminating the subject, and presence or absence of glasses affects the size and the position of a face detected by the locating algorithm. It was, therefore, found to be more appropriate to crop out face images from the group images, using the face location algorithm instead of performing this task manually. In this way, it was possible to incorporate the behavior of the face location procedure in the training images. All the face images so collected were scaled up/down to a standard size of 40x50 pixels to have a single PCA training image database. We tested the recognition performance of our



Figure 5.1: Variations of light, orientation, and background in the images collected for the experiments.

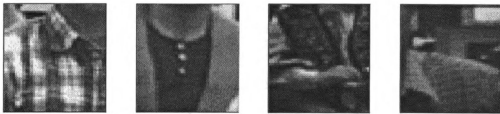
system for different sizes of the training set. These training sets had face images of 10 subjects; first set consisted of 15 images of each subject, the second set 20 images per subject, and the third set 25 training images per subject.

5.1.3 Test Set

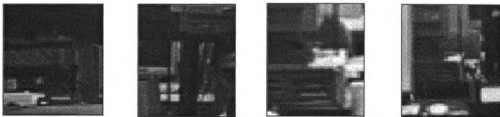
Images used for the test phase were not included in the training set. We collected a total of 510 images of 10 subjects, out of which training sets of three sizes were drawn and the remaining images were used to test the recognition performance. In addition to these face images, we tested the system for face images of 10 additional subjects who were not included in the training set. In other words, these faces were unknown to the system. This test set of the so called “impostors” consisted of 140 face images. Since our aim was to test the system under realistic environments, which are expected to prevail in a typical home TV lounge, we did not try to eliminate the complex background at the time of data collection. A natural effect of cluttered background is the large number of false alarms generated by the face detection algorithm. Obviously, such false alarms are passed on to the recognition algorithm as face images for recognition during the test phase. It was, therefore, important to investigate the system behavior on such non-face images. We collected a set of 130 such images making an effort to have a representative of all types of these false alarms resulting from our entire image set. To summarize, our test image set had the following three categories: (1) 260-360 face images of known subjects, depending upon the size of the training set, (2) 140 face images of 10 unknown subjects (impostors), and (3) 130 non-face images resulting from false alarms of the face detection algorithm.

5.2 Experimental Setup and Decision Strategy

Two important face recognition strategies can be found in the pattern recognition literature. First and more popular method is to find the best match in the training set for a given test face image and classify the test face as the best or first match. Most of the face recognition algorithms follow this decision methodology because the representation of a test face is always included in the training set. Recognition results are reported for correct or incorrect first match found by the algorithm. There is no option in this method to reject a test face image and all the test images are classified. This decision technique was not suitable in the solution of our problem. There is always a possibility that an unknown face may appear before the camera for which the images were not included in the training set. Secondly, a false alarm generated by the face location algorithm is also bound to be classified as one of the training faces when passed on to the recognition algorithm for classification. There may be a possibility to include the images of all the “unknown” persons under one category and train the system to classify them as unknown based on the first match. There is, however, no way to ensure an adequate representation of non-face objects which may be a potential source of false alarms. Such image patterns can result from background for which one has a reasonable chance to collect representative samples of false alarms to be included in the training set provided the background was fixed. However, a very important source of these false alarm is the print patterns on the fabric worn by the subjects. Obviously, collecting a representative data for these types of visual patterns is not possible. Examples of few non-face images detected by the system are shown in Figure 5.2 to highlight this point.



(a) False alarms generated due to fabric print patterns.



(b) False alarms generated due to objects in the background.

Figure 5.2: Examples of non-face images accepted by the face location algorithm.

The second recognition strategy is based upon some kind of a confidence measure. The technique is to classify an object only if it meets some pre-selected threshold values, otherwise it is rejected. All the images, selected for classification, are assigned a class label by the classifier which could be correct or incorrect. While working with PCA methodology, the distance computed between the test image and its first match from the training set can serve as a measure of confidence. A schematic layout of this decision mechanism is shown in Figure 5.3. We adopted this scheme for image classification which requires an accurate selection of a threshold value, for rejecting impostors and non-face objects.

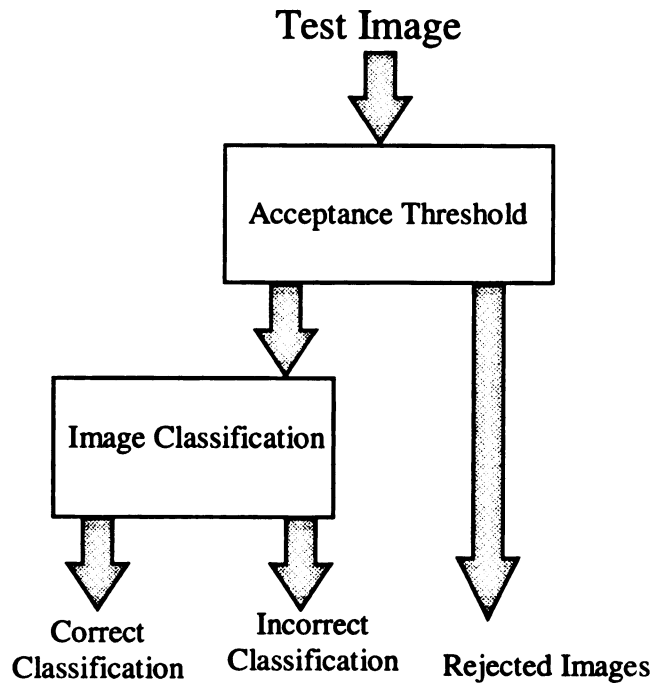
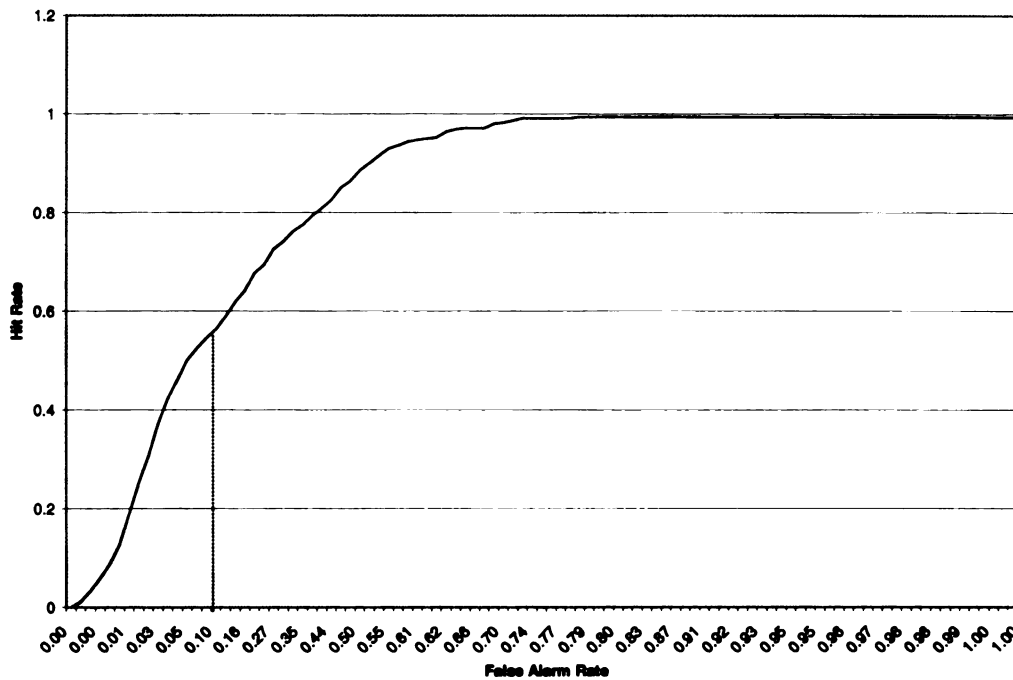
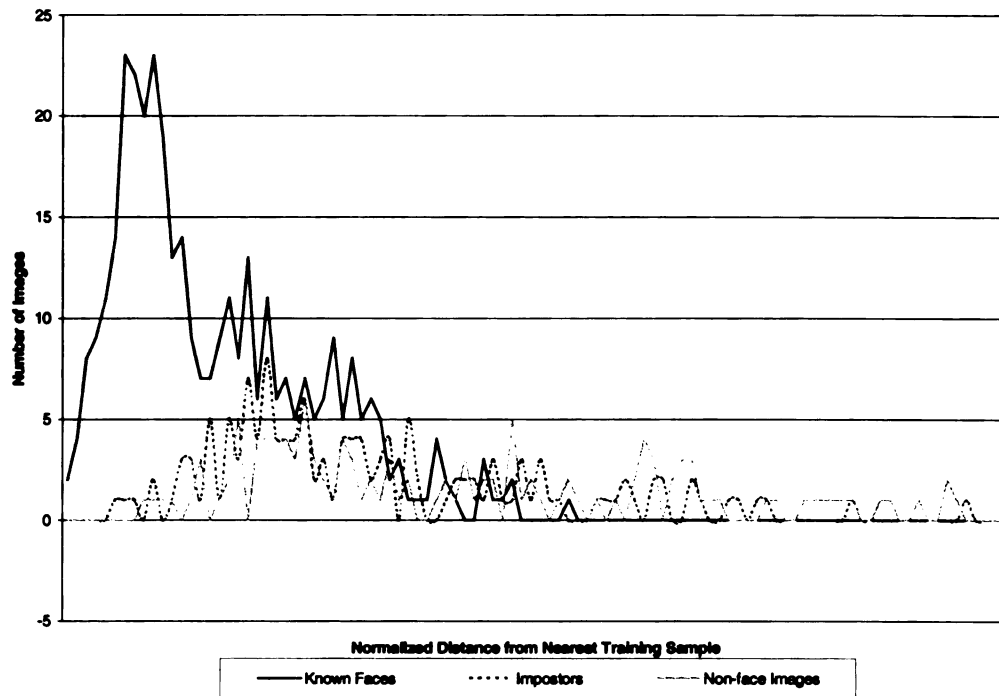
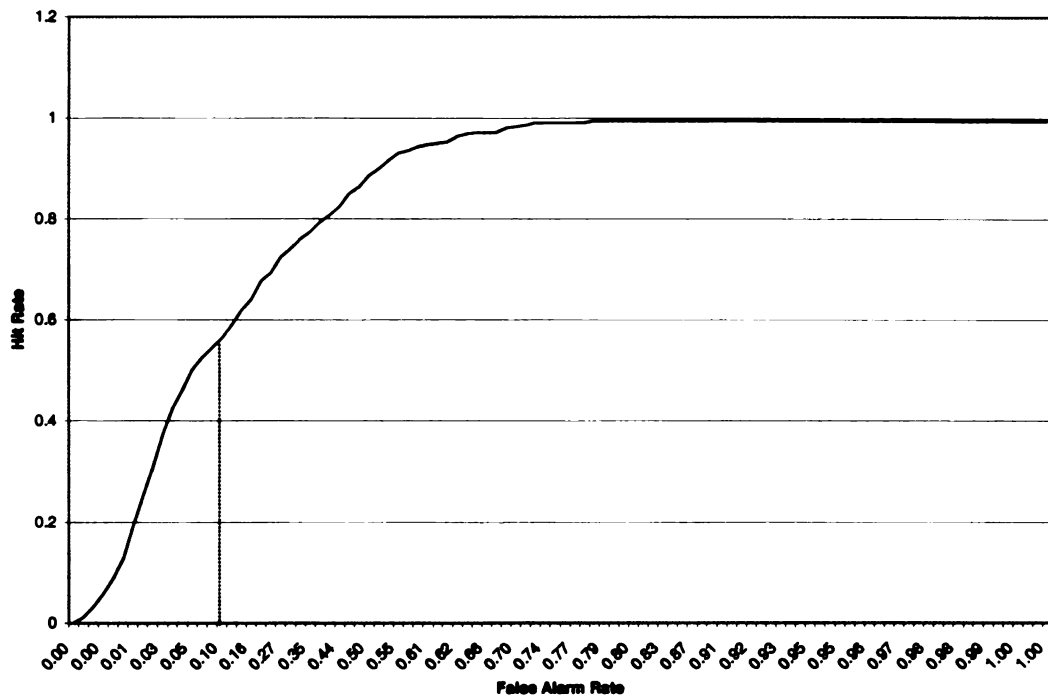
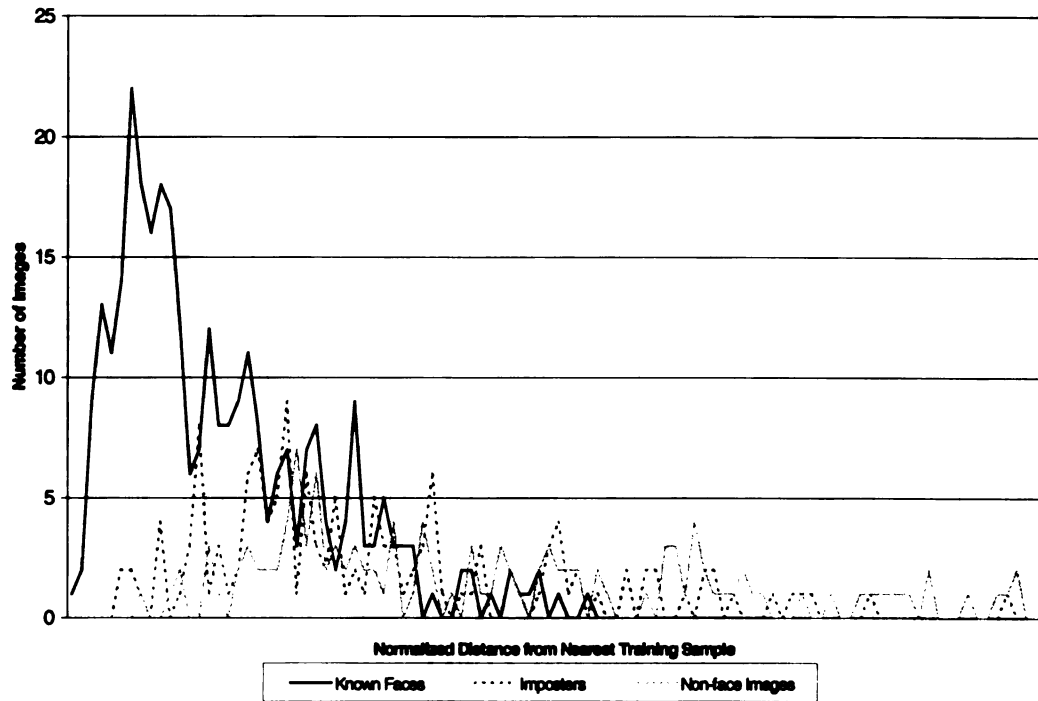


Figure 5.3: Decision methodology based on acceptance threshold.

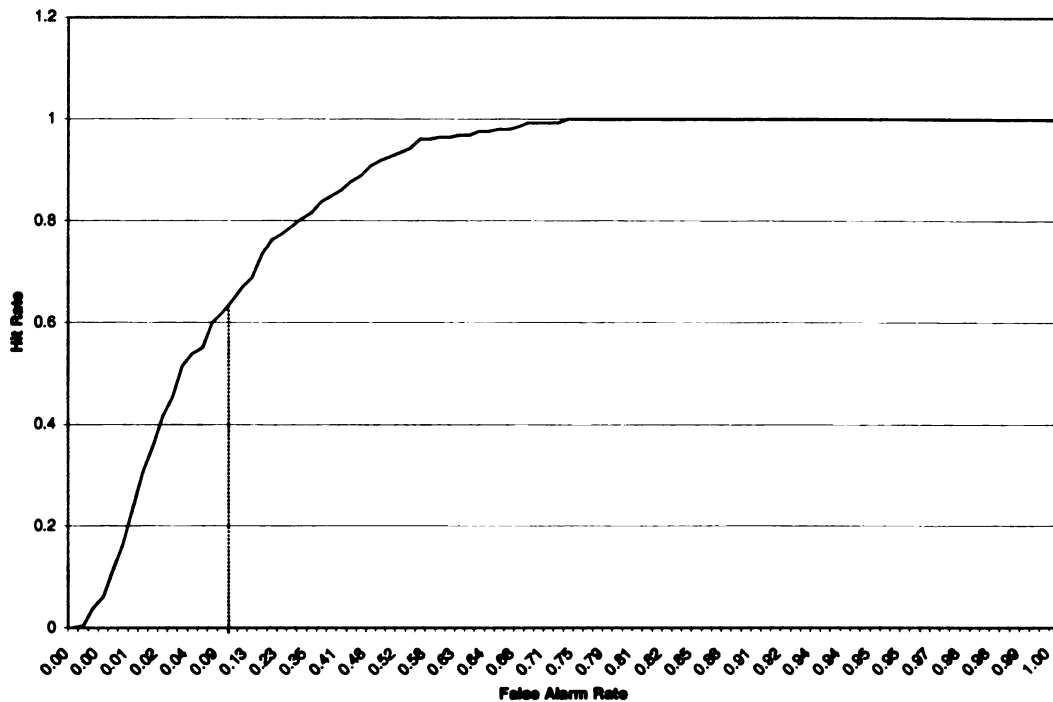
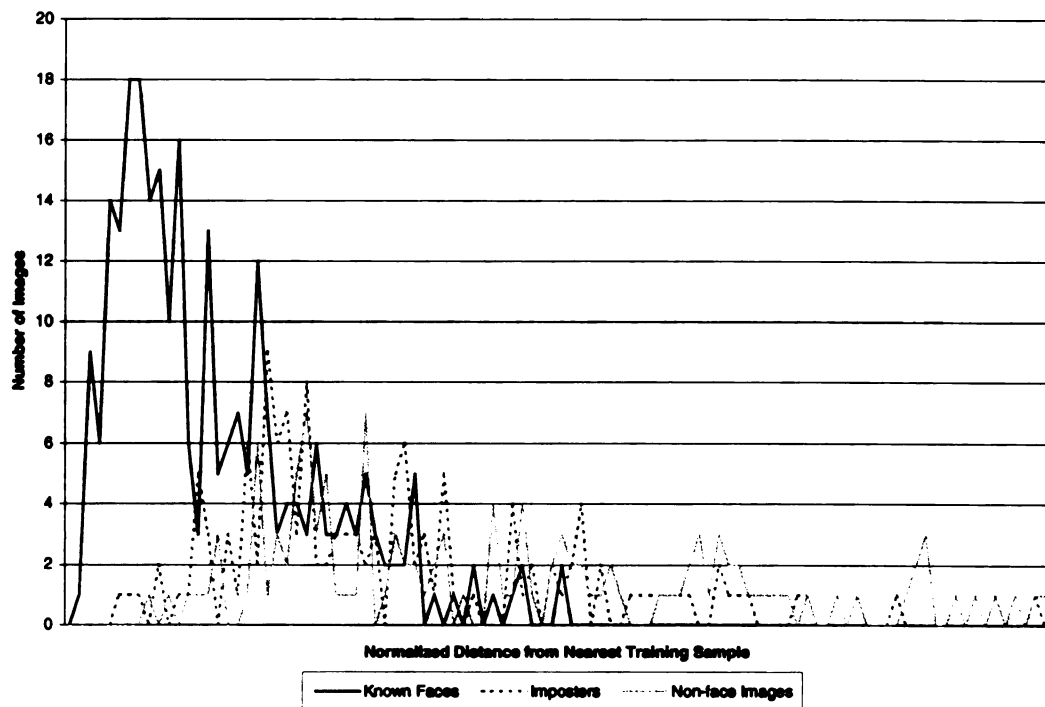
We have tested the recognition algorithm on test images belonging to all the three categories by finding its nearest training sample and distance to the nearest training sample. This data was plotted in the form of histograms depicting distribution of test samples from each category. Figures 5.4 (a) through (c) show these histograms and the Receiver Operating Characteristic (ROC) curves for training set sizes of 15, 20, and 25 training images per subject, respectively. It is evident from these plots that there exists a large overlap among the known faces, impostors and non-face images. An appropriate threshold was selected for each training set from the ROC curves and incorporated in the recognition algorithm to achieve the best possible performance. Results described in the following section are based upon the thresholds selected for each training set.



(a) Recognition histograms and ROC curve for 15 training images per subject.



(b) Recognition histograms and ROC curve for 20 training images per subject.



(c) Recognition histograms and ROC curve for 25 training images per subject.

Figure 5.4: Histograms and Receiver Operating Characteristic (ROC) curves for various sizes of training set depicting recognition behavior.

5.3 System Performance

In the normal operating environments, being considered for our recognition system, we expect that a large number of images will be available as the subjects are watching the TV program. Depending upon the frame grabber, one can expect to have up to 30 frames per second from the output of the camera. Thereafter, the overall system throughput will solely depend upon the processing speed of face location and recognition algorithms. Running at any given frame rate, it is expected that the camera will capture face images with all possible orientations, expressions and illumination. All these factors dictate that in order to achieve a reasonable system performance, we must accommodate such variations in the training set. So, it is appropriate to start with a large training set right from the outset. Performance of the system was evaluated for various sizes of the training database. Results recorded for various sizes of the training set are given in the following subsection.

5.3.1 Recognition Performance for Various Sizes of Training Set

Our first training set had 15 training images for each of the 10 subjects representing various age groups. Second and third sets had 20 and 25 training samples per subject, respectively. Recognition results mentioned in Table 5.1 are based on the first match. These results are based on only those subjects which had a representation in the training set. We used 510 images for this experiment, so our test set consisted of 260-360 images depending upon the number of images used for training.

The other set of results, presented in tables 5.2 through 5.4, is based upon recognition performance after incorporating the rejection option. Same images of the so called known

subjects were in the test set along with two other categories, i.e., impostors and non-faces.

Number of Training Images per Subject	Number of Test Images	Correct First Matches	Incorrect First Matches	Percentage Correct Recognition
15	360	263	97	73.0%
20	310	239	71	77.1%
25	260	219	41	84.2%

Table 5.1: Recognition results for 10 subjects based upon the first match.

A total of 140 images of 10 subjects, labeled as impostors, as none of their images were included in the training set and 130 images of false alarms generated by the face location algorithm, were labeled as non-faces. The test image set was large enough to evaluate the algorithm under various possible situations. We achieved a correct recognition rate of 92% using 250 images for training (25 per subject), after rejecting those images which were not suitable for classification. Rejection rate in tables 5.2-5.4 depicts the percentage of images which were not accepted for classification and rejected altogether. Reasons for rejection are analyzed in Section 5.4.1. Images shown in Figure 5.5 illustrate the face detection and recognition performance of the algorithm on multiple subjects. The face detection algorithm takes up to 90 seconds to process one such image on a Sun Ultra-1 workstation and recognition computations take 0.28 seconds of CPU time for each face with a training set of 250 images.

5.4 Analysis of the Results

Results of the experiments show good overall system performance with a high degree of accuracy in classifying impostors and non-face images. By increasing the size of the training set, recognition performance improved significantly. This improvement can be attributed to the fact that due to relatively uncontrolled imaging environments and inconsistencies in cropping faces from original images, training and test images had lot of variations. Increasing the training set resulted in a more comprehensive representation of each subject. An interesting point to note here is that the error rate for impostors and non-faces remained almost stable with a slight performance improvement. Apparently one may expect performance deterioration on such images as the training set is increased because increased variations in the training set can result in misleading distributions and outliers can reduce the distances between learnt classes and unknown probes. This recognition consistency resulted due to the fact that increased training set caused an overall reduction in the distance between known test face images and training samples due to large training data, so no significant change in the recognition performance resulted for impostors and non-faces. A comparison of the plots shown in Figure 5.4 clearly supports this conclusion.

5.4.1 Rejection Rate

As shown in tables 5.2 through 5.4, about half of the images that belong to known subjects were rejected during classification. This behavior was due to the fact that the distributions of known faces and images from two other classes had a large overlap as

shown by the histograms. Those images of known subjects, which were at larger distances (in the feature space) from the training samples, contributed to this rejection.

Test Image Class (No. of Images)	Correct Recognition	Error Rate	Rejection Rate
Known Faces (360)	83.0%	17.0%	50.9%
Impostors (140)	-	6.4%	93.6%
Non-faces (130)	-	3.1%	96.9%

Table 5.2: Recognition results for 15 training images per subject with reject option.

Test Image Class (No. of Images)	Correct Recognition	Error Rate	Rejection Rate
Known Faces (310)	86.1%	13.9%	48.7%
Impostors (140)	-	9.3%	90.7%
Non-faces (130)	-	3.1%	96.9%

Table 5.3: Recognition results for 20 training images per subject with reject option.

Test Image Class (No. of Images)	Correct Recognition	Error Rate	Rejection Rate
Known Faces (260)	92.0%	8.0%	46.7%
Impostors (140)	-	5.0%	95.0%
Non-faces (130)	-	2.3%	97.7%

Table 5.4: Recognition results for 25 training images per subject with reject option.



Legend

1. Non-face accepted and classified as a face.
2. Known faces correctly classified.
3. Correctly classified non-faces.
4. Rejected known faces.
5. Incorrectly classified impostor.
6. Correctly classified impostors.

Figure 5.5: Examples of the images processed by the system. Single boxes depict the faces detected by the face detection algorithm and double boxes show recognized faces.

Few examples of such rejected images are shown in Figure 5.6 to illustrate variations in the images due to large orientation changes, unwanted contribution of background, and face location inconsistencies that caused rejection. Although it may be possible to reduce the reject rate by changing the acceptance threshold, but this will then increase the incorrect classification of impostors and non-faces as well. On the other hand, as we expect to get multiple images of the scene each second, it will not be a serious problem to reject about half of these images and classify the rest of them with a high accuracy. The performance will solely depend upon the speed at which the images, captured by the camera, are processed.



Figure 5.6: Examples of test images of known faces rejected by the recognition algorithm.

5.4.2 Effects of Background Masking

In order to get rid of the unwanted effects of cluttered background, two methods were considered. One option was to multiply the image with a Gaussian window centered at the area of interest in the cropped image, thereby retaining most of the useful area in the resulting image and fading out the areas away from the center. Although this process does not eliminate the effect of the background completely, it does reduce its unwanted

contribution. The second and a more direct approach is to use a binary oval template which masks the background and eliminates its effect completely, without modifying the area of the image which is useful for classification. We tried the later approach to get rid of the undesired background. Images resulting from this masking procedure are shown in Figure 5.7.

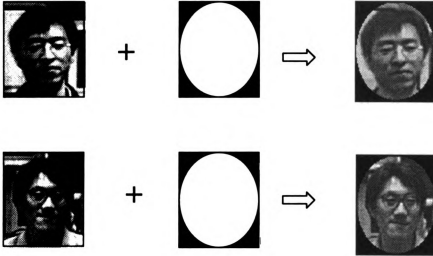


Figure 5.7: Masking scheme to reduce unwanted effects of cluttered background.

The algorithm was trained using 25 images of each of the 10 subjects and then tested on the remaining images in the data set. Results summarized in Table 5.5 show that the overall rejection rate of known faces reduced by about 4% as compared to the results obtained for unmasked images for the same size of training set (Table 5.4), but on the other hand the recognition performance suffered a bit. Similarly, more number of images from the categories of impostors and non-faces were also incorrectly classified. These results can be explained based upon the characteristics of the image database. The similarity between the images belonging to different categories increases with masking as the masked area of all the images becomes the same. The behavior of the cropping algorithm is affected by orientation, illumination, and complexion of the subject,

therefore it is not necessary that only the background is deleted by the mask, rather some useful image information is also lost in the process. As a result, the recognition rate drops and additional impostors and non-faces are accepted as the images of known subjects. The best option was to select the cropping size such that only a minimum portion of the background was included in the resulting image. This selection, however, was based on experience and is specific to the image database used in these experiments.

Test Image Class (No. of Images)	Correct Recognition	Error Rate	Rejection Rate
Known Faces (260)	89.4%	10.6%	42.9%
Impostors (140)	-	16.4%	83.6%
Non-faces (130)	-	5.4%	94.6%

Table 5.5: Recognition results for 25 training images per subject after masking the background.

5.5 Summary

The experimental results show good performance of the proposed automatic face recognition system on our image database. The image database, comprising of 300 group images of 20 subjects, was acquired in an imaging environment which is expected in a typical system application. Our results also demonstrate the requirement of a large training database to achieve an acceptable recognition performance.

Chapter 6

Conclusions

6.1 Summary

In this thesis we have addressed the problem of automatic face recognition under relatively unconstrained imaging environments. The motivation of the work was to implement and test an automatic face recognition system in as realistic conditions as possible. Subjects could appear before the camera in groups as casually as one sits in front of a home TV set. The system can detect the faces appearing in an image, extract them, and recognize by comparing against a face image database learned during the course of training. The algorithm used for face detection works on the principle of texture analysis which was adopted from [16] whereas face recognition was based on PCA [12]. The image database used for training and evaluation of the system performance was collected under minimum possible constraints to ensure realistic operating environments. The database consisted of 300 images of 20 subjects who appear as a group of 3 to 4 persons in each image. The system was trained to recognize 10 subjects using a training database of 250 face images which were cropped out from the group images. Test image database consisted of 530 cropped face images comprising of the following three

categories: (1) 260 images of those 10 subjects for which the system was trained, (2) 140 images of those 10 subjects for which the system was not trained, and (3) 130 non-face images or false alarms produced by the face detection algorithm. Our results show that the system was able to classify the last two categories with an accuracy of over 95%. Images of the subjects from the first category were recognized with an accuracy of 92% after rejecting about 47% of the images which were found unsuitable for classification. We attribute this high rejection rate to the cropping inconsistencies resulting from changes in face orientation, subject illumination, and undesired effects of cluttered background.

6.2 Future Research

There are a number of research issues which need attention in future work. The overall system performance can be improved by working on the weaker links observed during the experiments. These areas are in both the stages of the system, i.e., face location and recognition. Speed and accuracy of the face location algorithm need more attention. One possible approach could be to work with neural network based classifiers, which can reduce the processing time [5]. Detection accuracy can be improved by a fine feature search in the image areas classified as containing a face image. Another methodology could be to subtract the background using motion tracking and discard the background from the image before processing it for face detection. In this way, we can get rid of those false alarms which are caused by the objects in the background and reduce the processing time required by the face location algorithm because plain background will be rejected by the first classification stage of the face location procedure. Background subtraction can also improve the recognition performance by filtering out background

which is an important reason for misclassification. Algorithms developed to track facial features like eyes can also be used to find the exact face position in an image if color images are used instead of gray scale images.

Bibliography

- [1] Alan Bloom, *The Republic of Plato*. Basic Books, New York, 1991.
- [2] Robin Waterfield and David Bastock, *Aristotle*. Oxford University Press, UK, 1996.
- [3] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern classification and scene analysis*. 2nd edition John Wiley & Sons Inc., 1998 (to be published).
- [4] Rama Chellappa, Charles L. Wilson, and Saad A. Sirohey, "Human and machine recognition of faces: A survey", *Proc. of IEEE*, vol. 83, no. 5, May 1995, pp. 705-740.
- [5] Henry A. Rowley, Shumeet Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Transactions on PAMI*, vol. 20, no. 1, Jan. 1998, pp. 23-37.
- [6] M. D. Kelly, "Visual identification of people by computers", Tech. Rep AI-130 Stanford AI Project, Stanford, CA, 1970.
- [7] A. Yuille, D. Cohen, and P. Hallinan, "Feature extraction from faces using deformable templates", *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 1989, pp. 104-109.
- [8] I. Craw, H. Ellis, and J. Lishman, "Automatic extraction of face features", *Pattern Recognition Lett.*, vol. 5, 1987, pp. 183-187.
- [9] P. J. Burt, "Multiresolution techniques for image representation, analysis, and 'smart' transmission", *SPIE Proc.: Visual Comm. and Image Process.*, vol. 1199, 1989, pp. 2-15.
- [10] J. W. Shepherd, "An interactive computer system for retrieving faces", in H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, (eds.), *Aspects of Face Processing*, Dordrecht, Nijhoff, 1985, pp. 398-409.
- [11] Saad A. Sirohey, "Human face segmentation and identification", Technical Report CAR-TR-695, Center for Autom. Res., Univ. of Maryland, College Park, 1993.

- [12] M. Turk and A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive and Neuroscience*, vol. 3, no.1, 1991, pp. 71-86.
- [13] L. Sirovich and M. Kirby, "Low-dimensional procedure for characterization of human face", *J. Opt. Soc. Amer.*, vol. 4, 1987, pp. 519-524.
- [14] Daniel L. Swets and John J. Weng. "Using discriminant eigenfeatures for image retrieval", *IEEE Transactions on PAMI*, vol. 18, no. 8, Aug. 1996, pp. 831-836.
- [15] P. Jonathon Phillips, Hyeonjoon Moon, Patrick Rauss, and Syed A. Rizvi, "The FERET September 1996 database and evaluation procedure", *Intl. Conf. on AV based Biometric Person Identification, Springer, LNCS 1997*, pp. 395-402.
- [16] Nicolae Duta and Anil K. Jain, "Learning the human face concept from black and white pictures", *Proc. of ICPR*, Aug., 1998.
- [17] T. Sakai, M. Nagao, and S. Fujibayashi, "Line extraction and pattern recognition in a photograph", *Pattern Recognition*, vol. 1, 1969, pp. 233-248.
- [18] Shang Hung Lin, Yin Chan, and S. Y. Kung, "A probabilistic decision-based neural network for locating deformable objects and it's application to surveillance system and video browsing", *Intl. Conf. On Acoustics, Speech, and Signal Processing, Atlanta, GA, 1996*, pp. 3,554-3,557.
- [19] Kah-Kay Sung and T. Poggio, "Example-based learning for view-based human face detection" *IEEE Transactions on PAMI*, vol. 20, no.1, Jan. 1998, pp. 39-51.
- [20] M. Bichsel and A. Pentland, "Human face recognition and face image set's topology", *Computer Vision, Graphics and Image Processing: Image Understanding*, vol. 59, 1994, pp. 254-261.
- [21] V. Govindaraju, S. N. Srihari, and D. B. Sher, "A computational model for face location" *Proc. of 3rd Intl. Conf. on Computer Vision*, 1990, pp. 718-721.
- [22] T. Kanade, "Computer recognition of human faces", *Basel and Stuttgart: Birkhauser*, 1977.
- [23] Jun Zhang, Yong Yan, and Martin Lades, "Face recognition: Eigenface, elastic matching, and neural nets", *Proc. of IEEE*, vol. 85, no. 9, Sep. 1997, pp. 1,423-1,435.
- [24] R. Brunelli and T. Poggio, "Face recognition: Features vs. templates", *IEEE Trans. on PAMI*, vol. 15, no. 10, Oct. 1993, pp. 1,042-1,053.

- [25] Steve Lawrence, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back, "Face recognition: a convolutional neural net approach", IEEE Trans. On Neural Networks, vol. 8, no. 1, 1997, pp. 98-113.
- [26] H. Bourland and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition", Biological Cybern., vol. 59, 1988, pp. 291-294.
- [27] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection" IEEE Trans. on PAMI, vol. 19, no.7, Jul. 1997, pp. 711-720.
- [28] Baback Moghaddam, Chahab Nastar, and Alex Pentland, "A bayesian similarity measure for direct image matching", Intl. Conf. on Pattern Recognition, Vienna, Austria, 1996, vol.2, pp. 350-358.
- [29] John J. Weng, "Cresception and SHOSLIF: Towards comprehensive visual learning", in S. K. Nayar, and T. Poggio (eds.), *Early visual learning*, Oxford University Press, New York, 1996, pp. 183-214.

MICHIGAN STATE UNIV. LIBRARIES



31293017014733