HUMAN FACE RECOGNITION BASED ON FRACTAL IMAGE CODING

TEEWOON TAN **B.Sc., B.E.**

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

School of Electrical and Information Engineering
The University of Sydney

August 2003

Abstract

Human face recognition is an important area in the field of biometrics. It has been an active area of research for several decades, but still remains a challenging problem because of the complexity of the human face. In this thesis we describe fully automatic solutions that can locate faces and then perform identification and verification. We present a solution for face localisation using eye locations. We derive an efficient representation for the decision hyperplane of linear and nonlinear Support Vector Machines (SVMs). For this we introduce the novel concept of ρ and η prototypes. The standard formulation for the decision hyperplane is reformulated and expressed in terms of the two prototypes. Different kernels are treated separately to achieve further classification efficiency and to facilitate its adaptation to operate with the fast Fourier transform to achieve fast eye detection. Using the eye locations, we extract and normalise the face for size and in-plane rotations. Our method produces a more efficient representation of the SVM decision hyperplane than the well-known reduced set methods. As a result, our eye detection subsystem is faster and more accurate.

The use of fractals and fractal image coding for object recognition has been proposed and used by others. Fractal codes have been used as features for recognition, but we need to take into account the distance between codes, and to ensure the continuity of the parameters of the code. We use a method based on fractal image coding for recognition, which we call the Fractal Neighbour Distance (FND). The FND relies on the Euclidean metric and the uniqueness of the attractor of a fractal code. An advantage of using the FND over fractal codes as features is that we do not have

to worry about the uniqueness of, and distance between, codes. We only require the uniqueness of the attractor, which is already an implied property of a properly generated fractal code.

Similar methods to the FND have been proposed by others, but what distinguishes our work from the rest is that we investigate the FND in greater detail and use our findings to improve the recognition rate. Our investigations reveal that the FND has some inherent invariance to translation, scale, rotation and changes to illumination. These invariances are image dependent and are affected by fractal encoding parameters. The parameters that have the greatest effect on recognition accuracy are the contrast scaling factor, luminance shift factor and the type of range block partitioning. The contrast scaling factor affect the convergence and eventual convergence rate of a fractal decoding process. We propose a novel method of controlling the convergence rate by altering the contrast scaling factor in a controlled manner, which has not been possible before. This helped us improve the recognition rate because under certain conditions better results are achievable from using a slower rate of convergence. We also investigate the effects of varying the luminance shift factor, and examine three different types of range block partitioning schemes. They are Quad-tree, HV and uniform partitioning. We performed experiments using various face datasets, and the results show that our method indeed performs better than many accepted methods such as eigenfaces. The experiments also show that the FND based classifier increases the separation between classes.

The standard FND is further improved by incorporating the use of localised weights. A local search algorithm is introduced to find a best matching local feature using this locally weighted FND. The scores from a set of these locally weighted FND operations are then combined to obtain a global score, which is used as a measure of the similarity between two face images. Each local FND operation possesses the distortion invariant properties described above. Combined with the search procedure, the method has the potential to be invariant to a larger class of non-linear distortions. We also present a set of locally weighted FNDs that concentrate around the upper part of the face encompassing the eyes and nose. This design was motivated

by the fact that the region around the eyes has more information for discrimination. Better performance is achieved by using different sets of weights for identification and verification. For facial verification, performance is further improved by using normalised scores and client specific thresholding. In this case, our results are competitive with current state-of-the-art methods, and in some cases outperform all those to which they were compared. For facial identification, under some conditions the weighted FND performs better than the standard FND. However, the weighted FND still has its short comings when some datasets are used, where its performance is not much better than the standard FND. To alleviate this problem we introduce a voting scheme that operates with normalised versions of the weighted FND. Although there are no improvements at lower matching ranks using this method, there are significant improvements for larger matching ranks.

Our methods offer advantages over some well-accepted approaches such as eigenfaces, neural networks and those that use statistical learning theory. Some of the advantages are: new faces can be enrolled without re-training involving the whole database; faces can be removed from the database without the need for re-training; there are inherent invariances to face distortions; it is relatively simple to implement; and it is not model-based so there are no model parameters that need to be tweaked.

Acknowledgements

I would like to thank Professor Hong Yan for supervising the research conducted in this thesis. His help and advice has been tremendously valuable, and I am deeply grateful and appreciative of all he has done for me.

I also wish to thank my colleagues in the Laboratory of Imaging Science and Engineering, namely Ju Jia Zou, Philip Mitchell, Qin Zhi Zhang, Yi Xiao, Dr. Chaminda Weerasinghe, Dr. Mustafa Sakalli and Associate Professor Alan Fu for their help and useful discussions on material directly and indirectly related to my research. I am very grateful to Inge Rogers for proofreading this thesis and for providing many valuable corrections.

I would also like to acknowledge and reciprocate the love, support and encouragement from my father (Eow Hooi Tan), mother (Yoon Ping Chin), brother (Tee Han Tan), sister-in-law (Vivian Tung) and other family members. To Geoffrey and Shirley, PFITA! My love and gratitude also goes to my partner, Jeanette Lee, for her constant love and support.

The research on which this thesis is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in; K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre; "XM2VTSbd: The Extended M2VTS Database, Proceedings 2nd Conference on Audio and Video-based Biometric Personal Verification (AVBPA99)" Springer Verlag, New York, 1999. CVSSP URL: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb".

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Using a computer to perform the task of human face recognition automatically is an active area of research, and has been for several decades. The aim is to enable a computer to recognise human faces from digitised images or video sequences. The human face is a form of biometric. Biometrics are a set of measurable physiological and/or behavioral characteristic properties of the human body that can be used to infer a person's identity [PBJ00, PC00, PMWP00, FD00, NCS+00, Way00]. Due to recent events there has been increasing interest in using biometrics to improve security in public places and to provide protected access to physical areas and resources. Applications of biometrics generally fall into two categories, identification and verification. The literature also refers to these as recognition and authentication respectively [TKP01]. We will use the former convention and the term recognition is used to refer to both identification and verification.

Verification is the task of verifying a person's identity given their biometric and their claimed identity. The claimed identity is the assumed identity of the person, which may be correct or incorrect. A person may present an incorrect identity in an attempt to gain access as another person. A good biometric system will compare

that claimed identity with the identity obtained from a biometric of that person and then reject that claim, preventing access for that person. The performance of a verification system is usually measured in terms False Accept Rates (FAR) and False Reject Rates (FRR). FAR gives an indication of the percentage of times a person with a false identity can be accepted by the system as a true identity. FRR gives an indication of the percentage of the times a person with a true identity can be rejected by the system as a false identity.

Identification is the task of finding a person's identity given just the biometric data. The unknown identity is compared to a database of known identities and the best, or best n, matches are returned by the system. Additionally, some systems can decide to reject the unknown identity if the matching score does not reach a certain threshold. This would ideally be the result when the true identity of the unknown identity is not in the database. The performance of an identification system is usually presented as the recognition/error rate as a function of the top n matches.

Examples of the use of biometrics include the recognition of fingerprints, hand, voice, signature, iris and face. Of all biometric systems the ones based on human face recognition are one of the least intrusive in that they do not require as much human interaction and participation. In fact, it can perform recognition without the person being aware of it. This is usually the case in surveillance applications. An example of this is the detection of criminals at the airport [Tit02], on the street or in the premises of a retail shop.

The problem of human face recognition remains challenging. The view of the human face through a camera is subject to variations in size, orientation, colour, facial expressions, pose, shadows, texture and lighting conditions. Although there are many commercial systems available and increasingly being used in practical applications, there is a general consensus that recognition accuracy must be further improved [CR02]. As it currently stands, automatic human face recognition in a surveillance environment has limitations. In an environment where there is heavy human traffic even a small identification rate will generate an unacceptable number

of false alarms. A system that has an accuracy rate of 99.0 % will still generate 10000 false alarms for a traffic of 1 million people. This is usually unacceptable in an airport environment where traffic is in the order of millions every year. A trial system at a United States airport was cancelled after it was deemed too inaccurate. Privacy issues have to be addressed and treated in each application of human face recognition. A face recognition system requires the use of cameras, and whenever cameras are installed there is the risk of camera abuse, where it is used for more than just its intended purpose. For surveillance in a public environment there is a risk of losing anonymity.

The accuracy of a face recognition system can never be stated precisely because it is affected by too many factors. It depends on the environment in which it operates, the type of images used in the database and the settings of various system parameters. A result quoted for one environment does not automatically translate to the system achieving the same performance in a different environment. And the applications in which face recognition is put to use operates in environments that are too wide and varied to be able to quote a result that are valid for all of them. This problem can be alleviated if the face recognition algorithms have reached a level where it can achieve an acceptable level of invariance under most environmental conditions, but as yet the technology has not reached that level. Nevertheless, standard testing procedures and images obtained in controlled environments are required to enable the comparison between different systems. There are many standard image databases (e.g. FERET, ORL and XM2VTS) and testing protocols (e.g. FERET and Lausanne Protocol for the XM2VTS database) that can be used to compare the performance of various systems in a controlled environment.

One of most challenging problems facing most human face recognition systems is the altered lighting conditions due to a change in environment. In many cases the environmental condition during the enrolment of faces differs from those during verification or identification. Most, if not all, algorithms cannot handle the change satisfactorily in a practical situation. That said, a face recognition system can still operate successfully, depending on its installation and how it is customised to

the application. Successful customisation requires the identification of weaknesses in the system and taking steps to minimise the effects of those weaknesses. For example, the lighting conditions where the system is installed should be analysed and then modified to maximise accuracy. The lighting conditions at the enrolment and identification/verification points should be matched. Reflections off spectacles, mirrors and glasses, the direction of shadows and light sources all have to be considered. Accuracy can also be improved if subjects are cooperative in removing hats and sunglasses and looking directly into the camera. Success also depends on performance requirements, and must integrate well into the existing security infrastructure. There are limitations in the use of face recognition to detect terrorists, mainly due to the availability of true facial images of those people. Facial images of all but the most wanted terrorists are unavailable. Even if they are available they are mostly taken in uncontrolled environments, so accuracy rates are generally not very good.

Other current uses of face recognition technology include the detection of fraudulent drivers' licenses or ID cards, access control to business premises, logging of work hours, identification of ship passengers and access to computer networks, individual PCs and peripherals.

Although much progress has been made in facial recognition technology, improvements are still necessary for it to be practical in some applications. There are two parts to a basic facial recognition system. The first is the face detection and extraction component. The second is the recognition component. In a practical system other components are necessary, such as a communication subsystem for tasks such as remote administration and monitoring and a separate subsystem that provides database access. The design of the latter also depends on whether the database is centralised or local to each processing station. The work in this thesis concentrates on face detection and recognition components.

1.2 Face Detection and Extraction

One of the first steps of a fully automated face recognition system is to detect and extract the faces in an image. Recognition can only be carried out once the face has been extracted and normalised. If the face database used for experiments requires the faces to be located prior to recognition, then the performance of any face recognition system is dependent, to some extent, on the accuracy of the face detection system. There are many publicly available face databases that can be used for experiments to facilitate the process of comparing algorithms devised by different researchers, for example, the Olivetti Research Laboratories (ORL) face database, extended M2VTS (XM2VTS) database and the FERET database. The ORL face database has faces roughly located without any background, which is why it is still used to test recognition algorithms. The FERET database also has eye locations provided in separate text files. The XM2VTS database requires faces to be located.

The aim of face detection is to locate all faces in a scene if there are any. Many factors affect the performance of a face detection system. There are countless numbers of articles in the literature describing face detection. Only a small sampling of those are described here.

The performance of face detection systems is affected by head size, location, background objects, head orientation with respect to in-plane rotation and out-of-place rotations, lighting, pose, facial expression and occlusions. Lighting has a significant effect on performance. Some algorithms rely on a set of training images and they have to be selected so that they cover different lighting conditions. The problem is obtaining enough images in different lighting conditions and whether the algorithms scale well to a large number of training samples. Some algorithms are built using expert knowledge, so they do not rely on training images, but changes in lighting conditions still have to be accounted for to some degree. Faces also change with eye glasses, hats and beards. There is a wide variety of eye glasses and reflections off light sources picked up by the camera that can pose problems for the face detector.

Hats can cast shadows and beards and moustaches can change the appearance of a face. A human face undergoes a large amount of non-linear distortions under different facial expressions. It is difficult to model these accurately without resorting to facial muscle models or similar. Background objects may look similar to a face, or depending on the detection algorithm, a part of the background with no resemblance to a face may be picked up because the algorithm is too general. Faces can also be partially occluded by objects between the face and the camera.

Face detection and face localisation are slightly different problems. The former describes the process of locating all faces in a scene, and the latter assumes that only a single face is present in the scene and the aim is to determine the position of that face. Face localisation is a simpler form of face detection. It should be noted that if any two face feature points are detected then that face can be located and normalised. Therefore, face detection/localisation can also be tackled via facial feature detection. Face tracking is face detection operating on a sequence of images.

A survey of face detection algorithms is given by Hjelmås and Low [HL01]. Yang et al. [YKA02] also present a good review of face detection methods, and classified face detection algorithms for single images into four categories, namely knowledge-based methods, feature invariant approaches, template matching methods and appearance-based methods. Knowledge-based methods use a set of rules developed from what humans know about the appearance of a face. Feature invariant approaches use structural features that are invariant to changes in pose, expression, or illumination. Template matching methods use a set patterns representative of the face, which are then correlated with the input image. Appearance-based methods perform face detection using models or templates learnt from a set of representative training images. Note that some methods can be placed into more than one category. We group the existing methods we are describing into those four categories and a summary is shown in Table 1.1.

Category	**Used by**
Knowledge-based	
Multiresolution	Yang et al. [YH94], Kotropoulos et al. [KP97]
Feature Invariant	
Geometry and Shape	Jesorsky et al. [JKF01], Maio et al. [MM00], Intrator et al. [IRY96], Brunelli et al. [BP93]
Skin Colour	Yilmaz et al. [YS02], [CN98, CB94, CB97], [YA98, YA99]
Template Matching	
Fixed Template	Zhang et al. [ZP96], Sinha [Sin94], Rowley et al. [RBK98b], Brunelli et al. [BP93]
Deformable Template	Yuille et al. [YHC92]
Appearance-based	
Neural Network	Féraud et al. [FBVC01], Lin et al. [LKL97], Rowley et al. [RBK98b]
Principle Components Analysis	Moghaddam et al. [MP97], Martínez [Mar02], Wong et al. [WLS01]
Filtering	Keren et al. [KOG01]
Genetic Algorithm	Wong et al. [WLS01], Lin and Wu [LW99]
Snakes	Lam et al. [LY96]
Wavelets	Kondo and Yan [KY99]
Hidden Markov Model	Samaria et al. [SY94, Sam94], Nefian and Hayes [NH98]
Statistical	Schneiderman and Kanade [SK00] Liu [Liu03]
Support Vector Machines	Osuna et al. [OFG97], Smeraldi et al. [SCB99], Romdhani et al. [RTSB01], Terrillon et al. [TSS+00]

Table 1.1: Summary of existing face detection algorithms.

1.2.1 Overview of Face Detection Algorithms

Féraud et al. [FBVC01] propose a face detector based on neural networks. Their aim is to determine whether a subimage of size 15×20 pixels within an image is a face or a non-face. The detector is composed of four filters. Each filter is applied in turn and they range from the simplest and least accurate but fastest, to the most complex and accurate but slowest. The first filter detects motion and extracts the parts that are moving. This assumes that the majority of motion is from the face. This first step eliminates 90 % of locations and scales that could be faces. The second is a colour filter that extracts skin colour pixels, eliminating 60 % of the hypothesis. A single multilayer perceptron (MLP), also known as a prenetwork [RBK98a] is used in the next filter stage. The network has 300 inputs, 20 hidden neurons and one output corresponding to a face or non-face decision. This network is trained with 8000 front and side views of faces, and 50000 non-face examples. The image samples are histogram equalised, smoothed and subtracted from the average face. This network has a very high false alarm rate and discards more than 93 % of possible face positions, but is usable when combined with the other two filters. The final filter stage used is the constrained generative model (CGM), which determines if a sub-image is a face. In this approach, an input sub-image is projected onto a face space as found using Principal Component Analysis [Jol86]. However, this algorithm has a time complexity that is of $O(n)$ with respect to the number of examples. The authors propose approximating the required operation of projecting an input example onto the set of faces with a neural network. Combinations of CGMs are also explored. The computational cost of the face detection process is reduced by a search algorithm. A grid is placed over the input image and each intersection point of the grid is tested for a face. An exhaustive search is carried out at those intersection points where the output of the system is high enough. Experiments on Test 1 of the CMU database using a search grid of 3 by 3 give a detection rate of 83 %. The algorithm can also detect side-on faces, with an accuracy of up to 23 % at the 90 degree pose angle.

Lin et al. [LKL97] use a probabilistic decision-based neural network (PDBNN). It is similar to a radial basis function network, but learning rules are modified and a probabilistic approach taken. Two feature vectors are obtained from intensity and edge images around the eyes, eyebrows and nose region. These are used by two PDBNNs and the two results are combined by fusion to obtain a decision.

Jesorsky et al. [JKF01] use a shape based approach to detect faces. Edges are obtained from grayscale images and the Hausdorff distance is used as a similarity measure to calculate the distance between possible face images and a general face model. The method operates on still grayscale images. For two finite point sets $\mathcal{A} = \{a_1, \ldots, a_m\}$ and $\mathcal{B} = \{b_1, \ldots, b_m\}$, the detection problem is formulated as

$$d_{\hat{p}} = \min_{p \in \mathcal{P}} H(\mathcal{A}, T_p(\mathcal{B})) \tag{1.1}$$

where the Hausdorff distance H is defined as

$$H(\mathcal{A}, \mathcal{B}) = \max(h(\mathcal{A}, \mathcal{B}), h(\mathcal{B}, \mathcal{A})), \text{ where} \tag{1.2}$$

$$h(\mathcal{A}, \mathcal{B}) = \max_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} \|a - b\| \tag{1.3}$$

where T_p is a transformation such as scaling and translation, and the parameters p are selected from the parameter space $\mathcal{P}$. Instead of using $h(\mathcal{A}, T_p(\mathcal{B}))$ the authors use the box-reverse distance h_{box}, the definition of which can be found in [Ruc96]. The detection process consists of coarse and refinement steps. In the coarse detection step, the Sobel operator is used to extract edges from the input image and a face model is used for localisation. In the refinement step the coarsely detected eye region is resampled and an eye model used to find the eye locations. Both the face and eye models are obtained using the average face image that has been optimised by genetic algorithms. More than 10000 face images are used for this. In their experiment with 1180 images from the XM2VTS database they obtain a face localisation accuracy of 98.4 %. An accuracy rate of 91.8 % is obained on the BIOID database consisting of 1521 images. In both cases they classify a face as being correctly located if the detected eye position is within 25 % of the eye to eye distance to the correct eye position.

Zhang et al. [ZP96] use a template based approach that includes *a priori* constraints. The constraints are the ratio of average intensity, chrominance and smoothness values. Their technique is an extension of the one introduced by Sinha [Sin94]. The method locates a face by identifying invariant relationships between different parts of the face. For example, the average intensity relationship between the cheeks and the eyes are identified to be invariant, to an extent, to different lighting conditions. The Generalised Ratio Template is introduced by Zhang to include other invariant relationships based on other properties such as chroma constraints and texture and frequency constraints. The relationship of average chroma values of different regions are used to further identify valid face regions. The variance is used as a measure of texture and frequency, that is smoothness, of the different regions. The best ratios of the average intensity, chrominance and variance values between the different regions of template are used to find the most likely location of a face. The template covers the forehead, eyes, cheeks and mouth regions of the face. Their experiment on a database of 400 images produces a detection accuracy of 95 %.

Rowley et al. [RBK98b] use a templated based face detector combined with neural networks to achieve rotation invariant detection. There are two networks, a router network and a detection network. The router network is trained with input subwindows of size 20×20 to locate faces and output their rotation angle. The subwindows are preprocessed by histogram equalisation. There are 36 output units, each representing an angle of $i \times 10°$. The network is trained with faces that have been rotated, slightly scaled and translated, giving a total of 15720 examples. Each output is trained to generate the value $\cos(\theta - i \times 10°)$ for a face at angle θ. The estimated angle for an input test face is then given by the average vector

$$\left(\sum_{i=0}^{35} \text{output}_i \times \cos(i \times 10°), \sum_{i=0}^{35} \text{output}_i \times \sin(i \times 10°)\right) \tag{1.4}$$

The router network returns many false positives, but this is filtered out with the detector network. This network is trained to output a value of $+1.0$ for a face input and -1.0 for a non-face input using a bootstrap method adapted from Sung [Sun96]. Experiments on two large databases show that their system achieves a detection rate of 79.6 % with a small number of false positives.

Moghaddam et al. [MP97] present a density estimation technique for target detection, recognition and coding. Central to the idea is the concept of distance-in-feature-space (DIFS) and distance-from-feature-space (DFFS). PCA is used to identify the subspace of face images. The component of a sample $\mathbf{x}$ in this space is referred to as the DIFS. The component of $\mathbf{x}$ in the orthogonal space to this face space is referred to as the DFFS. Derivation of the technique begins with the estimation of high-dimensional Gaussian densities. The likelihood of an input pattern belonging to a certain class is characterised by the Mahalanobis distance. An estimator for this distance is formulated, which is expressed as the summation of two components, one corresponding to the principal subspace and the other to the orthogonal complement. This method is also extended to multimodal densities. The density estimate $\hat{P}(\mathbf{x}|\Omega)$ is used for face detection. Subwindows $\mathbf{x}$ at each spatial position (i, j) is extracted and the density estimate calculated. The position (i, j) that maximises $\hat{P}(\mathbf{x}|\Omega)$ is selected as the most likely face location. Experiments on the MIT Media Laboratory's database of more than 7000 images show that the system has a highest detection rate of 95 %. It outperformed the DFFS and sum-of-squared-differences (SSD) methods.

Martínez [Mar02] describes a set of methods aimed at alleviating the problem of imprecisely localised, partially occluded and expression-variant faces from a single image per class. They use the method described in Moghaddam et al. [MP97] to locate the eyes and mouth of each face. Once located the face is warped so that the eyes, nose and chin are at standard positions. The imprecision of localisation is modeled by a Gaussian distribution. A mixture of Gaussians is also considered. Dimensionality reduction using PCA is used to improve the computational cost of calculating and storing each Gaussian model. An unseen image is classified by searching for the closest Gaussian model in the reduced eigenspace using the Mahalanobis distance. Experiments show that the method improved performance with respect to invariance to localisation errors, which is reflected in their recognition subsystem. To alleviate the problem of occlusion the face image is divided into k local parts. A similar method to localisation is used, where k eigenspaces are

created, each with Gaussian distributions. Experiments on occluded images show an improvement in the recognition rate. It was discovered that the region around the eyes carries more discriminating information than the mouth region, which is reflected in improved results when the mouth is occluded. However, in tests using duplicate images the reverse is true, that is, the mouth region has more discriminating information. To account for facial expressions they note that different parts of the brain are responsible for different emotions. For example, anger and sadness involve the right hemisphere of the brain. The left hemisphere of the brain is more active than the right. Correspondingly, the logical assumption is that the right side of the face will be more expressive for anger and sadness, and the left side of the face will be more expressive for happiness. Therefore, in addition to the probabilistic approach described above, weighting is used for different parts of the face to decrease sensitivity to facial expressions. Experiments show that this weighted eigenspace representation gives better results than the unweighted version.

Keren et al. [KOG01] introduce the concept of antifaces. It is a template-based method where instead of concentrating on detecting objects of interest, non-objects of interest are detected and eliminated, thus the name antiface. A series of simple filters are applied as inner products with an input image. Central to the design of the filters is the observation that "the absolute value of the inner product of two smooth vectors is, on the average, large." Proof is provided based on the Boltzmann distribution. They called the collection of templates that should be used for detection "multitemplates". The filters are designed to yield small values for multitemplates but large values for "random" natural images. Three constraints are used in the design of the filters: the filter must be unit norm; the dot product of the filter with a multitemplate should be small; and the filter should be made as smooth as possible. The filter is found using an optimisation process in the Discrete Cosine Transform (DCT) domain. The next filter is found by imposing an additional condition that the conjugate of the dot product of the first and second filter be equal to zero. Other filters are found in a similar fashion. The resulting filters are applied sequentially, and are more or less independent of each other. Experiments show

that only three to four filters are required to locate faces. Detection under varying illumination is achieved by using the multiplicative nature of the reflectance model of light reflecting off a flat object.

Maio et al. [MM00] describe a face location technique based on directional image information. Their system is composed of two stages. The first stage produces approximate face locations, followed by a stage with a finer search algorithm that performs finer face location and verification at the same time. The first stage calculates the directional image using the method proposed by Donahue et al. [DR93]. Faces in the directional image correspond to elliptical blobs. For this they use the generalised Hough transform [Bal81] with the elliptical annulus as a template to detect the locations of these blobs. In the second stage, a mask is used in an orientation-based correlation in an area around the roughly located face location. The mask describes the global aspect of a human face. In experiments the faces in 69 out of 70 images are correctly located.

Yuille et al. [YHC92] introduce deformable templates to extract features from the human face. A template for the eye and mouth are proposed. The eye template is defined in terms of parabolas and circles that describe an outline of the shape of the eye. A separate template is used to model a closed mouth and an open mouth. Both use parabolas to described the outline of the mouth. The circles and parabolas are controlled by a set of parameters. Changing these parameters result in a change in the shape of those templates. An energy function is defined in terms of those parameters and the peak, valley, edge and image intensities. The energy function is defined such that its gradient is computable, therefore the method of steepest descent can be used for optimising the parameter values. The method is used successfully to locate the shapes of the mouth and eyes in their test images.

Brunelli et al. [BP93] compare the use of geometric feature-based matching versus template matching for face recognition. They also describe a method of detecting facial features. In the feature-based approach, the gradient of the input image is projected onto the vertical and horizontal axes. Detection of the mouth, nose and

eyebrows is found by examining the peaks and valleys of the projected gradients and comparing them against anthropometric standards. The face outline is fitted using an ellipse over the gradient intensity map and optimised using a cost function and dynamic programming. The template matching strategy involves the correlation of eyes, nose, mouth and face templates.

Yilmaz et al. [YS02] use skin colour as a first step in extracting an elliptical region of the face. Facial features such as the eyes and eyebrows are detected from the contrast in the skin colour. Training templates of the eyes and eyebrows are used to calculate the weighted probability that a colour belongs to a particular feature. This weight is a function of the distance from the center of the feature. A model distribution is created for each facial feature, which is then used for comparison with an input distribution. Edges found using the Sobel edge detector are used to locate lines that connect the eyes, eyebrows and the lines perpendicular to them running through the center of the eyes. Motion information is utilised by using confidence measures that are functions of the previous frames. In their experiments facial features are correctly located in 1018 out of 1358 images. Other methods of face detection based on skin colour have been investigated [CN98, CB94, CB97]. Gaussian density functions [YA98] and a mixture of Gaussians [YA99] are also used to model skin colour.

Wong et al. [WLS01] use the genetic algorithm (GA) to search for possible face regions in an image. A pair of eyes are selected by the GA by using the valley points in the image. The candidate face is then extracted and projected onto an eigenspace representing the space of faces. The distance from this space is calculated, from which a fitness function is derived. The symmetry of the face is also measured and used for refining the feature locations. Experiments are performed on the MIT face database, and 100 % detection rate is obtained for frontal faces with uniform lighting. With heads tilted a detection rate of 95.3 % is obtained. Genetic algorithms are also used to extract facial features by Lin and Wu [LW99].

Lam et al. [LY96] use snakes to locate the boundary of the head. Using anthropometric standards, a search is performed inside the rough locations of the eyes. The corners of the eyes are detected using curvature, orientation and region dissimilarity properties. More accurate eye locations are obtained by using the deformable template proposed by Yuille et al. [YHC92].

Wavelets are used by Kondo and Yan [KY99] for face detection and for detecting the symmetry of the face. The Haar wavelet is used for its simplicity. The horizontally high-passed and vertically low-passed, plus the vertically high-passed and horizontally low-passed images are used for symmetry detection. In particular, they are used for edge detection, followed by the extraction of edge blocks and facial edge blocks, where T-shaped blocks are detected and passed as possible face regions. Combined with gradient orientation a symmetry map is created. A model face template is cross correlated with the reduced number of possible face locations to find the most likely face location. Using edge information, the technique is robust to illumination variations. Detection rates of 100 % are achieved for head-on illumination, and 90 degree illumination, of faces from the MIT database. A detection rate of 93.33 % is achieved for 45 degree illumination.

Intrator et al. [IRY96] use the Generalized Symmetry Transform [RWY95] combined with the Radial Symmetry [RWY95] as the first steps in detecting the locations of the eyes and mouth. The highest peaks from those transforms are detected and the midline of the face image is found from the peak value of the autocorrelation function of the edge image. The assumed geometry of the eyes and mouth are then used to locate those features using the midline.

Yang et al. [YH94] use a hierarchical approach for detecting faces. A search is performed at three different resolutions. At the coarsest resolution a search is made for what a face looked like. At the finest resolution facial features are searched. Experiments are performed on 60 images, and faces in 50 images are detected accurately. False alarms occur in 28 images. A similar approach is presented by Kotropoulos et al. [KP97].

Hidden Markov Models (HMM) have been used for face detection and recognition. Patterns are characterised by parametric random processes. The parameters of an HMM can be optimised for training data using standard Viterbi segmentation and Baum-Welch algorithms [RJ93]. One dimensional and pseudo 2D HMMs are used for extracting facial features and for face recognition by Samaria et al. [SY94, Sam94]. HMMs are also combined with the Karhunen Loève Transform (KLT) to perform face localisation and face recognition by Nefian and Hayes [NH98].

Support Vector Machines (SVM) are used by Osuna et al. [OFG97] for face detection. SVMs are learning machines that implement the principle of structural risk minimisation, where the training process aims to minimise the upper bound on the expected generalisation error. This is in contrast with many other methods of training that aim to minimise training error. SVMs can be linear or non-linear in the mapping of input samples to the feature space, depending on the type of kernel used. Some of the most commonly used non-linear kernels are the homogeneous and inhomogeneous polynomial and the radial basis function kernels. An optimal decision hyperplane, linear in the feature space, is found by solving a linearly constrained quadratic programming problem. In the system described by Osuna et al. a database of face and non-face examples of size 19×19 pixels are trained using an SVM with a second degree polynomial kernel function. Bootstrapping is also used during the training process. Face detection then involves running the SVM over 19×19 subwindows in the input image at different scales to account for size variations. In tests with 313 images containing a single face a detection rate of 97.1 % is obtained with 4 false alarms. In a separate test with 23 images with a total of 155 faces a detection rate of 74.2% is obtained with 20 false alarms.

Smeraldi et al. [SCB99] use a log-polar retinotopic sampling grid, where the Gabor decomposition is used at each grid point for feature extraction. Facial features are extracted and trained using SVMs. They also employ a saccadic search strategy whereby the focus of the grid is centered on the grid point with the highest response. The sampling grid is denser at the center than the outer regions. In experiments using the M2VTS database containing 349 images, the system locates the eyes and

mouth in 91 % of these images to within 3 pixels. Reduced set vectors are used by Romdhani et al. [RTSB01] for face detection.

1.3 Face Recognition

Once a face image has been extracted and normalised, recognition can be carried out. As mentioned above, depending on the application, recognition may be used for either identification or verification. Identification is mainly used to identify an unknown individual, and verification is mainly used in access control. Identification systems are usually measured in terms of recognition rate, and the cumulative recognition rate of the top n matches. Verification systems are usually measured in terms of False Accept Rates (FAR) and False Reject Rates (FRR). Current state-of-the-art technologies in face recognition include those that use linear and non-linear discriminant analysis, statistical methods aimed at efficient class separation, and elastic graph matching based methods to name a few. There is a tremendous amount of material in the literature related to face recognition, a small sampling is described here. We categorise existing methods in a way similar to that used for face detection algorithms described in the previous section, but we leave out the knowledge-based category because it is not as suited for recognition. The remaining categories are: feature invariant approaches; template matching methods; and appearance-based methods. Feature invariant approaches use structural features that are invariant to changes in pose, expression, or illumination. Template matching methods use a set of patterns representative of the face, which are then correlated with the input image. Appearance-based methods perform recognition using models or templates learnt from a set of representative training images. Note that some methods can be placed into more than one category. We group the existing methods we are describing into those three categories and a summary is shown in Table 1.2.

Category	Used by
Feature Invariant	
Feature Points	Lam and Yan [LY98]
Geometry and Shape	Brunelli and Poggio [BP92], Hjelmås and Wroldsen [HW99]
Template Matching	
Fixed Template	Brunelli and Poggio [BP93, BP92]
Elastic Graph Matching and Dynamic Link Architecture	Lades et al. [LVB$^+$93], Wiskott et al. [WFKvdM97], Tefas et al. [TKP01], Ma and Tang [MT03]
Appearance-based	
3D	Voth [Vot03], Ansari and Abdel-Mottaleb [AAM03], Lee and Ranganath [LR03]
Neural Network	McGuire and D'Eleuterio [MD01], Er et al. [EWLT02], Haddadnia et al. [HFA03], Salah et al. [SAA02], Lawrence et al. [LGTB97], Palanivel et al. [PVY03]
Fuzzy	Haddadnia et al. [HFA03]
Near-infrared	Pan et al. [PHPT03]
Principle Components Analysis (PCA)	Turk and Pentland [TP91a], Zhang et al. [ZPZP00], Li et al. [LQLP02], McGuire and D'Eleuterio [MD01], Er et al. [EWLT02], Lanitis et al. [LTC02], Gunturk et al. [GBA$^+$03], Wang and Tang [WT03], Liu et al. [LCT03]
Kernel PCA	Kim et al. [KJK02], Yang [Yan02]
Density Estimation	Moghaddam et al. [MP97, MWP98, MJP00]
Linear Discriminant Analysis and Fisher's Linear Discriminant (FLD)	Etemad and Chellappa [EC97], Belhumeur et al. [BHK97], Liu and Wechsler [LW01], Er et al. [EWLT02], Lu et al. [LPV01, LPV03a]
Kernel FLD	Yang [Yan02]
Local Feature Analysis	Penev and Atick [PA96]
Independent Component Analysis	Havran et al. [HHC$^+$02], Liu and Wechsler [LW03]
Group Decision-Making	Jing et al. [JZY03]
Robust Correlation	Matas et al. [MKK97], Savvides and Kumar [SK03]
Hausdorff Distance	Gao et al. [GHF03], Lin et al. [LLS03]
Illumination Cone	Belhumeur and Kriegman [BK98], Zhang and Samaras [ZS03]
Wavelets	Lai et al. [LYF01]
Hidden Markov Model	Eickeler et al. [EMR00], Liu and Chen [LC03], Kim et al. [KKL03]
Observable Markov Model	Salah et al. [SAA02]
Support Vector Machines	Tefas et al. [TKP01], Smeraldi et al. [SCB99], Lu et al. [LPV01], Cui and Gao [CG03]
Fractals and Fractal Image Coding	Kouzani et al. [KHS97, KHS00, KHS99], Chandran and Kar [CK02], Ebrahimpour-Komleh et al. [EKCS01b, EKCS01a], Neil et al. [NC96, NCC96],

Table 1.2: Summary of existing face recognition algorithms.

1.3.1 Overview of Face Recognition Algorithms

Lades et al. [LVB+93] introduce the concept of the dynamic link architecture (DLA) for object recognition. Central to DLA is a grid of vertices, each of which is allowed to move independently. Gabor decomposition is carried out at each vertex, which uses the Gabor-based wavelet defined as

$$\psi_{\vec{k}_{\nu\mu}}(\vec{x}) = \frac{\vec{k}_{\nu\mu}^2}{\sigma^2} \exp\left(-\frac{\vec{k}_{\nu\mu}^2 \vec{x}^2}{2\sigma^2}\right) \left[\exp\left(i\vec{k}_{\nu\mu}\vec{x}\right) - \exp\left(-\sigma^2/2\right)\right] \tag{1.5}$$

where $\vec{x} = (x, y)$ is a vector representing the position of a pixel in the image. The Gaussian envelope function in Equation (1.5) is calculated at each vertex point $\vec{x}_0$ for each $\nu \in \{0, \ldots, 4\}$ and $\mu \in \{0, \ldots, 7\}$, where

$$\vec{k}_{\nu\mu} = k_\nu e^{i\phi_\mu} \text{ with } k_\nu = k_{max}/f^\nu, \phi_\mu = \frac{\pi\mu}{8} \tag{1.6}$$

where the spacing factor between kernels in the frequency domain is denoted by f. The authors investigate the effects of using different values for f and k_{max}. An example of the values used is $f = 2$ and $k_{max} = \pi/2$. Feature vectors at each position $\vec{x}_0$, called a "jet", are formed. Each jet has a length of $|\nu||\mu| = 5 \times 8 = 40$. The graph of a model is created by placing an undistorted grid of vertices over the image representing the model. Jets are created at each vertex of the model graph. Jets are also created for an input image to which the model is to be matched. A similarity function is defined as the normalised dot product between a jet in the image domain with a corresponding jet in the model domain. Each vertex has four neighbours, and the four edges are labeled by the Euclidean distance vector from one vertex to the other. A cost function related to the amount of graph distortion between the image and model is defined as the squared norm of the difference vector between the corresponding Euclidean distance edge vectors in the image and model graph. A total cost function is defined as

$$\mathcal{C}_{\text{total}}(\{x_i^I\}) = \lambda\mathcal{C}_e + \mathcal{C}_\nu \tag{1.7}$$

where $\{x_i^I\}$ is the set of vertex positions in the image domain that should be optimised, $\mathcal{C}_e$ is the total cost of edge deformations, λ controls the rigidity of the graph

and $\mathcal{C}_\nu$ is the total similarity between corresponding jets in the image and model domain. The optimal vertex positions in the image domain are found using simulated annealing at zero temperature. Significance and acceptance criterias as functions of the total costs are defined and used for recognition. In their experiments with 87 people, accuracy rates up to 88 % are achieved. A similar method using elastic bunch graph matching is described by Wiskott et al. [WFKvdM97]. Support vector machines (SVM) are used with weighted morphological elastic graph matching for face verification by Tefas et al. [TKP01], and they obtain an equal error rate of 2.4 % on the M2VTS database.

Brunelli and Poggio [BP92] perform face recognition using geometrical features. The eyes are detected from a normalised image using an eye template scanned through the image at five different scales. Separate left and right eye templates at different scales are used to improve the location of the eyes. The horizontal and vertical projections of the edge map are generated. For an image $\mathcal{I}(x, y)$ in the domain $[x_1, x_2] \times [y_1, y_2]$, the horizontal projection is given by $H(y) = \sum_{y=y_1}^{y_2} \mathcal{I}(x, y)$ and the vertical projection is given by $V(x) = \sum_{x=x_1}^{x_2} \mathcal{I}(x, y)$. The projections are analysed and twenty two geometrical features are extracted, which include features such as: eyebrow thickness and vertical position of the eye center; nose vertical position and width; mouth vertical position, width and height; chin shape, described by eleven values; bigonial breadth; and zygomatic breadth. The distance between two feature vectors $\{x_i\}$ and $\{x_i^{'}\}$ is defined as

$$\sum_{i=1}^{n} w_i \left(\frac{|x_i - x_i^{'}|}{\sigma_i} \right)^{\alpha} \tag{1.8}$$

where $\{\sigma_i\}$ is the inter-class dispersion vector, $\{w_i\}$ is the weight vector and various values of α are tested to achieve the best performance. This distance measure is used for recognition. This method is compared to a template based method in [BP93]. Eye, nose, mouth and whole face templates are used for correlation using various types of normalisations, ranging from no preprocessing to a Gaussian regularised image.

Turk and Pentland [TP91a] introduce the concept of eigenfaces. Their work is

inspired by the method proposed by Sirovich and Kirby [SK87] using principal component analysis (PCA) to efficiently represent human faces in a lower dimensional space. Turk and Pentland extend its use to human face recognition. The process of computing the face space begins with an ensemble of face images. Principle components are found that maximise the variance along each coordinate. The coordinates are orthogonal to each other. They find that faces can be approximated by M coordinates, where $M \ll N$ and N is the dimension in the original space. Therefore, face images can be efficiently represented by M weights. Training faces belonging to different classes are projected onto the face space, then recognition is performed by minimising the Euclidean distance in the face space. That is, the best matching class is found from

$$\arg\min_k(\epsilon_k) = \arg\min_k \|(\mathbf{\Omega} - \mathbf{\Omega}_k)\|^2 \tag{1.9}$$

where $\mathbf{\Omega}$ is the projection of the unknown face image into the face space, and $\mathbf{\Omega}_k$ is a vector belonging to the kth face class. This vector can be obtained from a single face belonging to the kth class or it can be the average of several eigenface representations in that class. Experiments are performed on 2500 images of 16 individuals and accuracy rates of 96 % averaged over lighting variation, 85 % averaged over orientation variation and 64 % averaged over size variation are obtained.

Dual eigenspaces are used by Zhang et al. [ZPZP00], which allows a coarse to fine matching strategy. The input face is projected onto the usual eigenface space. A few potential faces are selected and projected onto class specific eigenspaces for further finer classification.

Linear discriminant analysis (LDA) is used by Etemad and Chellappa [EC97] for face recognition. Within and between class scatter matrices are used to derive their own measure of discrimination power. A transformation matrix is constructed that maximises class separability, which at the same time performs dimensionality reduction. This type of analysis is possible because they use 1500 images for training and each sample is 25×50 pixels, that is the dimensionality of the samples is 750. There are more samples than the dimensions of the samples, so there are no

problems with singular matrices. The training set is increased by including mirror images and noisy versions of the original images. The ORL database, with some manually extracted faces from the FERET database are used in experiments. The system achieves an accuracy rate of 100 % on the training set and 99.2 % on the test set.

It is now well known that the eigenface method in the original form is very sensitive to illumination. The projection produced by PCA is along the direction of maximum variance, not class separability as required for recognition. The projection in fact retains unwanted variations due to lighting and facial expression as described by Belhumeur et al. [BHK97]. The authors describe an alternative method, which they called Fisherfaces, for face recognition. Fisher's Linear Discriminant (FLD) [Fis36] is used to maximise between-class scatter and minimise within-class scatter. The optimal projection is given by

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{1.10}$$

where S_B is the between-class scatter matrix and S_W is the within-class scatter matrix. Equation (1.10) reveals that S_W is often singular because the number of images in the learning set is usually much smaller than the number of pixels in the image. The authors overcome this problem by projecting the image set to a lower dimensional space using PCA before applying the standard FLD. They call this method Fisherfaces. Recognition is carried out by using the nearest neighbour classifier in the space generated by Fisherfaces. In experiments with the Yale Database using the Leave-One-Out method, an error rate of 0.6 % is achieved. The authors also point out that the Eigenface method is equivalent to correlation when the number of principal components used equals the size of the training set. However, as pointed out by Liu and Wechsler the performance of this method is only superior to the Eigenface method if the training images are representative of the range of class variations, otherwise the performance difference is not significant [LW01]. To overcome this problem they propose the use of the Enhanced Fisher Classifier (EFC). They improve the generalisation capability of the FLD by the simultaneous diagonalisation of the two within- and between-class covariance matrices. The method is

tested on human face recognition using manually annotated facial features. Shape and texture features are extracted and concatenated into a single feature vector, which is used with the EFC to perform recognition. Six hundred images of 200 subjects from the FERET database is used to evaluate the system, and a recognition rate of 98.5 % is reported when 25 features are used. This method is extended by the same authors to use Gabor features [LW02]. An augmented Gabor feature vector is created from each input image using Gabor decomposition. This vector is used with the EFC for face recognition. They also experiment with four different similarity measures, the L_1, L_2, Mahalanobis and cosine similarity measures. Again using 600 images of 200 subjects from the FERET database the system achieves a recognition rate of 100 % when 62 features are used.

As described in the previous section, Moghaddam et al. [MP97, MWP98] present a density estimation technique for target detection, recognition and coding. They use the concept of distance-in-feature-space (DIFS) and distance-from-feature-space (DFFS). The density estimate $\hat{P}(\mathbf{x}|\Omega)$ is used for face detection. Subwindows $\mathbf{x}$ at each spatial position (i, j) are extracted and the density estimate calculated. The position (i, j) that maximises $\hat{P}(\mathbf{x}|\Omega)$ is selected as the most likely face location. For recognition, a similar method is used to detect four facial features within the face image. The features are the left eye, right eye, tip of the nose and the center of the mouth. These locations are used to normalise the face, which is then projected onto a custom set of eigenfaces. A view-based approach is used for recognition. For N individuals and M distinct views, M independent subspaces are created, each representing a different view. In tests with 189 images of 21 people an average recognition rate of up to 90 % is obtained. The authors also try using just the features for recognition, called eigenfeatures. In tests with 45 individuals a recognition rate of 95 % is achieved. Complementary subspaces are also used in a Bayesian approach to face recognition by Moghaddam et al. [MJP00]. A probabilistic similarity measure is derived based on a Bayesian belief that the intensity difference image between two faces is characteristic of typical variations in an individual's

appearance. Experiments on the FERET database demonstrate that the system achieves an accuracy of 95 % in returning the correct match as a first rank.

Li et al. [LQLP02] use the orthogonal complement PCA (OCPCA) for recognition. Two orthonormal bases S_B and S_{AB} are generated using PCA with the sets of images D_B and D_{AB}. The image set P_B consists of pairs of images from the same person, and P_{AB} consists of pairs of images from different individuals. The difference set D_B is generated from the differences between pairs of images in P_B, and D_{AB} is generated from the differences between pairs of images in P_{AB}. The orthogonal complement of S_B is found in the space S_{AB}, which is used to calculate a recognition distance for each input image based on multiple eye positions. Experiments on the FERET database give recognition rates of up to 5.4 %.

McGuire and D'Eleuterio [MD01] use a method called eigenpaxels and neural networks to perform image classification. A paxel is defined as a pack of pixels of any shape. The authors used square paxels of size 16×16. Paxels are localised with respect to the original image. Ten thousand random paxels are obtained from 200 images from the ORL face database. Principal component analysis is performed on this ensemble of paxels, thus the name eigenpaxels. Input images are of size 92×112 pixels. Paxels are extracted at all positions where they overlap by 12 pixels in the horizontal and vertical directions. Each paxel is projected onto the eigenpaxel space and the first ten eigenpaxels are retained. Ten images of size 20×25 are then obtained, where each image corresponds to the responses obtained at paxel positions within the input image along a particular eigenpaxel coordinate. These ten images are subsampled by a decimation factor of two. Each response is connected to an Error Correcting Network (ECN) with 40 fully connected neurons. The ECN is trained with 200 images. Recognition experiments are performed and a recognition rate of $2.9\% \pm 1.1\%$ is obtained.

Er et al. [EWLT02] use an RBF neural network classifier to recognise features extracted by an FLD operating on the PCA projection of the original image data. Their method allows them to circumvent some of the problems associated with an

RBF classifier when applied to face recognition, namely the problems of overfitting, overtraining, small-sample effect and the singularity of the covariance matrix. In face recognition the original data space is usually very large compared to the number of samples available, and using the PCA plus FLD alleviates this problem. A new clustering technique is introduced to cluster homogeneous data and to achieve a compact structure with limited mixed data. Criterias are used to estimate the initial widths of RBF units to control the generalisation of the classifier. Finally, a hybrid learning algorithm combining the gradient and the linear least square (LLS) paradigm is used to adjust the learning parameters. An average recognition error rate of 1.92 % is obtained in experiments using the ORL database.

A fuzzy hybrid learning algorithm is used with the RBF neural network by Haddadnia et al. [HFA03]. The number of neurons in the network are determined using cluster validity indices with majority rule. Fuzzy clustering is used to initialise the characteristics of the hidden neurons. The fuzzy hybrid learning algorithm combines the gradient and linear least squares methods for adjusting the RBF parameters and the neural network weights. Experiments using the ORL database, where the number of training and testing images are 200 each, show that their system achieves an error rate of 0.45 %.

Local feature analysis (LFA) is used by Penev and Atick [PA96] for face representation. The authors point out that PCA does not preserve topography, in that nearby values in the image space are not related to nearby kernels in the reduced space. Topography is imposed in LFA by labelling the kernels in the reduced space representation with variables from the input space rather than the eigenmode index. Output is decorrelated by minimising an appropriate function. They demonstrate that the LFA kernels correspond to features in the face. This technique is used in a commercial system by Visionics Corporation [Kro02] (merged with Identix Incorporated as of June 26, 2002).

Kernel PCA is used to perform face recognition by Kim et al. [KJK02]. A kernel function maps the input image space to a higher dimensional feature space on which

the standard linear PCA is computed. A facial feature vector is extracted by kernel PCA and classification is via linear SVMs. Experiments are performed on the ORL database using 20 randomly selected training and test sets of faces. Half of the faces are used for training and the other half for testing. An average error rate of 20 % is obtained. The use of kernel PCA and kernel FLD is explored by Yang [Yan02]. Experiments on the ORL database reveal that kernel PCA achieves an error rate of 2.0 % and kernel Fisherface achieves an error rate of 1.25 %.

As mentioned in Section 1.2.1 Smeraldi et al. [SCB99] use saccadic exploration of a scene based on a log-polar retinotopic grid and Gabor feature extraction for facial feature detection. That method is extended to face authentication by Smeraldi and Bigün [SB02]. Once the eyes and mouth are detected, the responses from three retinotopic sensors are used for classification. They experiment with Nearest Neighbour (NN), K Nearest Neighbour (KNN) and SVM classifiers. The output from the three classifiers are combined by fusion. Majority voting is used with the NN classifier, the scores from the KNN classifiers are added together, and the output from the SVMs are combined using the nonlinear tanh function. They use the XM2VTS database for their experiments, with the dataset partitioned according to Configuration II as described by the Lausanne Prototcol [MMK$^+$99]. However, they did not perform the experiments as strictly as described by the Prototcol in that the evaluation set is not used to set their system parameters. The training of each client is performed by taking all other clients as negative examples. Experiments on the client and impostor test sets show that the system achieves an equal-error rate of 0.5 %.

SVMs are also used by Lu et al. [LPV01] to perform recognition. Feature extraction is performed using their proposed technique called DF-LDA. They note that with traditional LDA algorithms the separability criteria is not directly related to classification accuracy, which may result in unreliable classifications. Fisherfaces perform an intermediate PCA step for dimensionality reduction to overcome the small sample size problem. The drawback with this intermediate step is that it is possible that some discriminatory information is discarded in the process [YY03].

In fact, in some instances, eigenfaces can outperform Fisherfaces [LW00]. The proposed method uses incremental dimensionality reduction, where at each step a weighted between-class scatter is recomputed and rediagonalised. The weights are recomputed such that classes that come together have larger weights. The effect is to avoid severe overlap between classes in the output space. The ν-SVM classifier is then used for training and classifying the extracted features. The SVM implements structural risk minimisation where the upper bound on generalisation error is minimised. In comparison many neural networks and other classifiers attempt to minimise empirical risk. The ν-SVM is used instead of the C-SVM. The parameters ν and C are training parameters, where C is known as the regularisation constant. Adjusting the parameter ν changes the upper bound on the fraction of margin errors and the lower bound on the fraction of support vectors. The advantage of using ν-SVM is that it permits the control of the number of support vectors and errors, whilst C does not have an intuitive interpretation. However, using the ν-SVM does not necessarily guarantee better performance. In experiments using the ORL database, half of the samples is used for training and the other half for testing. A recognition error rate of 3.125 % is achieved. The same authors also describe a method of using kernel direct discriminant analysis [LPV03a]. Direct LDA overcomes the problems of small sample sets whilst at the same time minimising the loss of discriminant information by only discarding the null space of the between class scatter matrix. The authors extend this idea to a nonlinear higher dimensional space by using kernels.

A recent study by Liu et al. [LHK02] using magnetoencephalography (MEG) finds evidence that a region in the human brain responds to faces in general whether or not the face is later recognized. This occurs at 100 ms after stimulus onset. Another region of the brain seems to perform categorization at a later stage to extract the identity of the face. Salah et al. [SAA02] develop a visual pattern recognition model based on primate selective attention mechanism. They attempt to simulate both the bottom-up and top-down visual processing that occurs in the brain. The model has three levels, attentive, intermediate and associative. Basic processing occurs at

the attentive level, such as line detection. A single-layer perceptron is used at the intermediate level for basic recognition. The single layer network is preferred over multilayer perceptrons because the latter has the tendency to overlearn, use longer training times and requires more parameters. The observable Markov model (OMM) is used at the associative level for recognition. They prefer using the OMM over the Hidden Markov Model (HMM) for its faster training time and the observability of the model parameters. This allows the observation probabilities to be calculated. Saccades are also simulated. For face recognition, they use Gabor wavelet filters as feature maps at the attentive level. In experiments with the ORL database they obtain a recognition accuracy of $92.0\% \pm 7.98\%$ using 10-fold cross-validation.

Lam and Yan [LY98] first locate the head boundary using snakes, an active contour model [KWT87]. The corners of the eyes and mouth are detected using a corner detection scheme proposed by Xie et al. [XSZ93]. Altogether, 15 feature points are detected. These points are then compared to those in a database. Similar faces in terms of these feature points are then selected for further recognition via correlation. The head is modeled as a cylindrical volume to account for out-of-plane rotations of the head. Three correlation windows are used. They are the eyes and eyebrows, the nose and the mouth windows. The results from the correlation and the feature point comparison are combined into a single score by weighted summation. With automatically detected feature points the system achieves a recognition rate of 84 %. With manually detected feature points the recognition rate is 89 %. The ORL face database is used in their experiments.

Matas et al. [MKK97] propose a novel correlation method that performs face localisation, normalisation and identification simultaneously. The normalisations used are geometric and photometric. The geometric transformations considered are translation, scaling and rotation. A score function is defined as

$$s(t) = \alpha s_{area}(t) + (1 - \alpha) s_{grey}(t) \tag{1.11}$$

where t is a geometric or photometric transformation and α is a constant between zero and one. The area score s_{area} is defined to encourage large overlaps between

the test and reference image,

$$s_{area}(t) = \frac{|S_t \cap S_r| - |S_t \cap S_r^c|}{|S_t|} \tag{1.12}$$

where S_r and S_t are the sampling sets for the reference and test images. The grey-level score is defined as,

$$s_{grey}(t) = \frac{\sum_{p_r \in S_r \cap S_t} f_k(f_i(I_r(p_r)), I_t(f_p(t, p_r))}{f_k^{max}|S_r \cap S_t|} \tag{1.13}$$

where the robust kernel f_k compares grey levels, f_k^{max} is the maximum response of the robust kernel, f_i is the intensity transformation that transforms a grey level g by $gt_{slope}+t_{offs}$ where t_{slope} and t_{offs} are parameters, I_r and I_t are the reference and test images, and f_p is a projection that rotates, scales and translates a pixel. The robust kernel is defined as $f_k(g_1, g_2) = -(g_1 - g_2)^2 + d_c^2$ if $|g_1 - g_2| < d_c$ and zero otherwise, where g_1 and g_2 are grey levels and d_c is the cut-off distance. The parameters t are optimised by random exponential perturbations similar to simulated annealing at zero temperature [KGV83]. To improve computation time the images are sampled at random positions as given by two-dimensional Sobol sequences [PTVF92]. In leave-one-out experiments using the M2VTS database an equal error rate of 5.4 % is achieved.

Lawrence et al. [LGTB97] use a hybrid neural network solution, combining local image sampling, self-organising maps (SOM) [Koh90] and a convolutional neural network. The SOM preserves the topological ordering of classes, that is, similar patterns in the input space are also close to each other after projection onto the map. The SOM is made up of nodes, and each node is represented by a vector of the same size as the input vector. During training the closest matching node to the input vector is found. A neighbourhood smoothing kernel function centered on the best matching node is used to weight the update values for each node. The SOM is used for dimensionality reduction and is trained using subimages of size 5×5 pixels extracted from a set of training images from the ORL database. They use a three dimensional SOM with five nodes per dimension. A 25 dimensional input vector is reduced to three dimensions, resulting in three maps. These act as inputs to a

convolutional neural network which makes a classification decision. In experiments using the ORL database they obtain a recognition error rate of 3.8 %.

Eickeler et al. [EMR00] use pseudo 2D Hidden Markov Models (HMM). An HMM is a statistical model with various states. A transition probability matrix determines the transition from one state to another. Pseudo 2-D HMMs are nested one-dimensional HMMs. Features are extracted using the DCT. In particular, compression is performed using the JPEG standard. The Baum-Welcm algorithm is used to train the HMMs. A common HMM model is generated from all the available training data, and this is used as the starting points for training the class specific HMMs. Experiments are conducted on the ORL database, from which half is used for training and the other half for testing. In addition to the original training data, mirrored images are also used. The authors also experiment with using the original uncompressed data for training, from which they obtain a recognition error rate of 0 %. An error rate of 0.5 % is obtained with DCT compressed features when the compression ratio is set at 7.5:1.

Independent component analysis is a data analysis tool used for source separation. That is, to recover original signals from known observations where the observations are a mixture of the original signals. ICA maximises the statistical independence of output variables by using contrast functions, such as Kullback-Leibler divergence, negentropy and cumulants [Com94, HKO01]. ICA has been adapted for face verification [HHC$^+$02], whereby faces are taken as the observation signals, and the linear combination of unknown independent source vectors is approximated, which can be used to reconstruct the observations. The source vectors, once approximated, are ordered according to a class separability criteria, which is usually the ratio of between class to within class variance. Some drawbacks of ICA include convergence problems and difficulties in handling high dimensional data. To overcome this the authors perform PCA before applying ICA. Face verification is performed by comparing feature vectors using the Euclidean distance and the angle between the vectors. The results are better when angles are used. In experiments using the

XM2VTS database consisting of 295 people, they obtain an equal-error rate of 5.62 % on the learning set. On the validation set they obtain $(FAR+FRR)/2 = 5.21\%$.

Lai et al. [LYF01] combine wavelets and the Fourier transform to achieve recognition that is invariant to scale, rotation and translation. The Daubechies wavelet D4 [Dau90] is used to achieve some invariance to facial expressions. This is demonstrated by visually observing the low frequency image of a three level wavelet decomposition. The factors affected by scale, rotation and translation in the Fourier transform are examined and negated by introducing opposing factors. To account for illumination differences, histogram equalisation is used and light intensity is normalised. Experiments are performed on the Yale and ORL databases. When one reference image is used per person, a recognition rate of 91.33 % is achieved on the Yale database, and 81.94 % is achieved on the ORL database. When three images per person is used as a reference from the ORL database, an accuracy rate of 94.64 % is achieved. When two reference images per person is used from the Yale database, an accuracy rate of 95.56 % is achieved.

Changes in facial illumination have a significant effect on recognition rates. Phillips and Vardi [PV96] normalise illumination on different parts of the face by transforming and matching histograms. A more sophisticated approach is taken by Georghiades et al. [GBK01]. It is observed that the set of images of an object in a fixed pose and illuminated from all possible locations is a convex cone in the space of images [BK98]. Furthermore, if the surface reflectance is Lambertian, the illumination cone can be constructed using a few images under variable lighting and the cone can be well-approximated with a low-dimensional subspace. A low-dimensional subspace is created for each face by applying SVD. Seven training images per face in a frontal pose under slightly different lighting conditions are used for recognition. Recognition is performed using the illumination cone representation. Experiments on the Yale Face Database B give recognition error rates of 0.9 % for the frontal pose, 2.7 % for 12° poses and 5.5 % for 24° poses. Shashua and Raklin-Raviv [SRR01] use a different method for achieving illumination invariance. They define a quotient image from which images with artificial illuminations can be generated. This image

is also illumination independent, so it is used for recognition. Two experiments using 1800 images prepared by Vetter et al. [VB98, VJP97] are performed and they obtain error rates of 0 % and 0.33 %.

Aging effects are explored by Lanitis et al. [LTC02]. Aging functions are derived from a set of images representing different ages ranging from 2 to 30 years old. PCA is used to create the model, and parameters are derived that correspond to different ages. This age simulation is used for face recognition, and accuracy rates of 71 % is obtained compared to 63 % without age simulation.

Kouzani et al. [KHS99] use the fractal dimension for recognition. The proposed image matching method compares two images by first calculating the fractal dimension of local areas. Each pixel is then characterised by the fractal dimension in its neighbourhood region. To account for different textures, the fractal dimension is calculated for regions of different sizes and then averaged. Problems associated with pixels at the borders of an image are alleviated by extending the original image using circular shifting. Based on the local fractal dimensions, normalised cross correlation is applied to measure the difference between two images.

The same authors use fractal image coding for recognition in [KHS97]. They use the fractal code, which is made up of the parameters of a Partitioned Iterated Function System (PIFS) code of images as features. Two neural networks are used in their system. One feedforward neural network is trained to find the best matching domain block for each range block. Another feedforward neural network is trained with the fractal codes of training images. Recognition is performed by feeding the fractal code of an unknown image into the neural network and observing the output. Fractal codes are also used for recognition in [CK02, EKCS01b, EKCS01a]. Using fractal codes in their raw form for recognition poses some problems, because one can generate many fractal codes that represent the same image. To use it properly the distance between codes and the continuity of parameters must be taken into account, as done by Chandran and Kar [CK02], who use it for image indexing in the content-based retrieval of images. Ebrahimpour-Komleh et al. [EKCS01a] extracted four

fractal features and they call them domain-range, brightness, rotation and contrast features. Each fractal feature is used as a vector, giving four vectors to represent an image. Classification is performed by calculating the Peak Signal-to-Noise Ratio (PSNR) between the unknown input feature vector and the reference feature vectors. Experiments using a subset of the MIT face database give a classification accuracy of around 85 %.

Fractal image coding is used for the recognition of binary images by Neil et al. [NC96, NCC96]. Here, instead of using the fractal codes as features, a reference database of fractal codes is applied to an unknown input image. The fractal code that changes the input image the least is taken as the best match.

The face recognition system described by Kouzani et al. [KHS00] compensates for illumination effects by using an embossing technique, which is achieved by suppressing colour information and outlining the area with a selected colour. The pose of an input face is detected and used to generate a 3D head. This head is then used to render a 2D front-view image of the input face. Recognition is performed using fractal image coding, which is similar to those described by Neil et al. [NC96, NCC96] except that greyscale images are used. They incorporate translational invariance by circularly shifting the image. Rotational invariance is achieved by projecting the input input face onto the complex plane z using the function $w = \ln(z)$. Here, a rotation in the input image corresponds to a translation in the output image. Scale invariance is achieved by using a multiresolution approach, which utilises the fact that if a domain block of size d^2 best approximates a range block of size r^2, then that same domain block of size $(d/s)^2$ in the image scaled by s also best approximates the corresponding range block of size $(r/s)^2$. The rotation and translation invariance processes are applied before the multiresolution procedure incorporating the domain-range block scale invariance property. Additional images obtained by manipulating the original images are also added to the reference database to increase invariance. In experiments using the Bern face database their method achieves a recognition rate of 98.11 %.

1.4 Overview of Thesis

As evident in the overview given in the previous section there are many algorithms in the literature describing face detection and recognition. The current trends are the use of elastic graph matching methods, linear and nonlinear discriminant analysis and statistical methods aimed at separating different classes effectively.

The work described in this thesis uses a distinctly different approach. Fractal image coding was traditionally used for image compression and analysis. As mentioned in the previous section, several authors have used it for recognition and classification. They are Neil et al. [NC96, NCC96], Kouzani et al. [KHS97, KHS00], Ebrahimpour-Komleh et al. [EKCS01b, EKCS01a] and Chandran and Kar [CK02]. Note that the material published by Kouzani et al. [KHS00] was developed concurrently with our work, without correspondence.

Generally, there are two approaches to using fractal image coding for recognition. The first type uses the fractal code itself, and discrimination comes from differences between the codes. The fractal code of an image is the parameters of a Partitioned Iterated Function System (PIFS) code generated for that image. The second type uses the decoding process of fractal image coding to perform recognition. It is more difficult to use the first type for recognition because fractal codes can change dramatically even between very similar images. The problem is that more than one fractal code can be generated for a given image. If we use fractal codes as features, we need to take into account the distance between codes, and to ensure the continuity of the parameters of the code [CK02]. Another disadvantage is that different parameters of the fractal code have different scales. An advantage of using the second approach over the first is that we do not have to worry about the uniqueness and distance between codes. We only require the uniqueness of the attractor, which is already an implied property of a properly generated fractal code. This thesis focuses on the second approach, from which we develop a distance measure that we call the Fractal Neighbour Distance (FND). Compared to other methods, research into fractal image coding based recognition methods has been scarce. One of the

reasons for this is that when adapted for recognition, the methods are not built with a statistical foundation designed for discrimination. However, preliminary experiment results are promising, and thus we deem the approach worthy of further investigation. The difference between our work and others is that we investigate in greater detail the inner workings of the FND. With better knowledge of the method we can exploit it further to achieve better recognition rates.

1.4.1 Aim and Motivation

The core aim of any face recognition research is undoubtedly to improve the recognition rate. In order to do that we have to examine the factors that stand in the way of achieving that goal. The factors are related to the complexities of the human face and the environment in which the images are taken. Some of the major contributing factors are variations in lighting, shadows, background, pose, facial expression, facial hair, hairstyle, eye-glasses, eye-patches, makeup, masks, aging, caps and hats, partial occlusions, viewing angle, scale and position. It is beyond the scope of this thesis to tackle all these problems. Instead, we focus on the factors related to linear and non-linear distortions of the face, and illumination variations. Furthermore, we also examine ways of reducing the effects of facial hair and expression.

Many existing face recognition methods improve the recognition rate by:

- using preprocessing to further normalise the image [BP92];
- using more real/synthetic training samples [KHS00, EMR00, EC97] and adding noise [EC97] to increase distortion invariance; and
- using multiresolution approaches [KHS00] to account for scale.

All these methods increase processing time. One of the advantages that our FND based classifier has over some of the other existing methods is that our method has inherent invariance to distortions and illumination changes, thus reducing the

amount of preprocessing required. These invariances are due to the nature of the method, and virtually come for "free" in terms of classification time. The work in this thesis explores these invariances, and we show how the invariances can be varied and adapted to different image types by changing the parameters in the fractal code. Furthermore, invariance to facial hair and expressions is obtained by using a weighted version of the FND. The weighted FND has the potential to be more invariant to non-linear distortions.

The FND based classifier also has advantages over existing methods such as eigenfaces, neural networks and those that use statistical learning theory. Some of these advantages are:

- the presence of inherent invariance to distortions and illumination changes that can be varied and adapted to different image types by changing fractal encoding parameters;
- new faces can be enrolled without re-training the whole database;
- faces can be removed from the database without the need for re-training;
- it is relatively simple to implement; and
- it is not model-based so there are no model parameters that need to be tweaked.

1.4.2 Organisation

Face localisation is an important part of a complete face recognition system, therefore a method of locating the eyes of the face is described in Chapter 2. Knowing the location of the eyes enable us to extract and normalise the face for scale and in-plane rotations. Currently, the system assumes that there is only one face per image. However, the method described in this chapter can be easily extended to perform face detection. The method is based on support vector machines (SVM)

and adapted for use with the fast Fourier transform (FFT) to achieve exhaustive searching for eye locations. This is straightforward for the linear SVM. For nonlinear SVMs we derive a novel and efficient approximate representation of the decision hyperplane, so that it can be combined effectively with the FFT to perform fast eye detection. In particular, we introduce the concept of ρ and η prototypes. We show how the decision function of an SVM can be reformulated and expressed efficiently in terms of these prototypes. This reformulation produces a more accurate approximation of the decision hyperplane than the well-known reduced set method proposed by Schölkopf et al. [SMB+99]. In adapting this reformulated decision hyperplane for use with the FFT, we introduce some kernel specific algorithms that improve the classification speed. The radial basis function (RBF) kernel receives extra treatment, where correlation redundancies are used to achieve a further speed increase. A map of possible eye locations is produced from the FFT-assisted SVM sub-system. The number of possible locations is further reduced by using an efficient two dimensional range-searching algorithm. An expert knowledge-based approach is then used to match up the left and right eye pairs. These eye locations are then used for face extraction and normalisation.

Chapter 3 describes fractals and how fractal image coding can be used for recognition. In particular, the Fractal Neighbour Distance (FND) is introduced. We investigate the mechanisms behind the FND that gives it inherent invariances to rotation, scaling, translation, illumination and perspective transformation. The relationship between the contractivity factor and the recognition rate is investigated, and this has not been done before. Changing the contractivity factor changes the invariances in the FND. Depending on the images used, using a larger or smaller contractivity factor can lead to better recognition rates. We identify the situations when smaller/larger contractivity factors should be used. Chapter 4 investigates the effects of changing encoding parameters on the recognition rate. In addition to the investigation of the effects of the contractivity factor, we also examine the eventual contractivity factor. We derive a method for extending our ability to control the convergence rate of the decoding process for a fractal code, and through

this demonstrate how better recognition rates can be achieved. We show that by adjusting the contrast scaling factor in a controlled manner we can control the eventual convergence rate of a fractal decoding process, which was not possible previously. In this chapter we also introduce two relative distances for measuring the performance of the FND. Experiment results evaluated using these two performance measures show that the FND in fact increases class separation, which is desirable in classification algorithms. Chapters 3 and 4 also demonstrate that using uniform range block partitioning gives significantly better results than using Quad-tree or HV-partitioning. The reason is that the latter methods impose a spatial structure that reduces the invariance of the FND to distortions.

A complete face verification and identification system is described in Chapter 5 combining the FND with the FFT-assisted SVM eye detector. A block diagram of the complete system is shown in Figure 1.1, where the *face recognition using fractal image coding* block includes both verification and identification systems. The FND is further customised separately for the verification and identification systems. Here, the weighted FND is introduced. The weighted FND incorporates the use of weights in a local region of an image. A local search algorithm is also introduced to search for a best matching local feature using this locally weighted FND. The scores from a set of these locally weighted FND operations are then combined to obtain a global score, which is used as a measure of the similarity between two face images. Each local FND operation possesses the distortion invariant properties described above. Combined with the search procedure, the method has the potential to be invariant to a larger class of non-linear distortions. We also present a set of locally weighted FNDs that concentrate around the upper part of the face encompassing the eyes and nose. This design is motivated by the fact that the region around the eyes has more information for discrimination [NS98][BP93]. Facial verification experiment results show that the weighted FNDs perform better than normal FNDs. We also introduce the use of normalised scores and client specific thresholds. Normalised scores improve our results to a point where it is competitive with some of the current state-of-the-art methods. Client specific thresholds give further significant

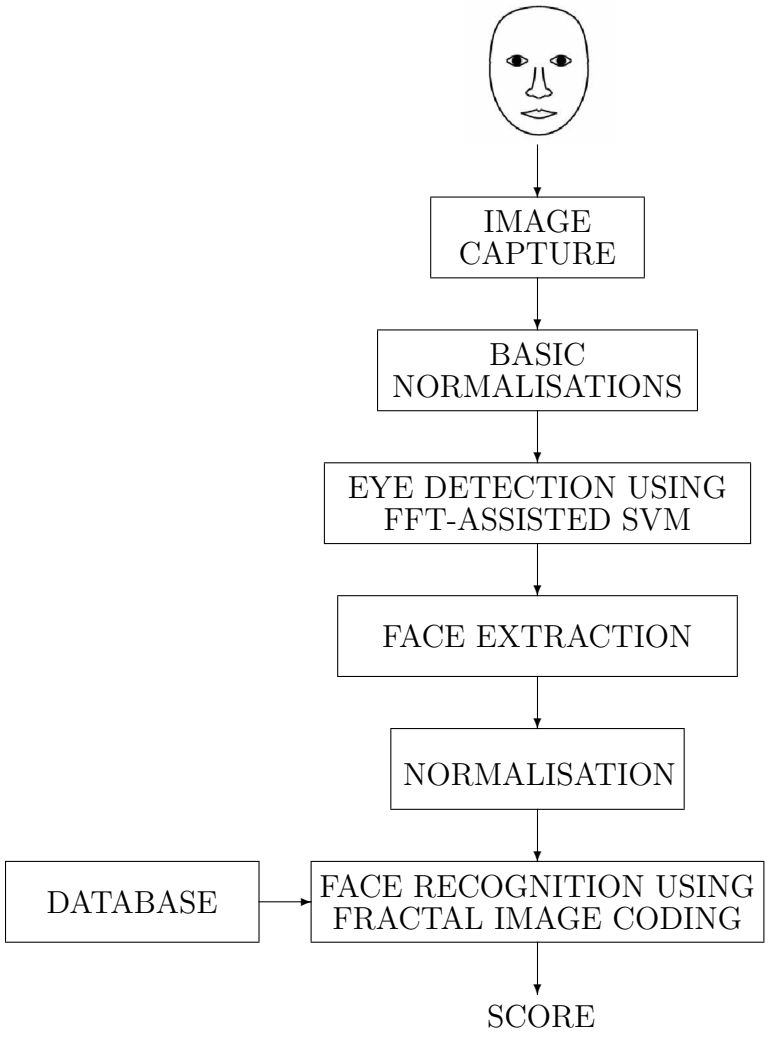

Figure 1.1: The complete face recognition system.

improvements on the evaluation set. Although results are slightly inferior on the test set, we show that potentially excellent results can be achieved by applying the client specific thresholds to this data set. The results demonstrate that the weighted FND does in fact create good class separation. The weighted FND based identification system still has short comings when some datasets are used, where its performance is not much better than the standard FND. To alleviate this problem we present a voting scheme that operates with normalised versions of the weighted FND. Although there are no improvements at lower matching ranks using this method, there are significant improvements for larger matching ranks.

Chapter 6 states the conclusions to this thesis. Future work and other applications using the methods described in this thesis are also discussed.

1.4.3 Independent Contribution

We now clarify the differences in contributions of this thesis compared to similar works in the literature. Our work is inspired by the works published by Neil et al. [NC96, NCC96] and Kouzani et al. [KHS97]. Kouzani et al. [KHS97] describe a recognition method that compares fractal codes directly using neural networks, which is published in 1997. This approach is different to ours, because we do not compare fractal codes directly. Instead, our method is based on the approach taken by Neil et al. [NC96, NCC96] where binary images are used in a fractal image coding based method for object recognition. Our method extends the ideas developed by Neil et al. [NC96, NCC96] to include the use of grayscale images. Our method is first published in an ICASSP paper in 1999 (see "Publications" at the end of this thesis). An almost identical approach is proposed independently by Kouzani et al.[KHS00] and published in the year 2000, after our initial publication. We developed our ideas independently and without collaboration with the Kouzani et al. research group. Our contributions to the method include a more thorough and methodical investigation of the mechanics of the approach. As a result, we are

able to develop more advanced variations of the basic fractal image coding based recognition method that achieve better performances.