Automatic Face Recognition using Radial Basis Function Networks

Andrew Jonathan Howell

CSRP 488

September, 1997

ISSN 1350–3162

UNIVERSITY OF

SUSSEX

AT BRIGHTON

Cognitive Science
Research Papers

Automatic Face Recognition using Radial Basis Function Networks

Andrew Jonathan Howell

Summary

It is well known that the task of automatic face recognition in dynamic environments is very hard. The key problem is that everyday influences, such as lighting, head pose and expression, can lead to greater variation between images of the same person than between images of different people. However, there must be some essential invariant set of features that allow us to recognise familiar faces. Automated face recognition systems must be robust with respect to everyday variability and capture essential similarities to identify individuals.

This thesis investigates the task of real-time face recognition within a small known group of people, using an example-based probabilistic learning scheme to learn and recognise individuals. The artificial neural network model used, the radial basis function (RBF) network, is an exceptionally fast classifier, both in training and subsequent classification phases. In addition, it provides a level of confidence in its output which allows ambiguous data to be discarded. Comparisons with other techniques using a standard database indicate the suitability of our approach.

Methods for view-based face representation are discussed and analysed, with emphasis on normalisation and preprocessing techniques. We then investigate how variations, such as pose and resolution, in face images affect recognition performance with RBF networks and explore the generalisation properties of the RBF network, looking specifically at pose, scale and shift invariance.

We present experimental work using a novel variant of the RBF network, the 'Face Unit' network, which learns to identify one particular individual. We then apply the RBF network to image sequences taken from a less 'constrained' environment to assess the suitability of the proposed approach for real-life applications. Finally, we look at the temporal learning abilities of a Time-Delay variant of the RBF network, focusing on simple behaviours based on head rotation,

Submitted for the degree of D. Phil.

University of Sussex

September, 1997

Acknowledgements

First, I would like to thank my supervisor, Hilary Buxton, for her constant advice, support and inspiration throughout the thesis.

My study for this work was made more enjoyable by the lively company and thought-provoking environment within COGS for which I am very grateful.

Special thanks must go to all at the Vision Group at Queen Mary and Westfield College, London for their invaluable discussion group, but particularly to Shaogang Gong and Stephen McKenna for their support and useful suggestions throughout the last two years. I would especially like to acknowledge their contribution in providing suitable image sequences used for part of the thesis, and the effort made by Stephen McKenna in providing positive comments on draft versions of the thesis.

Many thanks should go to David Young at COGS for his valuable help in providing POPLOG library functions for DoG and Gabor mask convolution.

Many thanks to the volunteers (victims) who allowed me to use their faces for *Sussex test data*: Rachel Bundy, Julie Coultas, Erica Morris, David Nicholson, Alex Payne, Rafael Pérez Y Pérez, Kathy Scott, Ben Shanks and Jabe Wilson, and for the *QMW image sequences*: Carla Benjamin, Gill Carter, 'J. J.' Collins, Shaogang Gong, Lorna Kyle, Stephen McKenna and Alexandra Psarrou.

For their emotional support, I would like to thank my family, Stephanie, Isadora and Rowan, without whom this would never have happened.

Financial support for this thesis was provided via a CASE studentship from the Biotechnology and Biological Sciences Research Council and GEC Marconi Ltd.

Preface

Some parts of the work reported here appeared in the following publications:

Howell, A. J., & Buxton, H. (1995a). Invariance in radial basis function neural networks in human face classification. *Neural Processing Letters*, *2*(3), 26–30.

Howell, A. J., & Buxton, H. (1995b). Receptive field functions for face recognition. In *Proceedings of 2nd International Workshop on Parallel Modelling of Neural Operators for Pattern Recognition*, pp. 83–92 Faro, Portugal. University of Algarve.

Howell, A. J., & Buxton, H. (1995c). A scaleable approach to face identification. In *Proceedings of International Conference on Artificial Neural Networks*, Vol. 2, pp. 257–262 Paris, France. EC2 & Cie.

Howell, A. J., & Buxton, H. (1996a). Face recognition using radial basis function neural networks. In Fisher, R. B., & Trucco, E. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 455–464 Edinburgh. BMVA Press.

Howell, A. J., & Buxton, H. (1996b). Towards unconstrained face recognition from image sequences. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pp. 224–229 Killington, VT. IEEE Computer Society Press.

Howell, A. J., & Buxton, H. (1997). Recognising simple behaviours using time-delay RBF networks. *Neural Processing Letters*, *5*, 97–104.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis investigates the task of automatic recognition of human faces in dynamic environments. By concentrating on face recognition, our work will cover only one of a larger set of techniques connected with the identification of people. The term 'biometrics' has come to be used for the study of automated methods for the identification or authorization of individuals using physiological or behavioral characteristics. Techniques such as speech recognition, iris scanning, hand geometry, fingerprint scanning and signature verification, as well as face recognition, can be combined to produce useful applications. In comparison to the other techniques, however, face recognition has the major advantage of being non-intrusive and requiring very little cooperation or modification of normal behaviour on the part of the subjects in order to collect useful data.

The real-life problems to be tackled here concern identifying individuals and their intentions in everyday settings, such as offices or living-rooms. The dynamic, noisy data involved in this type of task is very different to that used in typical computer vision research, where specific constraints are used to limit variation. Historically, such limitations have been essential in order to limit the computational resources required to process, store and reason about the visual data. However, enormous improvements in computers in terms of speed of processing and size of storage media, accompanied by progress in statistical techniques and neurobiology, now allow more efficient handling of such data.

The development of intelligent environments has been highlighted recently by the 'Smart Rooms' projects (Pentland, 1996) at the MIT Media Lab, which enable novel forms of interactive control for computer systems. Our particular focus within such an area is the role of adaptive learning techniques in recognising the individuals and simple movement-based gestures like head rotation. Unfortunately, the relatively unconstrained appearance of faces of individuals in video scenes makes this a particularly difficult problem.

There is great commercial interest in logging and interpretation of activity within domestic or commercial environments. Applications include access control and personalisation of domestic appliances such as computers, telephones and televisions. Burglar alarms could be improved so that they not only identify when unidentified people are in the house, but also record their activities for evidence. In addition, the logging of shoppers' interest and behaviour patterns in shops would be of interest to marketing and consumer research groups. Although this latter task does not require explicit identification of individuals, short-term memory of what individuals in a room look like could be used to connect what a particular person does for the time they remain there.

This chapter introduces the general concepts concerning our face recognition task. We first outline relevant computational approaches, then go on to discuss the particular process of recognising faces. The final section describes the specific structure of the thesis, with details of each chapter and appendix.

1.1 Computational Approaches

Many vision researchers, following Marr (1982), believed that the ultimate product of any visual system was some type of *three-dimensional* reconstruction of its environment. Although the lower levels of Marr's visual 'pipeline' scheme were clearly defined, the specific detail for higher-level processes, such as visual recognition of 3-D objects, were quite vague. This was mainly due to lack of evidence, as the computational effort required to implement full object recognition schemes was not available at that time. Once such systems using full 3-D models of objects were used to carry out useful recognition tasks, it became clear that representations simpler than full 3-D reconstruction may be more appropriate and make the task more computationally tractable.

The common tool of computer vision research is the video camera, which will only ever give a *two-dimensional* view of the world. As the input data is already 2-D, a direct technique for object recognition that takes advantage of this is to take a *view-based approach*, allowing the system to learn the task through experience. This type of recognition of 3-D objects is still difficult, as their appearance, when seen as 2-D views, will vary greatly when seen from different angles, especially when self-occlusion obscures characteristic features, or in different lighting conditions. However, it has been established that the combination of a low number of such views can be sufficient for 3-D object recognition (Ullman & Basri, 1991). Occlusion from other objects in the field of view can also add to the potential range of variation. Such view-based recognition schemes are at odds with traditional computer vision, which strive for the most accurate general techniques. We take the view that both representation and reasoning need to be tailored to the needs of a real-life task to obtain a flexibility and robustness that allow it to work in chaotic and unpredictable situations.

The view-based approach represents all potential variations in object appearance in example views within a view-sphere (see next chapter) sufficiently that test views are able to be matched to example views of the same object. If we combine view-based representations with connectionist approaches to learning, the example views can be used to train a neural network. If treated as preclassified, the training examples allow a classification to be learnt directly by the system. This type of supervised, *adaptive learning* can be characterised as function approximation. The trained neural network can then generalise to previously unseen test data to classify these examples too.

1.2 Recognising Faces

Recognising people in day-to-day life is generally effortless and unconscious. The ease with which humans manipulate such visual data makes it easy to underestimate the difficulty and complexity of such processing. The problem of automatic face recognition has stimulated lively debate and research in computer vision for many years, but it is only recently that techniques have become sufficiently robust to allow useful application systems to be developed. This is because, in reality, recognising a face poses several severe tests for any visual system, such as the high degree of similarity between different faces, the great extent to which lighting conditions and expressions can alter the face, and the large number of different views from which a face can be seen. Indeed, variations in facial appearance due to lighting, pose and expression can be greater than those due to identity (Moses et al., 1994).

As we discussed in the last section, a major distinction in object recognition is between 3-D and 2-D representations. The former, being *object-centred*, try to represent the object structure from all views, whilst the latter (which encompass the view-based representations and photometric approaches), being *viewer-centred*, try to establish characteristic or canonical views. Two-dimensional representations have been particularly popular for face representation, as faces can be treated as being approximately flat, although this severely restricts the usable pose range. It might seem more natural to form 3-D, rather than 2-D, object representations based on our apparent ability to mentally visualise 3-D manipulations of objects, but this may be a confusion of cognitive levels. Psychophysical experiments (Bülthoff & Edelman, 1992; Bülthoff et al., 1995) suggest that generalisation in recognition of unfamiliar views relies on interpolation between stored views. Although we may be

capable of some high-level mental manipulation and visualisation, it may well be that our everyday visual processing of objects is done using simpler representations and reasoning.

The issue of *invariance* has to be considered carefully for any task, though this rarely needs to be an absolute invariance. However closely in time two images of the same person are taken, there will always be some differences between them. The goal of any face recognition system, natural or artificial, is to associate some previously learnt identity with a previously encountered appearance of that same identity. The representation and similarity metric chosen, therefore, must be sufficiently invariant to changes between two images of the same person that they look more 'similar' to each other than to one of another individual. Prior knowledge about image variations likely to be useful for a task allows these variations to be explicitly highlighted within the representation during preprocessing and those that are not useful to be suppressed. Specific examples of this are shown in the requirements for our task in the next chapter.

Dynamic information is very important for biological vision, but is lost when single face images are used, although the consensus is that static images alone are sufficient for face recognition. The use of image sequences allows movement information to be extracted for gesture recognition. In addition, the use of image sequences as a data source permits a low false-positive/high discard strategy, where a large proportion of classifications are discarded to leave only those in which the system has a high level of confidence. This could be extremely helpful in chaotic or unpredictable environments, where there are likely to be ambiguous test images. For instance, in normal activity, a person may move to a variety of positions and poses, not all of which will be facing the camera. It has to be accepted that there will be instances where useful information about an individual's identity simply will not be available, and that it has to be assumed that, as a tracked body moves from an identifiable pose to an unidentifiable one, identity will maintain a coherence as long as the physical body does. In other words, that a person will not transform identity as they turn away from the camera.

Most research in automatic face recognition has been concerned with comparisons of single 'snap-shot' images. Although we are capable of recognising people in photographs, it is obvious that everyday human recognition is not carried out in this way, as it often takes a few seconds for someone coming into a room to be recognised. For applications monitoring an environment, techniques which take advantage of the abundance of information contained in sequences of images could be used to enhance face recognition performance.

An important influence in human face recognition is *context* (Young et al., 1985). Cues such as 'Who do I expect to see in the office at this time of day?' will limit the range of people we expect to recognise and their voice, posture, height, gait and how they are dressed will also greatly affect our judgement. For instance, we may not recognise someone we see every day at work if we encounter them in an unfamiliar situation, even though their facial appearance is exactly as it always is. Humans do not constantly monitor other people to assess their identity – spatial constraints in our 'world knowledge' will prevent us expecting people to swap identities. An automatic system which monitored and tracked individuals could use such prior knowledge to minimise the complexity of a recognition task and allow recognition over the whole time they are present.

1.3 Outline of the Thesis

In this thesis, the practicalities of computer-based human face recognition in domestic environments are explored. Artificial neural networks are used to learn and recognise individuals using an example-based learning scheme. Thus, wherever 'neural networks' are mentioned, it should be assumed that this refers to artificial neural networks, not any type of natural or biological neural network.

The thesis focuses on recognition techniques with segmented face images and image sequences, but prior tracking and localisation is required to create suitable input data (Appendix A describes how this has been done for the experimental data used here).

Chapter 2 describes the task of face recognition in unconstrained environments in detail and draws up specific requirements to fulfill it. Previous work on face recognition is then examined, looking at computational models and psychological and psychophysical evidence about face recognition in biological vision systems. This is followed with discussion on how task requirements affect the suitability of techniques and a direct comparison of performance and generalisation in several approaches using the same face database, using published research results and our own experimental data.

Chapters 3 to 7 give details of the five main experimental areas of research. Chapter 3 introduces our pose-varying Sussex database and discusses methods for face representation, normalisation and preprocessing techniques. Variations in face images are also studied to analyse how they affect recognition performance, with particular reference to the Euclidean distance measure for image comparisons. The contribution of the radial basis function (RBF) network is also analysed and compared with related classifiers.

Chapter 4 explores the generalisation properties of the RBF network, looking specifically at pose, scale and shift invariance. This is important, as it determines the accuracy of face segmentations required for data to be learnt or recognised.

Chapter 5 presents experimental work using a novel variant of the RBF network, the 'Face Unit' network, which learns to identify one particular individual only. This is useful for future applications as it gives an alternative, parallel method of learning tasks which can then be used as additional evidence for identity.

Chapter 6 explains how the RBF network can be applied to image sequences. The data used here was much taken from a much less 'constrained' environment than the other face recognition databases, so that the suitability of the proposed approach to real-life applications could be assessed.

Chapter 7 explores the temporal learning abilities of the RBF network. We focus on simple behaviours, based on head rotation, using a Time-Delay variant of the network to give a fast and effective classification over time within image sequences.

Chapter 8 concludes the thesis, summarising contributions to the field of automatic face recognition, and discussing directions and issues for future work.

In addition, there are three appendices, giving technical details to support the experimental work. Appendix A describes the face databases and image sequences used for the experimental sections. In addition to data collected specifically to answer questions raised in our research, we have also tested with standard data from other research groups for comparison of our approach to other previously published techniques.

Appendix B describes the specific implementation of standard RBF network used for the experiments. Details of the Face Unit and Time-Delay RBF models are included in Chapters 5 and 7 respectively.

Appendix C describes the specific implementation of the two preprocessing techniques used for tests in the thesis: single-scale Difference of Gaussians and multi-scale and orientation Gabor wavelet filtering.

Chapter 2

Background

This chapter first outlines our task requirements. We then go on to survey general theories of object recognition, including a review of psychological evidence, and computational research within face recognition from the perspective of acquisition, representation and reasoning. The final section will apply our proposed face recognition scheme to a standard database, giving comparisons with published results for other approaches.

The particular face recognition task considered here concerns a known group of people in an indoor environment such as a domestic living-room. Within such a task, it cannot be assumed that there will be clear frontal views of faces at all times. Therefore, it is important not to lose such vital information, which may only be present for a split-second if the subject is moving fast. To effectively tackle such a task requires the combination of three real-time processes: tracking of individuals as they move around the room, detection and localisation of their faces, and recognition of the final, segmented face information. Each of these three processes currently occupies a large area of research within computer vision. It is very difficult to consider an overall solution, and this thesis is mainly confined to the process of face recognition from video images, with the assumption that other processes, eg. McKenna et al. (1996), Gong et al. (1996), McKenna and Gong (1996), will provide suitable segmented face images and image sequences from our target environment.

In order for the system to be suitable for domestic environments, it needs to be as automatic and robust as possible. It cannot rely on monitoring or tuning from technical staff, nor should it constrain or require any particular actions from people in that environment. Any information collected should be from normal everyday behaviour. This is in contrast to security access control systems using face recognition, where the users are often required to stand in front of the system for several seconds in controlled lighting and pose after giving an ID card for verification. The uncontrollable nature of our potential subjects means that the system needs to be able to collect and process data as quickly as possible (close to frame rate), and make reasonable 'guesses' where information is confusing or missing.

Applications suitable for mass-market domestic use require economical solutions. Computational techniques have to be simple enough to be accomplished on standard serial processors (such as used in PCs). Data collected with a simple, fixed camera system with low-cost frame grabber must be sufficient for the recognition process. This data may need to be monochrome at present, although colour could soon be cheap enough to be used instead.

In terms of how many people the system should cope with, the maximum number distinguished does not have to be particularly high, as most family groups contain no more than 15 individuals, even including relatives. A face recognition system that could effectively discriminate a moderate number of individuals, for example around 40–50, could also be useful for monitoring other small groups, such as offices or small factories. If people are expected to stay in a room for at least a minute, for instance, a frame capture rate of 25 frames per second will provide 1500 test images for

identification purposes during that minute. It is not required that all of these are correctly classified: if we can accurately discard 1499 of them which the system finds ambiguous and correctly identify the one remaining frame, that will be sufficient data for logging the activity of that person (assuming that occlusion or multiple people do not undermine identity constancy). If the person stays longer than a minute, then hypotheses about identity can be confirmed and confidence increased.

Limiting expectations as to how many people can be expected to be recognised makes explicit how different this *low precision, high discard* task is to *high precision, low discard* tasks, such as police records analysis. The former type of task is expected to distinguish a small number of people from a large amount of potentially ambiguous data, most of which it is allowed to throw away (if a clear classification cannot be made), whereas the latter require hundreds of thousands or even millions of people to be distinguished unequivocally, using very small amounts of image data (usually single frontal and/or profile views). However, such an application would be able to take advantage of the highly constrained nature of the face images (each having fixed pose and lighting) and almost unlimited computing time (in comparison to the inter-frame period required here).

A major difficulty in tackling identification of individuals in dynamic, real-life environments is that it is not known exactly how many people there will be in the picture or whether they will be standing, sitting, etc. Additionally, even if a person can be tracked and their head localised, their face could be facing in any direction. If a simple, single fixed camera system is being used, a high-resolution wide-angle view of an average domestic room, when heads have been localised, will provide fairly low-resolution data (certainly under 100×100 pixels, and generally around 50×50), but, as we will show, this is sufficient for identification of familiar individuals.

The lighting in normal, everyday environments is obviously less controlled than that encountered in laboratory conditions (the standard for most current research face recognition systems). Not only will the number of light sources be variable (often varying from one moment to next), but they will be of different types, such as natural light from a window (varying from direct sunlight to overcast diffuse light) and spotlights, and shadows and illumination will change due to reflections and people moving around the room.

There are many types of variations in facial appearance that can occur: some arise out of the location itself, such as variable direction and contrast in lighting, and occlusion from objects and other people. Variation due to pose will occur as the subjects are free to stand in whatever position relative to the camera that they like. Day-to-day changes will be encountered after the system has learnt a person's appearance, as details such as hair styles, beards or stubble, makeup and jewellery will all vary for each individual. In addition, if daylight is present, the lighting due to this will vary according to weather and time of day. Longer-term changes will also have to be tolerated by the system (or at least the system would have to 'relearn' people periodically). These include ageing, weight change and facial changes, such as scars.

An important characteristic of the output is that it should be accompanied by a level of 'confidence' in that output, as it is essential to be aware of possible confusion in the classification process. 'Forced' classification, where a decision is given regardless of confidence, would not be appropriate or useful for our task, and the statistical validity of the approach is important so that classifications can be analysed effectively. This means that 'black box' solutions do not fare well as engineering solutions, since performance parameters will not be available, and it will not be known under what circumstances the system will be able to work.

In addition, the system ought to be capable of detecting if a viewed person is not from the group that it has learnt to identify. Such 'strangers' could then be monitored with temporary identities (to allow more than one at a time to be distinguished) which could then be assigned permanent titles if they needed to be 'remembered' for more than a short time (as determined by the user's requirements). Of course, deciding that someone is 'unknown' is a very much more difficult task than identification within a known group, where all classes will have explicit examples, as the system is trying to identify (as a member of a general group of classes, not as an individual) an almost infinite number of face classes that it has not previously seen. In a full system, a higher-level process

would be required to monitor day-to-day events and allow some behavioural reasoning to help with ambiguous data. This could allow expectations of who is likely to be present at a particular time of day, and to assess the likelihood of encountering unknown people and conduct re-learning of the database of distinct individuals (known and unknown) as required.

2.1 Task Requirements

The requirements for a useful, commercial face recognition and identity logging system for small groups of known individuals in busy, unconstrained environments, such as domestic living-rooms or offices, can be split into groups: there are *general requirements* that need to be satisfied by all parts of the system, *acquisition requirements* concerned with monitoring and extraction of useful information, *face recognition requirements* for the recognition stage and *identity requirements* which are concerned with how the recognition information is used.

1. General Requirements:
   (a) Computation involved possible on low-cost standard serial processor.
   (b) Robust performance with noisy, real-life data.
2. Acquisition Requirements:
   (a) Real-time tracking of individuals, with the ability to deal with multiple identities and occlusion.
   (b) Real-time detection and localisation of faces.
3. Face Recognition Requirements:
   (a) Fast learning and real-time recognition of faces, with a minimum of tunable parameters, of a moderate number of individuals (under 50).
   (b) Ability to work with low-resolution (under 50×50 pixels) face images, segmented from a single, wide-angle view.
   (c) Invariance to typical variations in images in such an environment, including:
      i. Minor variation in shifted position and scale of class information (in this case, faces) in the segmented image (dependant on accuracy of Requirement 2b).
      ii. Moderate variation in lighting direction, contrast, brightness and spectral composition.
      iii. Minor occlusion by another object (self-occlusion is addressed in Requirement 3(d)ii).
      iv. Any variation in background areas of image.
   (d) Invariance to typical facial variations in such an environment, including:
      i. Moderate expression variation. This would include changes due to talking, eating or chewing, etc. but not extreme facial contortions.
      ii. Head pose orientation, within a range of angles that allow some facial area to still be seen in image (for example, not the back of the head). Note that this will need to accommodate self-occlusion.
      iii. Day-to-day facial differences due to glasses, makeup, skin tones, facial hair and head hair style. Note that this too may create some self-occlusion.
      iv. Long-term, permanent facial changes due to ageing, weight change, scars, etc.

(e) Level of confidence in output available to allow discard of erratic or ambiguous data. Note this should be able to reduce 'false positive' results without creating a large proportion of 'false negatives'.

(f) Ability to detect, but not recognise, unknown individuals (that is, people from outside the learned group).

4. Identity Requirements:

   Ability to adapt the known group of individuals (the new information coming from a mechanism to handling 'strangers'), including:

   (a) Learning a new individual.

   (b) Forgetting a currently known individual.

   (c) Learning the new appearance of a currently known individual.

   (d) Identifying different types of 'strangers' (people not previously encountered):

      i. Authorized strangers, who are subsequently added to the known group and require an ID label from the user.

      ii. Authorized temporary strangers, who are, for instance, recognised for set period of time, such as the rest of the day, and then forgotten.

      iii. Unauthorized strangers, who have not been given permission to be in area.

This thesis will be primarily concerned with satisfying the requirements from Groups 1 and 3, as it will be assumed that those in Group 2 have been previously fulfilled, and suitable data provided to our system. Note that some collaborative work to establish the compatibility of the approach to state of the art tracking and localisation has taken place (McKenna & Gong, 1996, 1997). In the final chapter, techniques for tackling the Group 4 requirements will be outlined for future work.

2.2 Object Recognition

There are a number of introductions and surveys to the theoretical background of object recognition, such as Bruce and Humphreys (1994) and Ullman (1996), so it is not necessary, nor within the scope of the thesis, to reproduce such information here. We will summarise the basic common categories that have developed to describe different approaches.

First, it has to be noted that there are several levels at which an object can be 'recognised'. The most specific would be identifying a unique instance of an object, such as 'my diary', whereas the same object could also be categorised simultaneously as a type, 'a diary', and a more general category, 'a book'. Animals are often grouped into general categories based on their subparts, for instance their number of legs, but can also be seen as mammals, reptiles, etc., and this process can be carried out even if we do not know the exact species the animal belongs to. For face recognition, specifically, identification of the face object class is generally termed 'face detection' (see Section 2.3.4), whereas what is termed 'face recognition' is the discrimination of subordinate identity classes within the general face class.

Recognition will always be dependant on context and expectations, and our division of the scene into individual parts can be subjective. For instance, whether we choose to see the apple tree as a tree or a source of apples depends on whether we wish simply to navigate around it or if we are hungry.

2.2.1 Approaches

This section will examine how computational object recognition can be tackled, and the contribution different approaches can make to a successful system. It should be appreciated that faces are a fairly specific subset of all possible objects, and that the task we are tackling is not of category

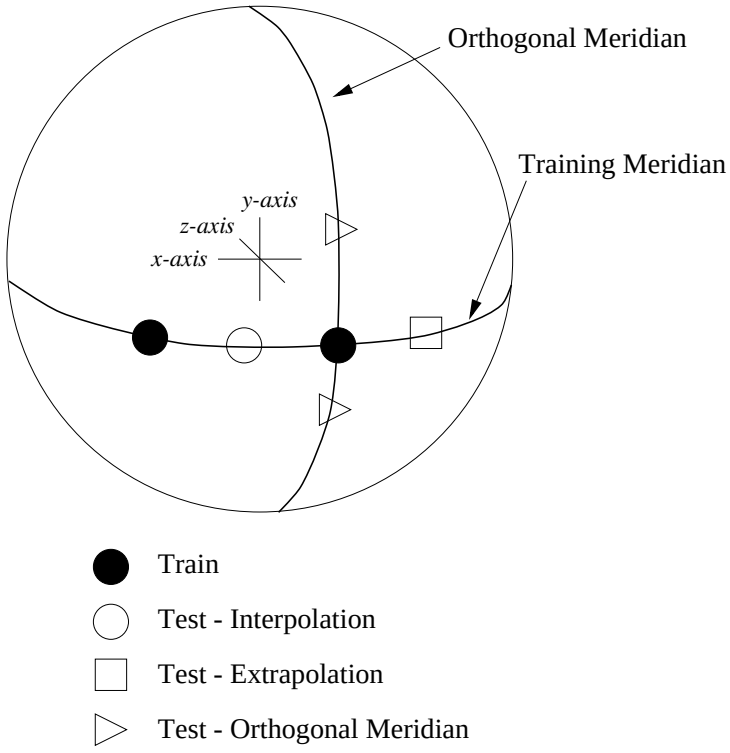


Figure 2.1: Illustration of view-sphere, demonstrating training view ranges and interpolating, extrapolation and orthogonal regions for test images in view-based recognition methods.
Adapted from Bülthoff & Edelman (1992) with permission.

classification (is it a chair or a table?), but of distinguishing very similar non-rigid, self-occluding objects (albeit from a smaller range of angles than is possible for most 3-D objects, as the face cannot be seen from the back of the head). It is assumed that all recognition requires a model which is matched to some form of representation, although this representation can be highly distributed or implicit.

Ullman (1989, 1996) divides approaches to object recognition into three broad areas: invariant properties, use of parts and structural descriptions, and alignment-based methods. The first two we will only be mentioning in passing. The use of *invariant properties* is based on the assumption that objects will have invariant properties that are common to all views of them, and was found to be useful for constrained recognition of flat, unoccluded objects. It is difficult to extend such an approach to more general applications, as it is unclear how invariant properties would be extracted from complex 3-D objects. Moses and Ullman (1992) argued that even approximate metric invariants do not exist in the general case for 3-D recognition.

The use of *parts and structural descriptions* allows a pose invariance through the use of 3-D structural graphs based on generic parts and relationships, for instance, in hierarchical arrangements of generalised cylinders (Marr & Nishihara, 1978) or geometric 'geon' primitives ('recognition by components') (Biederman, 1987). This approach is discussed in depth in Marr (1982). The use of parts is more useful for distinguishing general object classes than faces, for instance, as different facial identities will be constructed from same basic parts. A disadvantage of this approach is that it is difficult to construct 3-D models from the information present in 2-D images, due to the ambiguity of occluded surfaces. Visual systems incorporating 3-D models should exhibit complete pose invariance as any view is equally possible to compute, but this is not supported by psychophysical studies (Bülthoff & Edelman, 1992; Bülthoff et al., 1995).

Alignment-based Methods

Alignment can be approached in two distinct ways, first, there is an explicit 3-D *alignment of pictorial feature descriptions*, where potential 3-D object models are transformed to maximise the degree of match between the image generated by the transformed model and the image of the input object. The calculation of appropriate transformations is dependant on establishing correspondence between the model and the image, often using edge-based features (Ullman, 1989). The disadvantage of this approach is the difficulty of selecting features common for all views of an object (which remain unique to that object): its reliability depends on extracting sufficient features for matching.

The second alignment approach, *alignment through the combination of views*, uses a combination of 2-D views of the object as the model. Ullman and Basri (1991) were able to show that the linear combination of a small number of views was sufficient to express a wide range of views of object transformations, such as 3-D rotation, translation and scaling. The modes of variation have to be expressed though variations shown in example views, as the approach has little tolerance to orthogonal variation (see Figure 2.1). This is a view-based approach which relies on geometric relationships between the object, the image and the transformations between the two, using 2-D structural information, such as x, y coordinates and segment lengths and angles.

The general group of view-based methods encompasses other approaches that deal with photometric representations which are computed directly from image intensity values rather than trying to extract explicit structural information, although some approximate correspondences will be often established for normalisation of the image. Breuel (1992) gives an overview of the advantages of view-based recognition of 3-D objects from 2-D images over 3-D model-based approaches, citing robustness and efficiency as major issues.

Two major concepts in view-based approaches are: 1) the *canonical view*, Palmer et al. (1981) found that certain characteristic 'canonical views' allowed viewers to name an object much faster than other 'non-canonical views' and view-based methods can take advantage of such views to improve the efficiency of recognition, and 2) the *view-sphere*, which is the range of views from which the object can be seen (see Figure 2.1). A general 3-D object will have a view-sphere covering 360° movement in x-, y- and z-axes, although prior knowledge can allow this to be reduced. This is especially clear in the case of faces, as facial information is only visible on a human head from (roughly) the front $\pm 120°$ of x- and y-axis movement, and z-axis movement is physiologically constrained to around $\pm 20°$ (when standing or sitting). For example, Beymer (1994) covered a facial view sphere of $\pm 20°$ x-axis and $\pm 30°$ y-axis with 15 example views.

Appearance-based Methods

Another major approach that can be distinguished is what we have called *appearance-based methods*. This seeks to capture separately the essential visual characteristics of each object class to be recognised, and aims to create a viewpoint-independent representation. The appearance-based approach treats generalisation across views as a function approximation problem (Poggio & Girosi, 1990a, 1990b). If an assumption of smoothness is made on the function, non-linear interpolation within example views (Poggio & Edelman, 1990; Brunelli & Poggio, 1991) can be used. Partially viewpoint-invariant representations of object classes can be formed by training an RBF network to interpolate a function in the space of all possible views of the objects (Poggio & Edelman, 1990). Such representations can be seen as 'grandmother cells' (dismissed by Marr (1982) and reinstated by Edelman and Poggio (1992)!).

2.2.2 Psychological Evidence

We review psychological and psychophysical results here not to construct a 'biologically plausible' model of face recognition, but to look at issues central for 'engineering' a solution for our particular task.

Cognitive studies of the way human faces are perceived (Bruce & Young, 1986; Ellis & Young, 1989; Hay & Young, 1982; Hay et al., 1991) have contributed to our understanding of the problems for automating face processing. Psychological theories on the processing and recognition of

objects and faces seem to point to different strategies for each; for a review, see Bruce (1988), Bruce and Humphreys (1994). They suggest that the general level of object recognition may use edge-based (intensity discontinuity) information, whereas face recognition may use surface-based (texture, shading) 'holistic' information. However, if all visual recognition is seen as a spectrum from the most general to the most specific, this division can be seen as a consequence of comparing tasks from opposite ends of such a spectrum rather than some inherent difference in the nature of the recognition process. It is clear that the task of recognising radically different general categories of object, such as trees and houses, requires different techniques than distinguishing those categories that are structurally very similar, such as faces from members of a family. Within familiar face recognition, there is evidence that we may use a kind of 'face recognition unit' mechanism, where each is tuned to recognising a known individual (Bruce & Young, 1986; Bruce, 1988).

The disproportionate effect on face recognition of *inversion* (making the photographic image negative) has been taken as support for special mechanisms in face processing (Hay & Young, 1982; Bruce & Young, 1986) separate from other other types of object. The inversion effect may be due to the type of object processing required for the task. The conclusion is that it is failure of the first-order feature information to distinguish class members that leads to use of second-order configural data, and it is the latter that is susceptible to inversion interference. However, Diamond and Carey (1986) showed that this may be the effect of 'expertise', rather than something unique to human face processing, in that dog breeders who are good at identifying particular breeds of dog are more adversely affected by inversion than dog 'novices'.

The *'caricature advantage'*, where a face is recognised more efficiently from a caricature than from a veridical (undistorted) representation (noted in computer-generated line images (Rhodes & McLean, 1990) and computer-generated photographic quality images (Benson & Perrett, 1994)) also points to configural processing. In addition, Bruce and Green (1990) point out that elements of composite faces are recognised more easily in tests if misaligned than if they are correctly aligned, which suggests that when two different halves are exactly lined up, a new facial 'configuration' is created which is difficult to decompose into elements.

Within face processing research, there is also support for treating face classification as a task separate from others using facial information, such as expression interpretation. Ellis and Young (1989) describe evidence for separate mechanisms being present in human vision for facial recognition and facial expression recognition. This is shown most clearly in people with cognitive disorders such as *prosopagnosia*, where they cannot distinguish individual faces, but can usually still 'read' emotional states from expressions. Bruce and Young (1986) surveys neuropsychological disorders that indicate that familiar and unfamiliar face recognition also proceeds separately. In addition, PET brain scans (Sergent et al., 1994) have provided further evidence for independent processing.

Psychophysical studies with monkeys (Logothetis et al., 1994) suggest that they use interpolation between 2-D, viewer-centred representations for object recognition. There is some psychophysical support for the appearance-based approach (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; Sinha & Poggio, 1996), since the results suggest that only partial view-invariance over the restricted view-sphere is found for pose in human vision. In contrast, 3-D models and linear combination of views predict full view invariance over either a full or restricted view-sphere respectively.

Bachmann (1991) has been able to establish the minimum resolution required for humans to recognise facial identity in a image at around 18 pixels horizontally and 24 pixels vertically. This extremely low level of spatial information indicates that local feature processing is not being used. This confirms that there is sufficient information in low-resolution face images for effective recognition. Several researchers have found that different frequencies of facial information can be informative for different tasks. Hancock et al. (1995) found that finer scales of facial information, were related to familiarity, whilst coarser ones were related to distinctiveness. Lando and Edelman (1995) used high-frequency receptive fields for detecting viewing conditions, such as illumination and pose, and low-frequency receptive fields for face identification.

Hancock et al. (1997) has compared eigenface and graph-based systems against human re-

sponses. Whilst finding that both approaches match the human results very closely, they concluded that graph matching was closer to face processing, as PCA was affected more by image variations, such as lighting.

In summary, three main points stand out from the psychological evidence, 1) a division of strategies for entry- (general object classification) and subordinate-level (face discrimination) object recognition, 2) that appearance-based methods are likely for human vision in at least some vision tasks, and 3) that a quite coarse image resolution is adequate for familiar face recognition.

2.2.3 Discussion

Ullman (1996, pp. 34-35) stresses two important points about object recognition. First, practical research will not in general be heavily committed to specific abstract philosophies, as hybrid schemes can frequently be more effective. Secondly, the relative merits of techniques will be different depending on the target application, in other words, no single approach can be a general solution to object recognition.

We will now go on, in Sections 2.3–2.5, to isolate the important issues in face recognition and explore how they have been dealt with in previous research projects. *Acquisition* – what are the important factors in the way face information is collected? *Representation* – what is it in a face that can allow a face recognition system to remember it when it sees it again and distinguish from others? *Reasoning* – how can a face recognition system compare faces most effectively? This is followed, in Section 2.6, by a practical evaluation of some of these techniques to determine their suitability for our task requirements, as described in Section 2.1.

We will not be presenting an exhaustive discussion of all current face recognition research, as that is outside the scope of the thesis. The field of face recognition has developed quickly and even recent surveys (Samal & Iyengar, 1992; Valentin et al., 1994; Chellappa et al., 1995) are perhaps already out of date. A more representative view of current research can be found in Bichsel (1995), Essa (1996) and the Face Recognition Home Page (Kruizinga, 1995).

2.3 Face Acquisition

This section looks at how the original data is acquired before the issue of representation is raised. How many and what type of face images are needed? How much and what kind of variation should be present in the images? As we are assuming our face images are pre-segmented (Requirement 2b has been fulfilled), we are treating face detection as an acquisition issue.

2.3.1 Design of Face Databases

The environment and manner in which a database of face images is collected is vital to the success of any face recognition system in which it is used. Almost without exception, face recognition research has been carried out with highly constrained data (Robertson & Craw, 1994), with variations due to lighting, expression and pose either fixed or within unrealistic limits (if compared to variations encountered in real-life data). Many variations important to long-term applications, such as how aging will affect recognition over months and years, are not allowed for in the system design. There are some exceptions, for instance, Bouattour et al. (1992) make a conscious effort to make their data more varied, with large rotations and occlusions due to hands in front of faces. Despite these limitations, we need to use standard databases to obtain comparative results.

Two UK databases are of note. First, the Olivetti Research Laboratory (ORL) face database (Samaria & Harter, 1994) is a small database of 40 people (400 images) showing some pose, lighting and expression variation, see Section A.1 in Appendix A for more details. The usefulness of the ORL database lies in having a large number of comparative results from different groups, discussed in Section 2.6. Second, the Manchester face database (Lanitis et al., 1995a) is larger (30 people, with 690 images). The training and test data have been deliberately kept separate to prevent systems using

spurious environmental details, such as lighting or background features, to classify individuals, and have at least 3 weeks between their collection for each person to introduce more realistic variability into the data. The training images have fairly constrained lighting, whereas the test images are more variable. Two levels of difficulty of test data are present, with different levels of variability. The first is fairly easy as the images are quite similar to the training images, only allowing changes to hairstyle, background and the wearing of glasses. The second is much harder, featuring occlusion with hands, dark glasses and covered hair. Although the Manchester database tests face recognition methods more thoroughly than the ORL database, it does not have so many published comparative results.

The largest collection of face images to date is the FERET database (Phillips et al., 1996, 1997), currently still under development by the US Army Research Laboratory. This has made great advances in constructing a standard by which competing face recognition systems can be compared, and conducting independent testing of leading algorithms. They have allowed pose movements from frontal to profile and limited lighting variations within the group of images for each person, and have required that data is collected over a period of time to allow changes in the person's appearance, clothing and lighting. The database evolves from year to year, but at present, no attempt has been made to collect data as image sequences or with sound/vocal information. Its major disadvantage is its unavailability to non-US institutions.

In summary, the ORL is the most useful standard database for our purposes, due to large amounts of comparative results. However, aspects of task requirements will remain that are not covered by such tests and where this is the case, it will be necessary to design and construct our own specific databases.

2.3.2 Multi-Modal Facial Information

Cues other than facial appearance play a part in human face recognition and could prove useful for automatic techniques. Speech patterns, body shape/height and posture/gait are all characteristic and easily collected alongside facial information (Brunelli & Falavigna, 1995).

The ATR Human Information Processing Research Laboratories in Japan are constructing a joint face and speech database of 60 people (ATR, 1996), however, the main face views are taken as still images rather than as continuous sequences. The European M2VTS multimodal face database (Pigeon & Vandendorpe, 1997) has been set up to test multimodal person verification strategies (Duc et al., 1997; Kittler et al., 1997). At present, the database contains 35 people with several image sequences of each. Synchronised speech is provided with at least one sequence for each person.

2.3.3 Space-Variant Sampling

Most view-based approaches have used the rectangular aspect provided by standard video cameras, although this does not necessarily provide the most useful representation. There has been interest in space-variant pixel arrangements, such as radial, logarithmically sampled (Young, 1987) or 'foveal' representations (Tistarelli, 1994), where pixels are concentrated at the centre, and cover a progressively larger area as they spread out, as this gives a natural rotation invariance. Rao and Ballard (1995) used a foveal grid arrangement for their iconic representation scheme, centred within a rough face boundary. It is not clear how useful such an approach is, as no comparisons were made with other sampling arrangements.

The disadvantages of spatially variant representations are 1) the extra computation required to remap pixels (unless dedicated hardware is used), and 2) that peripheral detail is sacrificed (although angular extent of the field of view can be increased). This can be useful for areas of vision such as autonomous robots, for example, Cliff and Bullock (1993), where a constant level of detail is not required over the whole visual field. In the context of our particular face recognition task, however, this loss of detail is more of a problem, as the face is already localised and uniform sampling within this region seems to have very little disadvantage.

2.3.4 Face Detection and Segmentation

Detection of faces using specific facial features will not be possible for our task, due to the low resolution of the data, see Section 2.4.1. An alternative is to use the whole face pattern as a holistic representation, such as with eigenface information (Turk & Pentland, 1991; Moghaddam & Pentland, 1995). A successful neural network face detector has been developed by Rowley et al. (1996), which also used receptive fields to give some translation and scale invariance. A bootstrapping algorithm is used to get around the problem of finding suitable 'non-faces' (negative examples) to train with by incorporating initial false-positives as subsequent training data. This use of only the most confusable near-face examples, rather than a potentially huge range from the whole spectrum of 'non-faces', can substantially reduce the size of training set required for good performance compared to earlier approaches.

Face detection in image sequences is very much easier, due to motion cues, than for single images and can be integrated into tracking techniques. Once a face been found in a frame, temporal correlations greatly reduce the search space in subsequent frames, for instance, McKenna and Gong (1997) were able to combine motion detection by spatio-temporal filtering with face detection with a neural network based on Rowley et al. (1996). More recently, they have been able to use colour to further reduce computation and give greater invariance to rotations in depth and partial occlusions (McKenna et al., 1997a).

Face detection not only includes finding a face in an image, but also determines how much of the face and background is actually segmented for further testing. The approach taken to face segmentation is important when assessing performance, as transitory details, such as hair style and background details, if included in training data, may be used as the most effective distinguishing detail. For instance, if one person stands next to a plant for a picture, whilst another does not, it is very much easier to check for the presence of the plant rather than to compare subtle facial details. Some groups, such as Craw et al. (1995), ignore higher performance of experiments conducted with face images with hair included, as this face representation is not seen as being sufficiently general for images taken over time, and prefer to cite poorer results for hair-free data. There is some psychological evidence that person-specific details such as hair may be used by humans for unfamiliar face recognition (Hancock et al., 1997), however, so the visual features that are used for recognition may well be dependent on the task.

In contrast, non-person-specific details such as background are more obviously spurious for recognition. Turk and Pentland (1991) acknowledged that the background surrounding the faces in their database was a significant part of the image data used to classify the faces. Of course, this must severely limit generalisation of such an approach when it is trained with data against one background and tested with data containing a different background.

2.3.5 Normalisation and Vectorisation of Images

Once a face has been localised and segmented within an image, the image itself must be standardised or normalised prior to further processing to improve the efficiency of matching. Sometimes such normalisation is just an adjustment of grey-level intensity values, but here we are considering adjustments to the image shape. This could be as simple as a rescaling to some standard size, or as complex as remapping each pixel.

The normalisation and vectorisation of an image are approximately similar processes. Image normalisation is generally taken to be a process of adjusting to allow particular areas in different images to line up when any two images are matched together. For example, face images are very commonly normalised via affine transform on the basis of the positioning of both eyes (and sometimes mouth or nose position). This can be taken further via the 'morphing' the face texture on the basis of a larger number of standard facial landmark positions. 'Dense correspondence' is the ultimate correspondence, where all elements of the image vector correspond to pixel information from the same object feature in scene, in other words, the process creates a feature-based representation from the pixel information (in the most abstract meaning of 'feature').

Two approaches have been taken to establish the required correspondence to a reference image for image vectorisation. 1) *approximate correspondence*, either using a low number, often two or three, anchor points as features (Craw et al., 1995) or an intermediate number contained in an active shape model (Cootes & Taylor, 1992; Cootes et al., 1993; Lanitis et al., 1995a) and 2) *dense correspondence*, where each pixel is a feature point, which can be solved through optical flow algorithms (Beymer & Poggio, 1995).

The vectorised representation contains two vectors: shape and texture. The *shape vector* contains the feature coordinates, either in absolute terms (widely used, for instance (Poggio & Edelman, 1990; Ullman & Basri, 1991; Craw & Cameron, 1991; Cootes & Taylor, 1992)) or relative to a standard reference shape (Craw et al., 1995; Beymer, 1995). The *texture vector* can be the original image, geometrically normalised or warped to the standard reference shape (Craw & Cameron, 1991; Bichsel & Pentland, 1994; Craw et al., 1995; Beymer, 1995) or local texture areas (Lanitis et al., 1995a).

The disadvantage of these approaches to the normalisation of images is that they cannot be applied under wide variations in pose, and can be computationally expensive. Task Requirement 3(d)ii will not be satisfied through the use of simple 2-D affine transforms, as they treat as 2-D approximately, and anchor points will not be available at all views over large pose ranges. In addition, it is not clear how one would go about normalising a profile view to match a frontal view. Our approach to tackling such problems is to use a 'nose-centering' technique, where the face images are centred on the tip of the nose, so that visible features on profiles, for instance, should be in roughly similar locations to those in frontal views of the same person.

2.3.6 Discussion

The publically available standard face databases are too constrained to be useful for testing real-life applications, and comparisons between techniques are therefore unreliable due to wide variations in databases used in published results. Therefore, we will start off our experimental work by testing our proposed approach with the standard ORL database (in Section 2.6), but then in the following chapters, go on to construct our own databases to specifically test pose variations and image sequence data.

As mentioned earlier, we will be using manually located faces for our initial studies, but will be taking advantage of state of the art tracking and localisation techniques (McKenna & Gong, 1996, 1997) for our work with image sequences. Segmentation and normalisation over large pose ranges is still an open issue, and we will investigate this later.

2.4 Face Representation

For a face recognition system to perform effectively, it is important to isolate and extract the salient features in the input data to represent the face in the most efficient way. The abstract elements of such a representation can be made up in a variety of ways, and it depends on the task which approach will be appropriate.

One of the main problems in computer vision, especially in face recognition, is dimensionality reduction to remove much of the redundant information in the original images. Simple mechanisms, such as sub-sampling, may give a rough reduction, but use of more specific prior knowledge to apply more sophisticated preprocessing techniques to an image is still required for the best results.

2.4.1 Simple Feature- and Template-Based Approaches

The feature-based approach requires the detection and measurement of salient facial points (see Samal and Iyengar (1992) for a survey). Kanade (1973) used geometrical distances and angles between primary facial features such as eyes, nose and mouth to classify faces using an economic representation of the face where the elements are based on their relative positions and sizes. A

disadvantage of such picture-plane measurements is that it is not obvious which features and configural information will categorise a face efficiently and accurately (especially if shape and texture variations are considered as part of a useful facial feature set), and so important data may be lost.

Automatic feature finding algorithms (Bennett & Craw, 1991; Craw et al., 1992; Brunelli & Poggio, 1992a, 1993) have been developed to locate facial 'key points'. However, this information has been more usefully used for normalisation through transformations prior to recognition (Craw & Cameron, 1991, 1992) than for identification itself. A problem in using such techniques for our task is that the low resolution data which will be available would make the accurate identification and positioning of small facial areas very difficult, if not impossible.

Template matching, involves the use of pixel intensity information, either as original grey-level or processed to highlight specific aspects of the data. The templates can either be the entire face or regions corresponding to general feature locations, such as eyes or mouth. Cross-correlation of test images with all training images is used to identify the best match. Brunelli and Poggio (1993) compared feature and template-based methods directly with the same database of frontal face views. Their template-matching strategy was based on the earlier work of Baron (1981), except that they automatically detected and used feature-based templates of the mouth, eyes and nose, in addition to whole face templates. These additional feature templates as well as the whole face image were used to give better performance. Geometrical alignment of the eyes to match test images with model views allows shift, scale and rotation normalisation prior to the recognition process itself.

The use of raw pixel intensity values will make the representation very intolerant to lighting conditions and variability, so Brunelli and Poggio (1993) compared several preprocessing techniques: none (plain grey-level values), intensity normalisation (using a neighbourhood average value) and the use of a gradient norm operator, which he found gave the highest recognition performance.

In summary, the simple use of templates or features will not be enough for real-life applications, but additional processes such as alignment (Beymer, 1994; Lando & Edelman, 1995) or filtering may be able to improve on this (see Section 2.5.1).

2.4.2 Deformable Templates and Active Shape Models

A priori knowledge of face variations and the expected shape of geometric features can be used to construct deformable (flexible) templates (Yuille, 1991; Yuille et al., 1992) to guide feature detection process. Size and shape parameters in such templates can be translated, rotated and deformed to fit the best representation of their shape present in the image and these variations give a feature description, allowing both detection and representation. Unfortunately, such approaches are critically dependant on appropriate starting positions for the template, and computationally expensive (5–10 minutes of Sun 4 CPU time to match one image was quoted). This use of hand-crafted templates, individually tailored for specific tasks, has been replaced by flexible shape models using a Point Distribution Model (PDM) (Cootes et al., 1992, 1993) which are learnt from examples. This, in turn, can be used together with a shape-free grey-level model, obtained by deforming and aligning each training face to the mean face, to give a combined face encoding scheme (Lanitis et al., 1995b).

Statistical methods are useful for modelling shape and grey-level appearance of images, as they can give a compact encoding of permitted variability (Lanitis et al., 1995a; Vetter & Poggio, 1996). Models containing prior structural knowledge of faces are learnt from a database of prototypical images. Such models can build flexible object representations (active shape models) (Lanitis et al., 1995b, 1997) through a linear combination of labelled examples, which can then be iteratively deformed to match image data. This requires under 100 parameters to describe each image with expression, lighting and limited pose ($\pm 15°$) invariance to produce the PDM representation, and has been used for tracking in image sequences (Edwards et al., 1996, 1997).

In summary, the deformable templates are computationally expensive and not robust to everyday variation. Both they and the simpler active shape models will have problems establishing matches for model points on low resolution images, such as provided in our task requirements.

2.4.3 Principal Components Analysis and 'Eigenfaces'

Principal components analysis (PCA), is a simple statistical dimensionality reducing technique that has perhaps become the most popular for face recognition. PCA, via the Kahunen-Loève transform, can extract the most statistically significant information for a set of images as a set of eigenvectors (Kirby & Sirovich, 1990) (usually called 'eigenfaces' (Turk & Pentland, 1991) when applied to faces), which can be used both to recognise and reconstruct face images. Eigenvectors can be regarded as a set of generalised features which characterise the image variations in the database. Once the face images are normalised for eye position, they can be treated as 1-D arrays of pixel values. Each image has an exact representation via a linear combination of these eigenvectors and an arbitrarily close approximation using the most significant eigenvectors (that is, those with the highest eigenvalues). The number of eigenvectors chosen determines the dimensionality of 'face space', and new images can be classified by a projection onto that face space. For example, Kirby and Sirovich (1990) chose the 50 most significant eigenvectors. Cottrell et al. (1987) and Fleming and Cottrell (1990) compressed face images using a simple neural network, the weights and hidden unit activations representing eigenvectors and eigenvalues, and moderate success was made in recognising novel images.

Comparisons can be made between pure image-based coding, which is effectively template matching with position and scale differences eliminated, and more extensive normalisations in which more shape variability was removed. Craw and Cameron (1991) morphed faces to an average shape before applying PCA, as the 'shape-free' images give a more linear space for analysis. Such normalisation of faces before extraction of eigenfaces is based on an assumption that faces lie within a low-dimensional manifold, linearly approximated by independent shape and shape-free texture. An eigenface coding of shape-free texture with manually coded landmarks has been found to be more effective for automatic recognition than correctly shaped faces, giving a higher-quality representation of the images in terms of facial variation (Craw et al., 1995). Although earlier work concentrated only on frontal views, Pentland et al. (1994) extended this to encode wide pose ranges, both parametrically (Murase & Nayar, 1995) (PCA calculated for all views together) and as modular view-spaces (PCA calculated separately for each view). Pentland et al. (1994) found a slight advantage for the latter approach. Assumptions have to be made about the suitability of the data before PCA is applied, hence the emphasis on normalisation. Akamatsu et al. (1992) used data in the Fourier domain to gain shift invariance in subsequent PCA. Oriented Difference of Gaussians convolution (Hancock et al., 1995) and Gabor wavelet transform (McKenna et al., 1996) have also been performed before PCA to provide a greater level of invariance than found using grey-level pixel information.

In summary, PCA is a very efficient signal encoder, and designed specifically to characterise and encode variations rather than ignore them. Thus, it may find the optimal low-dimensional representation, but this may be more useful for reconstruction rather than recognition (O'Toole et al., 1993). In addition, the eigenface method is not invariant to image transformations such as scaling, shift or rotation in its original form and requires complete re-learning of the training data to add new individuals to the database. Instead, we prefer to overcome both image variation and the problem of picking out important information using receptive field functions and adaptive learning.

2.4.4 Receptive Field-Based Approaches

The receptive field (RF) of a visual neuron is the area of the visual field (image) where the stimulus can influence its response. For the different classes of these neurons, a receptive field function $f(x,y)$ can be defined. Precomputed filters can simulate such fields when applied to locations across the image. This type of preprocessing is more biologically motivated than simple edge detectors or intensity normalisation, as there is psychophysical and physiological evidence for orientation and spatial frequency specific channels in biological visual systems (Daugman, 1988).

Gaussian Receptive Fields

A simple dimensionality reduction strategy is to use receptive field (RF) responses. Edelman et al. (1992) used the responses of 75 asymmetrically-positioned oriented Gaussian RFs arranged around the image as input for the RBF classifier system which learnt from examples. Lando and Edelman (1995) were able to use a similar arrangement to generalise from a single view of a face, using high-frequency RFs for detecting viewing conditions, such as illumination and pose, and low-frequency RFs for face identification. However, such simplistic filtering may not be making the input representation explicit enough. A Gaussian function will smooth the image at a given frequency, so it is good for removing noise. However, the disadvantages inherent to using raw pixel values will apply here, such as low tolerance to lighting conditions.

Natural Basis Functions

Rao and Ballard (1995) used the dimensionality reducing properties of PCA via a fixed set of learned basis functions extracted from natural scenes, which appear to match V1 simple cell responses quite closely. Earlier, Hancock et al. (1992) had found that the eigenvectors of patches of real-world images were close approximations of derivative of Gaussian filters. These filters can be applied at different orientations and scales to provide feature jets, similar to Gabor jets (see below). The advantage of this approach over simple PCA on the dataset itself is that the basis functions do not need to be recalculated to accommodate new faces.

Laplacian/Difference of Gaussians

Retinal ganglion cells and lateral geniculate cells, early in biological visual processing, have receptive fields very similar to the Laplacian operator. This can be implemented as Difference of Gaussian (DoG) filters (Marr & Hildreth, 1980) which combine edge boundary detection with Gaussian smoothing. The output of this process is then in a suitable form for detecting *zero-crossings* – locations where the second derivative of the intensity values in the image undergo a sign change, such as used in the *primal sketch* representation (Marr, 1982), which can be useful for object segmentation. The idea of DoG-style *valley-detecting* convolution, where the 'width' scale is adjusted to be sensitive to face-sized features, has been proposed by Bruce (1988) as being particularly useful for face processing. Scales and orientations can be introduced into the filtering process; for example, the Cresceptron network (Weng et al., 1993) used 8 directions of oriented zero-crossings at 2 scales for input representation.

The idea of edge information as a basic object representation is common, either unoriented as DoG filters, or oriented (Ballard & Rao, 1994) (steerable filters) to give more specific information. Kanade (1973) applied a Laplacian to binarize the grey-scale values, but then used projection analysis to extract feature information. However, specific positions of 'edges' may be too precise for generalisation, as matching will be 'brittle'. Edelman et al. (1992) found edge magnitude values from standard edge detection algorithms, such as the Canny operator, actually reduced performance when distinguishing faces. They thought such precise operations made generalisation under pose and lighting variation difficult, and found a directional derivative more useful than either raw intensity values or intensity gradient magnitudes. For this reason, we use binarised gradient information rather than zero-crossings contours in our work (see Section C.1.3 for details).

Gabor Wavelets

The receptive fields of the simple cells in the primary visual cortex (V1) of mammals are oriented and have characteristic spatial frequencies. Daugman (1988) proposed that these could be modelled as complex 2-D Gabor filters, which have been found to be efficient in reducing image redundancy and robust to noise (Bossomaier, 1989). Such filters can be either either convolved (Petkov et al., 1993) or applied to a limited range of positions, such as for 'jets' (Daugman, 1988; Manjunath et al., 1992; Würtz, 1994; Konen & Schulze-Krüger, 1995), where a region around a pixel is described by the responses of a set of Gabor filters of different frequencies and orientations, all centred on that pixel position.

Petkov et al. (1993) implemented a face recognition scheme based on Gabor wavelet input

representations to imitate the human vision system. Unlike most preprocessing techniques which try to reduce the amount of input data, the full convolution of the face image with a set of 64 Gabor functions (8 orientations and 8 scales) gives a very much larger representation than the original grey-level image. No learning algorithm was used to train the system: the test for recognition was based on simple comparisons of each image with all of the other images in the database. A successful match was where the closest (in terms of Gabor coefficients) to the test image was of the same person as the test image. The calculation of the Gabor coefficients as a complete convolution (rather than sparse sampling) of the image was reported as extremely computationally expensive, processing of the 64 Gabor functions for a single image taking about half an hour on a fast workstation. The approach was extended to use a Kohonen-style self-organising network used as classifier (Petkov, 1995). Gabor coefficients can, in addition, be used as data for PCA to provide a greater level of illumination invariance than found using grey-level pixel information (McKenna et al., 1996).

Summary

Filter-based preprocessing of the images is an important intermediate step in image-based techniques, as the input representation contributes a great deal to the learnability of the task. It is important to highlight relevant parts of the information (leading to reduction in the dimensionality of input) and provide moderate invariance to normal environmental illumination (Marr & Hildreth, 1980). This is in contrast to tackling strong, incidental lighting, which is very much more difficult (Moses et al., 1994), but luckily not expected in domestic environments. The approach can both suppress variation that is not important for the task, such as illumination variability, and highlight those variations that are useful, via, for example, orientations and scales used for Gabor filters.

2.4.5 Dynamic Link Graphs

The dynamic link approach to object recognition can be seen in two lights. First, as a theoretical model of biological vision (Lades et al., 1993; Wiskott & von der Malsburg, 1996), and second, as an algorithmic form which has been shown to perform extremely well on the standard databases (Wiskott et al., 1997). The important features of the approach are labelled graphs containing layers of Gabor feature jets and the dynamic links within the graphs that establish the image/model correspondence match. The process can be extremely computationally expensive, taking 10-15 minutes of SPARC 10 CPU time to recognise one face from a gallery of 111 models (Wiskott & von der Malsburg, 1996).

Objects can be described by both shape and texture information using elastic graphs (Manjunath et al., 1992; Würtz, 1994; Konen & Schulze-Krüger, 1995) of local features. This process uses a rectangular graph laid over the training images, the graph edges represent the distances between features (the geometric data), and the graph vertices hold coefficients from Gabor 'jets' (see Section 2.4.4) applied to the image at the feature locations. An alternative to the rectangular grid is to use manually constructed 'face bunch graphs' that are specific to faces, using fiducial landmarks, such as eyes, mouth, etc. (Wiskott et al., 1997). A coarse match of the graph onto the test image is made first with fixed parameters, followed by finer matching using a cost function to offset graph distortion against object distortion. This approach has some similarities to flexible templates (Section 2.4.2), as the matching algorithm is in terms of geometrical deformation and similarity of Gabor coefficients.

Some pose invariance for the elastic graph models can be gained by global transformations to the feature jets to account for changes in view (Maurer & von der Malsburg, 1995a, 1995b). This accounted for rotation up to half-profile (45°), but separate, manually-designed face grids and graphs have to be used to cope with self-occlusion at greater pose ranges (Krüger et al., 1996; Wiskott et al., 1997). Matching times for a single image for this type of approach was reported as 15–20 seconds (Maurer & von der Malsburg, 1995a).

However, these highly specialised representations clearly illustrate that the boundary between representational issues and reasoning issues is hard to define. They seem too committed and computationally expensive for our purposes.

2.4.6 Discussion

In general, approaches relying on simple templates or features alone will not be sufficiently robust under pose and lighting variations for our requirements, especially as the extraction of common features under all poses will be hindered by self-occlusion. This means that a photometric view-based or appearance-based representation is likely to be the most useful for our task.

Methods that rely on locating specific facial features, either for classification or normalisation, may turn out not to be efficient when applied to low resolution images. As mentioned earlier, although standardising face images (especially to an average shape) can be an extremely efficient representation for frontal views, it is not clear how such a process could be carried out for large pose variations, as there are no common features for all facial views.

PCA is an efficient way of reducing dimensionality, but has the drawback of being more sensitive to image variations than to facial characteristics. Its performance is dependent on the accuracy of normalisation, and the process has no inherent invariance to translation, scale or rotation. Lighting can severely disrupt matching (Hancock et al., 1997), and although pose can be dealt with, it cannot be accommodated easily.

Other representations approaches to improving generalisation through learning other aspects of the task are possible, such as low-dimensional object representations from examples (Edelman & Intrator, 1997), class-based image transformations (Lando & Edelman, 1995), or specific invariances (Simard et al., 1992), but their computational expense made them unattractive for our specific task.

The way representations are devised is primarily led by the need to reduce dimensionality to reduce complexity and computation. There is an implicit assumption that much of image data is redundant or irrelevant. Obviously, the dimensions discarded should be from this category in order to emphasise the useful data left over. This is the major reason for using filter-based representations, as one can specify the nature of feature that should be extracted. This has been observed in biological systems, where parallel processes can deal with different aspects of the images which were specifically extracted at at an early stage.

We will take a filter-based approach, which is fast and yet fairly general. We regard this filtering as an early stage of representation for identity, which we will develop further using adaptive learning in the next section about reasoning. Finally, in Section 2.6 we provide a comparison of other techniques to our proposed approach with a standard database.

2.5 Face Reasoning

Once a database has been collected and a representation decided upon for the images, the method of comparison between exemplar and test faces has to be determined. This reasoning can be simple matching if the representation extracted is extremely face specific or can be be very adaptive if a more generalised representation (not very discriminable) is chosen. It can be seen that the type of representation has determined a 'face-space' in which distance comparisons can be made. Standard distance metrics, such as Euclidean or Mahalanobis (for eigenvector spaces) (Craw et al., 1995), can be used for matching, whereas simple weighted sums may be more suitable for internal 'hidden' representations.

Learning is an important factor in any useful application, to avoid the 'brittleness' commonly found in manually extracted rule systems. Even simple vision tasks are of such complexity that original assumptions in manual systems turn out not to be valid or only partially valid in certain circumstances. In addition, such an approach is neither scalable nor modifiable in day-to-day operation. For example, if the task changes from the original specification due to different people or rooms being involved, the system should be able to automatically relearn the task, rather than require an operator to reprogram new rules to cover the changed circumstances.

2.5.1 Matching Techniques

Traditionally, matching has been found very useful in low-level vision for localising and identifying patterns, based on simple correlation of the image vectors. Such simple approaches were discussed in Section 2.4.1.

A different approach to matching is taken by Beymer (1994) who uses examples of faces in varying pose to learn a pose-invariant face description in terms of shape and texture vectors. This is essentially an alignment-based extension to traditional template-matching approaches (Baron, 1981; Brunelli & Poggio, 1993), but the model can solve the correspondence problem between face images in different poses, which can be used both for face image analysis and synthesis. Two methods were developed, an interpolating multiple view recogniser, and a virtual views approach. For the latter, Beymer and Poggio (1995) described how such synthetic views could complete a view-sphere for example-based learning where insufficient real views were available. Affine transform and optical flow were used to bring image templates into registration (they termed this process 'vectorisation'), and normalised correlation determined the best match. Despite good results, this approach is fairly slow and can take several seconds of processing time per test image.

Beymer (1994) used simple normalised correlation with example templates of eyes, nose and mouth, following a two-stage geometrical registration step. This was originally done with 15 example views of each person to be recognised, but this was adapted to work with virtual views. This 'analysis through synthesis' approach, recognising faces from one original and several, synthetic views (Beymer & Poggio, 1995; Ezzat & Poggio, 1996) is extremely useful where data is very sparse. This is, however, extremely intensive computationally, taking up to half an hour to analyse one image, and therefore not applicable to our task. This low data, high computation is quite the opposite to our high data, low computation task requirements.

2.5.2 Early Connectionist Approaches

Neural networks have a long history of being used for face recognition, though computational limitations of the time seem to have restricted the amount of testing that was possible.

The Kohonen associative networks (Kohonen et al., 1981; Kohonen, 1989) were able to demonstrate quite early on one of the main advantages of the distributed processing in neural networks, which is a tolerance to noisy or incomplete test data. They could classify grey-level images of faces when a forcing stimulus (the desired output activity) was provided along with the stimulus pattern (the input data). These values were clamped until a steady state of activations was reached. The idea was that, when unclamped, the network would converge when given the original input to give the desired output values. It could also generalise in classifying new views of learnt faces by interpolating within the range of angles already seen, but could not extrapolate to images outside this area. Millward and O'Toole (1986) used a Kohonen memory model to encode zero-crossing edge segments rather than grey-level values. The results, though better, are difficult to assess, as a greater pixel area was used to extract edge segments than was used for the pixel intensity values (Bruce, 1988, p. 107).

WISARD is a pattern classifier system that uses a neural network-like approach. It has a single layer local adaptive network with an n-tuple selection mechanism which is used to recognise human faces and expressions, and is able to distinguish between smiling and frowning faces (Stonham, 1986). WISARD was trained with many binary exemplar images of each face, input to the system from real-time video until a sufficiently high recall was achieved. This gave reasonable results, although it was intolerant to scale, 3-D rotation, and lighting or background variation. WISARD was not a distributed model, as trained concepts (individual faces) were held locally. The output for a particular image was a numerical representation of the detector responses, rather than a classification against trained input. This could form a personal identification code, either for confirmation of known faces or for matching instances of unknown faces. This approach has been used again for face recognition recently (Lucas, 1997), see Section 2.6.

Multi-layer Perceptrons and Associative Networks

The Multi-layer Perceptron (MLP), commonly trained using gradient descent with error back-propagation, is capable of good generalisation for difficult problems, but is notoriously difficult to ensure global convergence under all training runs, as the non-linearity of the hidden units and the nature of the input-output mapping lead to a large number of local minima, and training times can typically be long. Cottrell et al. (1987), Fleming and Cottrell (1990) used multi-layer networks with target output equal to input (auto-association) in order to compress photographic images. The network was trained on random patches of image. The compressed signal could be taken from the hidden layer of units (these values were effectively eigenvalues, the eigenvectors, called 'holons' here, being contained in the weight values between the unit layers), and these values could, in turn, be put back in to decode or uncompress the original image as output values.

Cottrell et al. (1987) found that the non-linear arrangement of their multi-layer network did not actually improve the compression of images when compared to networks using linear units. For this reason, all following networks used for PCA, such as Turk and Pentland (1991) for instance, have used simpler linear associative networks. However, Valentin et al. (1994) suggests that while linear associative networks and MLPs using back-propagation which calculate PCA can be effective for single-viewed classification tasks, they may not be as effective as HyperBF networks (Poggio & Edelman, 1990; Brunelli & Poggio, 1991) in a nonlinear mapping task, for example the classification of people with varying head pose (see Section 2.5.4).

2.5.3 Hierarchical Neural Networks

The Cognitron (Fukushima, 1975) and Neocognitron (Fukushima, 1988) were biologically-inspired self-organising hierarchical approach to object recognition. The neural network structure had successive layers of cells of increasingly large receptive fields with a cascaded grouping of features, which allowed it to become invariant to scale, rotation, and translation. The recognition of analogue input has been developed by Ting and Chuang (1993), but training still requires binary patterns, and the approach has only been used on 2-D objects, such as numerals and digits, so it is not known how such an approach would behave with 3-D variations.

The Cresceptron (Weng et al., 1993) had a similar retinotopic structure to the Neocognitron, but differed in that its configuration could be automatically determined during learning. The higher-level layers of units can be regarded as increasingly complex receptive fields, in that they become more and more specific to the training objects. Information can be 'grown' incrementally, with new network units being added as new concepts are detected. It was trained on complex images containing faces from TV news programs, and appeared quite robust to expression and minor pose variation (greater pose ranges could be explicitly learnt as different instances of the same object). However, the approach is computationally expensive and has not been tested with large numbers of objects or under large image variations, such as illumination.

Neurophysiological evidence has come from Perrett et al. (1989) for image-based coding in face-sensitive neurons in the macaque STS area of the brain, showing a viewer-based, rather than object-based, representation for faces. The view-invariance seen in some of the face cells has been supported by work on high order cortical sensory areas by Rolls (1994). Wallis and Rolls (1997) have also created a neural network simulation, 'VisNet', for learning spatio-temporally invariant object representations based on observed responses of temporal cortical visual neurons. Hierarchical layers of competitive networks are used, with short range mutual inhibition within each layer. This multi-stage feed-forward architecture was able to learn invariant representations of objects, including faces. A wide range of invariances have been observed, including spatial-temporal, translation and view, using a modified Hebb-style training rule incorporating a temporal 'trace' of each cell's previous activity (Wallis et al., 1993). This approach is useful for simulation purposes, but its complexity would not make it suitable currently for real-time applications.

The hierarchical style of network structure has had considerable success in overcoming rotation and scale differences, but this type of processing requires considerable computational effort even

to train with small amounts of data, due to the large numbers of layers the information has to passed through. It is clear that 3-D objects can be invariantly represented in such structures (Rolls, 1995; Wallis & Rolls, 1997), but at present the computational load precludes them from real-time applications. They would be suitable for a parallel process, but the specialised hardware required would exclude them from task suitability this time through cost (Task Requirement 1a).

2.5.4 Radial Basis Function Networks

One can implicitly model a view-based recognition task using linear combinations of 2-D views (Ullman & Basri, 1991) to represent any 2-D view of an object. A simpler approach is for the system to use view interpolation techniques (Poggio & Edelman, 1990; Brunelli & Poggio, 1991) to learn the task explicitly. Radial basis function (RBF) neural networks have been identified as valuable adaptive learning model by a wide range of researchers (Moody & Darken, 1988; Broomhead & Lowe, 1988; Poggio & Girosi, 1990b; Musavi et al., 1992; Ahmad & Tresp, 1993; Bishop, 1995) for such tasks. Their main advantages are computational simplicity, supported by well-developed mathematical theory, and robust generalisation, powerful enough for real-time real-life tasks (Pomerleau, 1989; Rosenblum & Davis, 1996). They are seen as ideal for practical vision applications by Girosi (1992) as they are good at handling sparse, high-dimensional data and because they use approximation to handle noisy, real-life data. The nonlinear decision boundaries of the RBF network make it better in general for function approximation than the hyperplanes created by the multi-layer perceptron (MLP) with sigmoid units (Poggio & Girosi, 1990b), and they provide a guaranteed, globally optimal solution via simple, linear optimisation. The RBF network is a poor extrapolator (compared to the MLP) and this behaviour can give it useful low false-positive rates in classification problems. This is because its basis functions cover only small localised regions, unlike sigmoidal basis functions which are nonzero over an infinitely large region of the input space.

Regularisation Networks are based on mathematical regularisation theory and include RBF and HyperBF (HBF) networks in configurations where the networks have an equal number of hidden units and training examples (Girosi et al., 1995). They can be seen as performing generalisation through non-linear view approximation (Bülthoff & Edelman, 1992), which has the advantage over linear interpolation (linear combination of views) (Ullman & Basri, 1991) in that it is less affected by variation orthogonal to learnt variation, see Figure 2.1. The RBF network can be considered as a special case of the more general HBF network (Poggio & Girosi, 1990b).

Once training examples have been collected as input-output pairs, that is, with the target class attached to each image, tasks can be simply learnt directly by the system. This type of supervised learning can be seen in mathematical terms as approximating a multivariate function, so that estimations of function values can be made for previously unseen test data where actual values are not known. This process can be undertaken by the RBF network using a linear combination of basis functions, one for every training example, because of the smoothness of the manifold formed by the example views of objects in a space of all possible views of that object (Poggio & Edelman, 1990).

Although Brunelli and Poggio (1992a) used a simple nearest neighbour classifier to discriminate feature vectors, their success with their HBF networks for object recognition (Brunelli & Poggio, 1991) led them to conclude that an HBF network would be a more effective solution to their template matching scheme (Brunelli & Poggio, 1993) for face recognition. Template matching is related to RBF and HBF Network schemes, with the difference that Gaussian, non-linear functions are applied to the correlation coefficients. The HBF network allows the use of non-radial basis functions and may find a more optimal solution than the RBF (Brunelli & Poggio, 1991), as more precision is available in the choice of basis function. They are less attractive for real-time applications, however, as the calculations for the higher-order centre functions are computationally more intensive than the simple Gaussian function used by the RBF network. The ability of such networks to train according to very specific tasks is shown by Brunelli and Poggio (1992b), where the HyperBF architecture was used to identity gender information from geometrical descriptions

very similar to those used in Brunelli and Poggio (1991).

Ahmad and Tresp (1993) trained a variety of nets to recognise stationary hand gestures from computer-generated 2-D polar coordinates of fingertips (not actual images). They achieved good generalisation in 3-D orientation and their system was able to cope well even when much of the data was missing. Their standard test data was best handled by a back-propagation net, but this performed badly with missing or uncertain (noisy) features, suffering a serious fall-off in performance as more elements were lost. They show, however, that a Gaussian RBF net can cope well with this type of data, with a success rate over 90% even with 50% of the features missing. This indicates that the RBF network would be suitable for learning the 3-D transformations and occlusion found in faces under large variations in head pose.

Lando and Edelman (1995) used RBF networks in two separate stages, to capture pose and lighting parameters (using high-frequency filters) and to classify individuals (using low-frequency filters). In between the two networks, a face class specific transformation was applied to the original image (using parameters from the first network) to align the test image with a single standard view for all trained identities (discriminated by the second network). This is related to the analysis by synthesis approach (Beymer & Poggio, 1995), in that there is only one training prototype of each class. The difference is that the method attempts to transform the test image to the single canonical view, rather than relying on interpolation between several views. Of course, this is not using the valuable view interpolation ability of the RBF network, and it is not clear how intensive such transformation are computationally.

A major advantage of the RBF over other network models, such as the MLP, is that a direct level of confidence is reflected in the level of each output unit. This is because regions in input space that are far from training vectors are always mapped to low values due to the local nature of the hidden units receptive fields, so that 'novel' input will give a low activation. This is in contrast to the global function approximation of the sigmoidal hidden units in the MLP, which can have spurious high output in similar regions of input space, allowing high confidence output. In addition, the normalisation of RBF hidden unit activities allow their output to represent probability values for the presence of their class (Moody & Darken, 1989). In light of the probabilistic nature of the RBF network's output, we will be using a discard measure in our work to exclude low-confidence output and reduce false positives.

2.5.5 Committees and Ensemble-based Networks

Committees of networks can be used to give a consensus opinion where each network is trained with different parameters or data examples. The combination of results from all the networks may be better than the use of the one that works best on test data, which may not generalise most efficiently (Bishop, 1995).

An ensemble network scheme was used by Edelman et al. (1992), who had a series of RBF networks, one for each person each trained on several images, with single output units. Output from each network was combined and used as input for a second stage ensemble RBF network which coordinated a final 'winner take all' classification. Each network only had one output, which signified the strength of classification for a particular individual.

An ensemble of RBF networks was used by Gutta et al. (1995) to identify faces, each network in the ensemble using different numbers of clusters and amounts of overlap. The number of hidden units was not related to training examples, so training is more computationally intensive than approaches using one unit per example, as extra effort is required to cluster the centre vectors. However, subsequent classification may be less intensive, as the system will have fewer hidden units. Although the FERET database used for testing includes a wide range of different pose views, only results for the frontal views were presented. The system has been updated (Gutta & Wechsler, 1997) to use a decision tree to coordinate the output from the ensemble for a contents-based image retrieval task. The decision tree component improved performance, but it is possible that a coordinating RBF network (such as used by Edelman et al. (1992)) could achieve similar results more

efficiently.

The interactive activation and competition (IAC) network model (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) has an ensemble style of organisation which can be used to account for psychological phenomena in face recognition (Burton, Bruce, & Johnston, 1990), based on the Bruce and Young (1986) face recognition perceptual framework model. 'Pools' of units, representing face recognition units (FRUs), person identity nodes (PINs) and semantic information units (SIUs), have inhibitory connections between their constituent units. These units then have excitatory links with specific units in other pools to allow the activation in a FRU, for instance, to activate that individuals PIN, to signal familiarity, and SIU, to allow more specific information about that person. The FRUs are view-independent, so that the unit can become active from any view of a particular face. This has been developed into interactive activation and competition with learning (IACL) (Burton, 1994) to allow unknown faces to learn their own FRUs. The model was not intended to be used as a functional system, but as a tool to confirm theoretical expectations.

Use of Negative Examples

Ensembles of networks, such as used by Edelman et al. (1992), rely on a second stage network to utilise implicit negative knowledge, where if one input has a large value it will act as a negative influence on all the other input units, because only one class can be present at any one time. This approach may be made more accurate if explicit negative (non-class) examples are learnt alongside the positive class examples. This technique has been shown to be of critical importance in building robust face detection systems (Sung & Poggio, 1994; Moghaddam & Pentland, 1995; Rowley et al., 1996). The selection of prototypical non-face training examples can be very difficult, as they have to represent the entire class space of non-face images, which is considerably larger than the class space of face images. Most successfully so far, Rowley et al. (1996) has trained networks with a 'bootstrapping' algorithm which adds previous false detections to the training set as the training progresses, which reduces the number of negative examples required. This shows that prior knowledge of confusion in the distinction of classes can be used to guide the choice of appropriate training examples. This issue is developed in Chapter 5, where we introduce the 'Face Unit' RBF network model, which uses this type of positive and negative evidence to signal the presence of one particular class.

2.5.6 Temporal Networks

Representations and reasoning only concerned with data from single points in time are ignoring potentially useful information occurring through time, including significant temporal correlations in image sequences. Study of the statistical properties of static images (Field, 1987; Hancock et al., 1992) has shown some regularities, see Section 2.4.4. This has been extended to image sequences (Dong & Atick, 1995) to show a high level of spatio-temporal correlation, showing that natural time-varying images do not change randomly over space or time. If our data source provides information over time, we would do well to take advantage of this.

Recognising simple temporal behaviours is an important capability in computer vision applications such as visual surveillance (Buxton & Gong, 1995b) and biomedical sequence understanding (Psarrou & Buxton, 1993). Dynamic neural networks for such tasks can be constructed by adding recurrent connections to form a contextual memory for prediction in time (Jordan, 1989; Elman, 1990; Mozer, 1994). These partially recurrent neural networks (RNNs) can be trained using back-propagation but there may be problems with stability and very long training times when using dynamic representations.

A limited alternative is to use a simple Time Delay structure which can provide fast, robust solutions. The Time-Delay Neural Network (TDNN) model (for an introduction, see Hertz et al. (1991)), incorporates the concept of time-delays in order to process temporal context, and has been successfully applied to speech and handwriting recognition tasks (Waibel et al., 1989). Its structured design allows it to specialise on spatio-temporal tasks, but, as in weight-sharing networks, the reduction of trainable parameters can increase generalisation (Le Cun et al., 1989) and give some

shift invariance when used as convolutional networks (CN) (Lawrence et al., 1997). A time delay variant of the RBF network, the TDRBF network, has more recently been developed for speech recognition (Berthold, 1994). This has the benefits of ordinary RBF networks over other models, such as low numbers of tunable parameters and fast training times.

An even simpler approach than Time-Delays for processing image sequences was used by Rosenblum et al. (1996) to create an ensemble of single emotion RBF networks. A decay constant was applied to encode temporal facial optical flow information from several frames into a single frame, so that earlier frames had less value than later ones when summed into the composite input frame. Although this reduces the representation size. it creates ambiguities between spatial and temporal changes as both have to be shown in the same space.

The capacity of temporal cortex for making associations (Stryker, 1991) has led research into using temporal relationships in patterns for learning, for instance, different face views (Bartlett & Sejnowski, 1996, 1997), using competitive Hebbian learning with a temporal 'trace rule' originally proposed by Földiák (1991). In contrast to previous models, these temporal learning rules use differences over time, rather than simple time windows, to directly learn those temporal relationships required for specific tasks.

2.5.7 Discussion

The limitations imposed by the requirements for the face recognition task prevents most of the computationally-intensive techniques from being used here. In particular, the most 'biologically-plausible' approaches, such as the VisNet network model (Wallis & Rolls, 1997), IAC model (Burton, 1994) and the full Gabor processing of Petkov et al. (1993), are the slowest in operation.

An example-based view interpolation learning approach using Regularisation Networks, especially RBF networks, is very attractive as a face recognition technique, due to its simplicity and ease of training. In addition, they provide fast and robust operation. We noted that there is evidence that we use some kind of 'face recognition unit' to recognise familiar faces (Bruce & Young, 1986; Bruce, 1988). In addition, primate vision systems seem to use some kind of view-based representations for recognition (Perrett et al., 1989; Perrett & Oram, 1993; Logothetis et al., 1994). These ideas are partially captured by the RBF network where the first layer of the network maps the inputs with a hidden unit devoted to each view of the face to be classified. The second layer is then trained to combine the views so that a single output unit corresponds to the individual person. If we regarded filter-based preprocessing as an early stage of representation of identity, we can now regard hidden unit output from an RBF network as a later stage of that representation, which has been transformed into a space of considerably fewer dimensions.

Up until this point, we have deliberately left out specific performance figures from our discussion of face recognition research, as it is extremely difficult to compare work from different groups when each chooses their own recognition task. Even current 'state of the art' tests, such as the FERET database, can be seen as too easy if their test images are compared to those encountered in realistic situations. Without dealing with expression, pose and lighting as confounding variables, a system can appear to work well, but turn out not to be useful in a practical application. As discussed in Section 2.1, truly robust systems would need to account for many other non-trivial aspects, such as temporal behaviours, occlusion and speech-related facial changes.

We believe that the best approach to satisfying our specific task requirements is to learn face classes over a wide range of poses with an RBF network. A preliminary test of the suitability of our approach is given in the following section, where we compare our results directly with published experimental results for other approaches using a common database.

2.6 Comparing Face Recognition Techniques

In the earlier sections of this chapter, we suggested that RBF network view interpolation with filtering preprocessing was a good way forward to meet our task requirements. The RBF network

Group	Technique	Images per Person					Processing Time	
		1	2	3	4	5	Training	Classification
Samaria & Harter	HMM	?	?	?	?	87	?	?
	pseudo 2-D HMM	?	?	?	?	95	?	4min
Lawrence *et al.*	Eigenfaces	61	79	82	85	89	?	?
	PCA + MLP	?	?	?	?	59	?	?
	SOM + MLP	?	?	?	?	60	?	?
	PCA + CN	66	83	87	88	92	?	?
	SOM + CN	70	83	88	93	96	4hr	<0.5sec
Lin *et al.*	PDBNN	?	?	?	?	96	20min	<0.1sec
Lucas	*n*-tuple	54	68	75	78	81	0.9sec	0.025sec
	cont *n*-tuple	73	84	90	93	95	0.9sec	0.33sec
	1-NN	?	?	?	?	97	0sec	1sec
Howell & Buxton	RBF before discard	49	65	72	80	86	8sec	0.01sec
	after discard	84	90	91	95	95	8sec	0.01sec

Table 2.1: Test generalisation (% correct) and processing times for various face recognition techniques used by various researchers using ORL Face Database of 40 people, averaged over several selections.

has been shown to provide robust classification even where data is noisy or partially missing (Ahmad & Tresp, 1993). Our original question (Howell & Buxton, 1995) was whether this ability can be used with complex 3-D objects such as faces, where the data varies in lighting, expression and pose. Here, we compare the RBF techniques with other methods using a standard database.

It is particularly important to establish that the RBF network is able to distinguish a useful number of face classes, as this will indicate its potential as a practical technique for future applications. A suitable source of data to test this is the Olivetti Research Laboratory (ORL) database of faces. This contains 400 images of 40 people, which is sufficient to satisfy Task Requirement 3a (see Section 2.1). Details of the database are in Section A.1 in Appendix A. It should be noted that comparisons with these separately published results is a quick and simple compromise, as the test results presented here were collected on different systems and under different testing regimes. However, they give a rough indication of comparative performance and suitability for our task requirements.

2.6.1 Results

Table 2.1 summarises the results from several published papers, plus our own tests. Test generalisation performance for systems with differing numbers of training images are given, together with times for the train and test (classification) stages (where available).

Hidden Markov Models

Samaria and Harter (1994) initially developed and experimented with the ORL database, using conventional Hidden Markov Models (HMMs) as a graphical probabilistic approach to encoding feature information. This approach used several subjective parameter selections, and gave a top performance around 87% for a system trained with 200 images. Further work using pseudo 2-D HMMs (Samaria, 1994) was able to improve this to 95%, but the computational complexity of this approach seems to count this out as a useful real-time technique, as 4 minutes is a long time to wait for a classification.

Eigenfaces

Both Samaria and Harter (1994) and Lawrence et al. (1997) tested the ORL database with the 'eigenface' (Kirby & Sirovich, 1990; Turk & Pentland, 1991) approach. Both report performance of around 90%, though the latter found that they could only get this by using separate training vectors for each image. This is in contrast to Turk and Pentland (1991), who averaged the eigenfaces for all images of each person in their tests. When tested with ORL data, this latter approach gave 74% for 5 training images per person (Lawrence et al., 1997). That this is much lower than the MIT results (where results over 90% are common) would seem to indicate that the ORL database represents a much harder task than the MIT face database (assuming the implementations were equivalent).

Convolutional Networks

Lawrence et al. (1997) used a self-organising map (SOM) to reduce the dimensions of the input representation, and a five-layer convolutional network (CN) to give translation and deformation invariance. This was faster than the previous HMM approach and performed equally well, but still required several hours training time.

They compared the dimensionality-reducing abilities of the SOM with principal components analysis (PCA), and the CN with a multi-layer perceptron (MLP). This latter approach gave very poor results, especially when several hidden layers were used. It should be noted that the figures for these approaches in Table 2.1 show the single best results from all combinations (which came from a MLP with one hidden layer) rather than average results (which are given for the other approaches from other groups).

Probabilistic Decision-Based Neural Networks

Lin et al. (1997) used a probabilistic, decision-based neural network (PDBNN) a modular network structure with non-linear basis functions (each sub-network similar to a HyperBF (HBF) network (Poggio & Girosi, 1990b)) that was able to train and classify much faster than the CN approach of Lawrence et al. (1997), while reaching a similar level of performance.

Continuous n-tuple Classifiers

The continuous *n*-tuple classifier (Lucas, 1997) is an updated version of earlier *n*-tuple classifiers, such as WISARD (see Section 2.5.2). The updating refers to speed and storage efficiency, so it is likely that this technique would suffer the same problems with image variations such as pose or lighting in real-world tests. However, the approach does train and classify quickly and provide a high level of performance.

The figures shown are for tests with 200 3-tuples (600 values) per image. Using 500 4-tuples (2000 values) per image improved recognition to 86% and 97% for the *n*-tuple and continuous *n*-tuple classifier respectively.

Nearest Neighbour Classifiers

Lucas (1997) was also able to achieve very high performance using a simple 1-nearest-neighbour (1-NN) classifier using a City-Block distance measure. The success of simple matching indicates how constrained the database is in terms of lighting and pose, as such techniques will not be invariant to such factors.

RBF Networks

To use the ORL data with the RBF network, we subsampled each image to 25×25 and applied 'A3' Gabor filter preprocessing (see Chapter 3 and Appendix C for more details). A simple discard measure, based on the relative magnitudes of the output units, was used to remove low confidence classifications (these being those where the highest output value was less than a certain ratio below the next highest). Each training example was used as a centre vector for a hidden unit.

The RBF network approach was fast in training and the fastest in classification of all the published techniques. Our experiments were conducted on a moderately fast Sun SPARC 20 workstation. Test generalisation before discard was fairly poor in comparison to the other approaches, though the results were well above random (2.5%). For 5 training examples per person, discarding

39% of results allowed performance to be improved from 84% to 95%, which was comparable with the best of the other techniques. The results after discard for the RBF network were especially good where lower numbers of training examples per person were provided.

2.6.2 Discussion

Table 2.1 shows that although, in pure generalisation terms, our RBF network approach is not the overall top performer, it does have a sufficient level of performance (95% after discard) for our target application where it will have to deal with image sequences. In this type of application, training data is relatively sparse (compared to the large range of variation in real-life images) and test data is abundant. The success of the RBF discard measure, which makes it the top performer where low numbers of training examples are available, highlights its efficiency in interpolating between even small number of views for reliable classification. Although discarding does reduce the number of useful classifications, a significant amount of data will remain when such techniques are used with image sequences. This issue is taken up in Chapter 6, where the RBF network is applied to real-life image sequences.

The ORL database is a highly constrained database and not designed to meet our task requirements. Thus, success in recognising the ORL faces does not necessarily indicate a suitability for our less constrained face recognition task. For example, the constant lighting conditions do not require an invariance to illumination to be developed, and thus no consideration has been made by the other approaches to the issue of preprocessing. We are not able to know how the other techniques would perform in the presence of variable lighting, but we believe that real-life applications would require some type of preprocessing to overcome this kind of variability.

A particularly important point is that all the other face recognition techniques gave processing times which are very much slower than the RBF network in classification of the ORL data. It is apparent that the RBF network provides a solution which can process test images in inter-frame periods on low-cost processors (Task Requirement 1a).

2.7 General Discussion

This chapter started by introducing our face recognition task and establishing its requirements. This was followed by a discussion of previous face recognition approaches, in terms of acquisition, representation and reasoning. We have evaluated a wide variety of general approaches with respect to our specific task requirements. It is perhaps not surprising that many of the approaches to face recognition discussed in this chapter do not fulfill these task requirements for the simple reason that they were not designed with such an application in mind. We stress that success of an application will be determined by relevance of the approach to the task.

The combination of Task Requirements 1a and 1b to be robust in the face of noisy and variable data, and yet fast enough to give results in inter-frame periods with standard processors is demanding, but not impossible. Simple filter-based preprocessing can give some invariance for the input representation, and RBF networks will give speed and robust performance for the recognition itself. The suitability of the RBF network approach for handling occlusion, covered by Task Requirements 3(c)iii and 3(d)ii, has been shown by results in Edelman and Poggio (1992) and Ahmad and Tresp (1993).

In summary, we can see that our proposed filter-based RBF view interpolation scheme appears to be very suitable for our target task: it combines fast training and testing times with the ability to cope with complex 3-D transformations. Results using the ORL standard database indicate that the network can discriminate useful numbers of face classes. We will establish further detailed evidence of requirement fulfillment in the following chapters.

The rest of the thesis will be concerned with how to apply this network to our target application of identity recognition in unconstrained domestic environments. The next chapter, in particular, will look at the Euclidean distance measure and how variations in the images, such as resolution,

pose and filter-based preprocessing methods, affect the distances between face identity classes. In particular, we will concentrate on how such distances are modified with pose variations, as this is crucial for our task. In addition, the reasoning component of the RBF network will be analysed and compared with related classification methods.

Chapter 3

Representations of Pose-Varying Faces

The previous chapter has shown the suitability of our proposed approach to the main task of face recognition, using a computationally efficient approach based on RBF networks with simple, receptive field-based preprocessing. This chapter will introduce our main test database, the 'Sussex database', which allows testing of face recognition techniques over moderate pose ranges. The database contains images of ten people in ten different pose positions from face-on to profile, 100 images in all (see Section A.2 in Appendix A for specific details).

The task requirements specify a tolerance of large pose variation (Task Requirement 3(d)ii), and it was necessary for us to create this specific database, as there is no other publically available data that systematically varies pose over a useful range for all individuals. Generally, it is desirable to have more widely-tested data, such as the ORL database (used in Section 2.6 of the previous chapter), so that comparable results are available.

The first section of the chapter investigates how the image data varies over pose for each individual, specifically, in relation to Euclidean distance comparisons between images of the same and other identity classes. We will be establishing how distinct such face-classes are under pose change. This will provide a context for the second section, which is concerned with analysing the individual classification components of the RBF network to see how (and whether) it can provide superior performance to simple, non-learning classification methods, such as 'nearest neighbour'. The third section looks in depth at how preprocessing helps the learning and generalisation process through modification of the face representation, using two specific receptive field function-based techniques – Difference of Gaussian filtering and Gabor wavelet analysis.

3.1 Euclidean Distances for Faces

With the Sussex database as a source of suitable data, we now want to establish how difficult it is to distinguish the individuals over varying pose. To do this, we can compare the Euclidean distances (defined in Equation B.3 in Appendix B) of *reference images* to all other images in the database (the *test images*), distinguishing two types of class distance:

Intra-class Distance The Euclidean distance between the reference image and a test image, where both are of the *same* identity class.

Inter-class Distance The Euclidean distance between the reference image and a test image, where both are *different* identity classes.

Obviously, if the former are less than the latter in all cases, the classification problem is solved, as perfect discrimination between the identity classes will be possible based on simple comparison alone. We do not expect this to be the case in practice, as real-life images are noisy and faces vary enormously over pose, expression and lighting.

We first look at the fundamental similarity mechanism we use for image-based face class discrimination, the Euclidean distance measure. This is applied to vectors of our basic representational 'feature', the pixel or pixel-based coefficient. The simpler 'City-Block' distance metric is tested later in Section 3.2.4.

3.1.1 Varying Face Resolution

It is important to first determine how the resolution of the data affects identity class discrimination through changes in relative intra- and inter-class Euclidean distances, so that a useful standard image size can be established for future tests. The original 100×100 window region is sub-sampled, using averaging, to a range of resolutions between 100×100 and 6×6, see Figure A.4 in Appendix A.

Preprocessing

We consider that preprocessing will be an important part of any real-life application, where lighting variation will be expected, and therefore needs to be included even in initial studies of the recognition process. Difference of Gaussians (DoG) filtering gives a useful level of invariance to lighting, which helps with Task Requirement 3(c)ii. Although Section 3.3 will deal with the effects of preprocessing in detail, for these initial tests we will use DoG convolution with a single scale factor (this will vary according to image resolution, see Table C.1 in Appendix C). The convolved values are thresholded to give binary information which emphasises the zero-crossing boundaries (see Section C.1.3 for specific details).

The representational ability of the smallest image resolution (6×6), which contains just 16 binary values after preprocessing, is obviously very limited, and we will not expect high generalisation performance when using this data. In comparison, the full resolution (100×100) will contain very much more information, but this may be more than is needed for the task, increasing computational complexity needlessly and even losing generalisation through a dependence on finer pixel-wise registration for image matching.

Results

Figure A.8 in Appendix A shows how the Euclidean distances vary through resolution for one specific image, pose 40° for one particular class (0). This one image is compared to all 100 images in the Sussex database, the zero value corresponding to where it is compared to itself, where the Euclidean distance is nil. It can be seen that there is a clear division between the same-class (shown by the thick line) and the other-class distances, but this is clearest only for the frontal range (0°–45°) of views. Interestingly, the resolution does not appear to greatly affect this degree of separability of the other images of the same identity class from those of the other classes, apart from the lowest resolution (6×6).

However, this intra-/inter-class distance distinction will vary according to the individual face-class. For example, Figure A.9 shows similar comparisons as for Figure A.8, except with pose 50° of a different class (5), and it is clear that there is much more inter-class confusion here than for the earlier example. Indeed, very few of the intra-class distances are lower than the corresponding inter-class distances, even at the highest resolution. Obviously, this indicates that some individuals will be harder to distinguish than others, but also, more importantly, that while a simplistic, 'winner-take-all' approach might work for one image (pose 40° of class 0), it cannot be assumed to work for other poses and other identity classes.

Table C.1 in Appendix C shows that the resolution of the face image has a great effect on the amount of data used to represent that image after preprocessing. Although the lower resolutions offer extremely compact representations, Figures A.8 and A.9 show that they may not be as robust in distinguishing individual classes as the higher resolutions. The 25×25 resolution is a good compromise between size and clarity of representation, having only 6% (5% after DoG preprocessing) of the data values of the 100×100 resolution, with only minimal loss in information.

In summary, the 25×25 resolution of face image has been found to be a useful compromise between a compact and a comprehensive representation for specific images. Section 3.2 will demon-

strate practically how the resolution affects the entire database when used to train a variety of classifiers.

3.1.2 Varying Face View

The pose view of the person is another factor, besides resolution, that affects the inter-class distinction. This is illustrated by Figure A.12 in Appendix A, which shows all Euclidean distances for six individual images at the 25×25 resolution from the Sussex Database, three each from classes 0 and 1 using pose angles of 0° (frontal), 40° and (a) 90° (profile). As in Figures A.8 and A.9, all 100 distances are shown on the graphs, connected by lines according to class, and the zero value can be seen where the image is compared to itself.

Results

The extreme profile view (90°) is less distinct than the centre views in Figure A.12 in Appendix A, and this will add to the problem of lack of interpolative data when we come to use these images with the RBF network, which largely relies on data interpolation. Because of this, we can expect that performance for the RBF networks using profile information will be significantly lower than for the central views and also lower than for the frontal (0°) view, where the intra-class views remain distinct for a greater range of views.

Intra-class Euclidean distances have been shown to be less, for some specific images in the Sussex database at least, than for inter-class comparisons for small pose angle ranges. This shows the potential of using such comparisons for recognition, especially where training examples can be provided at regular pose intervals.

Figure A.12 shows some bias in intra-/inter-class distinction for the frontal range (0–45°) over the profile range (45–90°). This may help to explain experimental results in unfamiliar face recognition, such as O'Toole et al. (1995), where no advantage was found for '3/4' views over frontal views (instead both were equivalent and much faster to match than profile views). Bruce (1988) took such results as supporting the view that 3/4 views were not serving as 'canonical' representations for recognition and that full-face and profile view might be separately represented. The mid-pose views used in Figures A.8, A.9, A.12(b)(i) and (ii) all show that same-class frontal views can quite often be discriminated from other-class views simply on the basis of Euclidean distance alone. This can be contrasted to Figures A.12(c)(i) and (ii), which show that class distinction is much lower when the extreme frontal or profile views are used as the reference.

3.1.3 Centralisation of Faces

The tests so far in this chapter have used face data which is centralised on the nose tip. This is based on an intuitive assumption that keeping such a facial feature fixed in the image would make comparisons over varying pose easier. This section is to check how our centering technique has affected the representation by comparing it with face data where pose has not been taken into account, using a simple framing of the head.

To create this alternate data set, we reprocessed the Sussex database images, adjusting the centering during face localisation to fit as much of the face onto the image, regardless of pose. This we term *face-centred* data, in contrast to the original *nose-centred* data, and is illustrated in Figure A.7 in Appendix A.

Results

Figure A.10 shows how the Euclidean distances vary through resolution for one specific image using the face-centering algorithm. This can be compared to Figure A.8, where the identical reference image was used, but with nose-centering. The face-centering appears to create smaller inter-class and greater intra-class distances, which indicate that the class distinction is not as effective as for nose-centering. An anomaly appeared in the face-centered 6×6 resolution graph, where the representation is so coarse that a neighbouring image (50°) is actually identical, and so has a zero Euclidean distance to the 40° image.

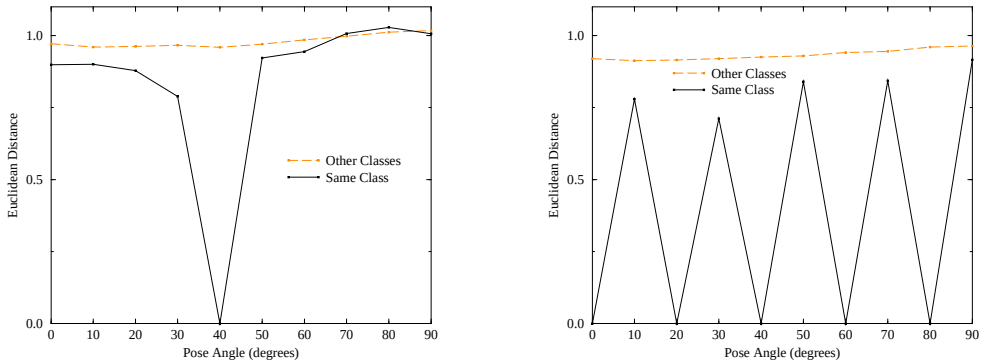
The nose-centering technique we have employed here is not a rigorous, mathematical vectorisation of the image, such as used by Beymer (1995), but the hand-alignment of facial features, such as the left eye and nose, over pose is shown to improve general class distinction (in terms of Euclidean distance). We will also be able to show later that the nose-centered representation can improve generalisation and reduce discard rates with RBF networks (see Table 3.2).

3.1.4 Discussion

This section has investigated how image resolution and face pose variations affect the distinction between Euclidean distances for images compared with others within its class and with those from other classes. We have been able to show that Euclidean distance comparisons can be used to distinguish between images from same and other classes within the Sussex database, at least for a few pose steps (each step is roughly 10°) closest to the pose angle of the reference image.

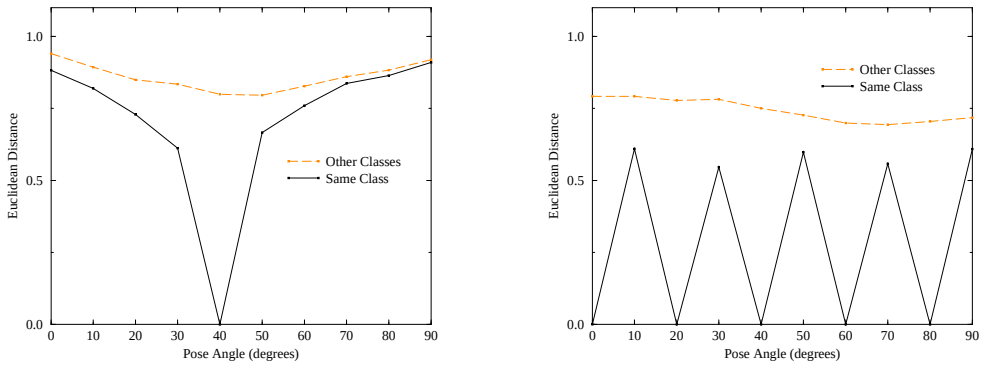
To see how different these distances are over the entire Sussex database, Figure 3.1 shows the overall average value for all intra- and inter-class distances, on a pose-by-pose basis. In (i), one image from a class (pose 40°) is compared to all others in the database. For clarity, the distances for the other classes are averaged for each pose position (examples of specific values for these classes can be seen in Figures A.8(c), A.12(b)(i), A.12(b)(ii) and A.13(b)) to provide a single line on the graph. For (ii), the same process is carried out for five images from one class, and the lowest distance value to those five images for each class at each pose position is then used as before (inter-class distances averaged before plotting). A clear division can be seen between the two lines in (i), indicating that the two types of distances can, in principle, be distinguished for most pose angles with a single reference image. As has been shown earlier, however, there are specific images where the distinction is unclear (for example, see Figure A.8(c)), and so contextual classification methods may be needed to disambiguate such data. The 5-example graphs (ii), which use five examples of each class as reference images, show a wider gap between the two average distance lines, indicating that the increased class knowledge will improve classification.

The next section will go on to show how pose and resolution differences affect the learnability of face classes by classifiers.



(i) (ii)

(a) DoG preprocessing



(i) (ii)

(b) Gabor preprocessing

Figure 3.1: Average Euclidean distances for 25×25 face images from Sussex database, with different preprocessing, between same and other classes whilst varying pose angle (i) compared to one pose angle, (ii) compared to 5 pose angles (using the lowest distance over the five).

3.2 Learning Identity

The previous section established that the intra-class Euclidean distances were shorter than the inter-class distances, but that this only was true on average. There will be individual images, perhaps even a majority, where this is not true, and so to take a simplistic, 'winner-take-all' approach to classification will be suboptimal.

We have identified the radial basis function (RBF) network as a suitable learning element for our task in Section 2.6 in the previous chapter. In this section, we will concentrate on investigating the individual parts of the RBF model. We can assess the contribution of 1) the Euclidean distance comparisons with simple 'nearest neighbour' (NN) classifiers, 2) the non-linear Gaussian centre function with Gaussian probabilistic neural networks (PNNs), and 3) the adaptive weight layer using a standard RBF network. Appendix B gives specific details on the implementation of the RBF network model used for the experiments.

We will be expecting a general improvement in generalisation performance as we progress through the three stages to the full RBF network. These tests will all use a '50/50' training configuration, with 50 images (5 for each class, taken from alternate pose positions) from the Sussex database being used for training and the other 50 for testing generalisation.

3.2.1 Nearest Neighbour (NN) Classification

The nearest neighbour (NN) classifier uses Euclidean distance comparisons to a set of 'training' reference examples to classify the test images. There is no training, except in the selection of five reference examples for each class.

The NN classifier is implemented via the basic RBF network structure, having altered its hidden-to-output unit weight layer. All the weights are set to 1 or 0 according to class. Weights connecting hidden units to the output unit of their class are set to 1, all others 0. The Euclidean distance is calculated by the hidden units, but no Gaussian is applied to these values.

Two types of NN classifier were used:

Winner-takes-all (WTA) Classifier This uses the single lowest Euclidean distance from the test image to all 50 reference examples to give a classification. This is implemented by setting the output of the winning hidden unit to 1, all others to 0.

Class-based Classifier This sums all Euclidean distance values for all 50 training images for each class, classifying on the basis of overall value for each class. As this combines evidence from several images for each person, we might expect it to outperform WTA.

Results

Table 3.1 shows the results for the two 5-example NN classifiers. The WTA arrangement gave better performance than the Class-based arrangement. The WTA had all hidden unit outputs, apart from the winning one, set to 0. This effectively removed any cooperative or undermining influences from the other hidden unit values. The lack of a trained weight layer may well mean that such contextual information is more confusing than helpful.

It can also be seen in Table 3.1 that generalisation performance is best at the highest resolutions and tails off as the resolution is reduced. This confirms the expectations about general class distinction trends arising from the Euclidean distance graphs seen earlier (such as Figure A.8 in Appendix A). For comparison, Table 3.4 shows how 1-example NN classifiers (trained with one example per class) were able to generalise with the same data. Obviously, the WTA and Class-based NN schemes are equivalent in the 1-example configuration.

In summary, the WTA and Class-based NN classifiers were able to classify the Sussex database to a reasonable level of performance without any learning. The simplest arrangement, WTA, performed best.

Resolution	% Correct		
	NN (WTA)	NN (Class-based)	PNN
100×100	88	78	70
50×50	86	76	80
25×25	84	70	70
12×12	78	70	74
6×6	48	42	54

Table 3.1: Test generalisation for 5-example simple nearest neighbour (NN) classifiers and probabilistic neural networks (PNNs) using DoG preprocessed images at varying resolutions (nose-centred).
WTA NN classification takes the closest Euclidean distance from the test image to the training examples to assign a class value. Class-based NN classification sums all distance values for all training images for each class, classifying on the basis of overall value for each class.

3.2.2 Probabilistic Neural Networks (PNNs)

To assess the contribution made by the Gaussian centre function (see Section B.1 in Appendix B for more details), we can use a radial Gaussian form of a probabilistic neural network (PNN) (Specht, 1990). As for the Class-based NN classifier, a network is implemented by using the RBF network model structure. This time, the only difference between the RBF network and the PNN is that the hidden-to-output layer of weights for the latter are all fixed at 1 or 0 according to class. The PNN separates all the hidden units (pattern units) for each class, their activations being fed to a summation unit, again separated by class.

Results

Table 3.1 shows the results for the PNN. The structure of the PNN is very similar to the Class-based NN classifier, apart from the non-linear process of applying the Gaussian to the Euclidean distance comparisons. Not surprisingly, therefore, the results are quite similar, showing an improvement (over the Class-based NN classifier) in all resolutions except the highest (100×100). They are not identical though, indicating that the Gaussian function has modified the group totals, and that this generally will lead to an improvement in generalisation.

In summary, the Gaussian function used by the PNN is still not sufficient to allow it to outperform the simplest model, the WTA NN classifier, except at the lowest resolution (6×6) of data. This may be due to the hidden unit activations being summed by class. This summing of all views for each class may 'blur' class distinctions, preventing effective generalisation.

3.2.3 Radial Basis Function (RBF) Networks

To assess classification of the face images with a full RBF network model, we use two sets of data, one nose-centred and one face-centred (for example, compare Figures A.5 and A.7). The main difference between the RBF and the Gaussian PNN used previously is that the RBF has an adaptive hidden-to-output weight layer (see Section B.2 in Appendix B for more details). This will allow a greater influence of 'context' from hidden units of same and other classes.

We can test the RBF network with all the resolutions of face data from the Sussex database used in Section 3.1 to check our initial predictions on the learnability of those representations which had been made on the basis of Euclidean distance comparisons alone. We can also check the hypothesis that the nose-centered data is a better representation for our task than the face-centered data.

Results

The results for the RBF network in Table 3.2 show that performance using nose-centred data is generally better than when using face-centred data for all resolutions apart from the lowest (6×6),

Resolution	Centering	Initial %	% Discarded	% After Discard
100×100	Nose	76	50	100
	Face	72	66	100
50×50	Nose	82	42	100
	Face	70	58	95
25×25	Nose	78	52	100
	Face	62	64	100
12×12	Nose	72	46	96
	Face	70	60	90
6×6	Nose	46	40	63
	Face	64	40	87

Table 3.2: Test generalisation for 5-example RBF network using DoG preprocessed images at varying resolutions and with nose- or face-centering.

which gives much poorer generalisation. These results match the relative class distinctiveness of each resolution shown in the Euclidean distance graphs in Figure A.8.

Table 3.2 shows how the nose-centered face data is a more efficient representation than the face-centered data, both for generalisation and for the proportion of low-confidence classifications. The curious exception to this is for the 6×6 data where the use of the face-centered data get better results than for the nose-centered data. Section 3.1.3 explains that this is due more to the coarse granularity of the representation, rather than some useful invariance, so it is unlikely that this type of data would be useful for any practical application.

Confidence Measure

As mentioned in Chapter 2, a useful feature of RBF networks is the availability of a level of confidence in the output. This can be derived from the relative values of the highest and second highest output units. The ratio between the two reflects how much contrast there is between the successful output units and the others, and so the larger this is, the more 'confident' we can say that the network is that the classification is correct.

The setting of a threshold on this ratio, below which the classification is discarded, allows greatly enhanced test generalisation performance. Table 3.2 shows the performance for RBF networks before and after discarding low-confidence classifications, using a threshold value of 1.8. The fairly large percentage discarded is in line with our task requirements, which favour a *low precision, high discard* approach.

This threshold value of 1.8 has been found to be a good compromise in practise, but it can be tailored to Task Requirements. Figure 3.2 shows how varying the ratio as a threshold to discard low confidence classifications affects the final generalisation performance, and indicates the suitability of a threshold between 1.5 and 2.0.

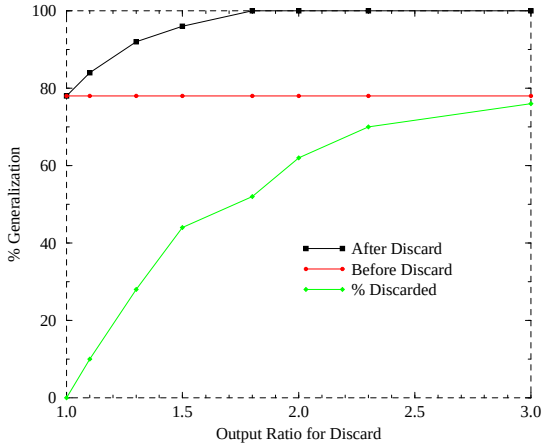


Figure 3.2: Effect on test generalisation (after discard) of changing the 'low confidence' threshold for 50/50 RBF networks trained with DoG preprocessed 25×25 faces images from Sussex database. The low confidence threshold is based on the ratio between highest and second highest output units, and a value of 1.8 has been found to be useful in practise.

Resolution	Pose Accuracy	Initial %	% Discarded	% After Discard
25×25	±5°	34	84	38
	±15°	70	84	88

Table 3.3: Test generalisation for 5-example RBF network trained on pose classes using DoG preprocessed images.
The network is trained on all 10 poses for 5 people (5 examples of each pose) and tested on all 10 poses from the other 5 people in the Sussex database. An accuracy of ±5° requires the exact 10° pose position in the image to be identified, whereas ±15° allows pose positions to either side of target as correct also.

3.2.4 City-Block *vs.* Euclidean Distances

The simpler 'City-Block' distance metric was also tried alongside Euclidean distance for comparison. This uses the summed total of the distances, but without the squaring used for the Euclidean metric (compare with Equation B.3 in Appendix B):

$$d_{CB}(\mathbf{i},\mathbf{c}) = \sum_{x=1}^{N} |i_x - c_x| \tag{3.1}$$

(notation defined in Appendix B). See Kohonen et al. (1981, Chapter 2) for more details of the mathematical background and relationship between the two metrics.

Results

Unlike reports where the City-Block metric has been found to be much better than the Euclidean for face classification, such as Lucas (1997), we found little difference in generalisation between the two. Table 3.4 shows an example comparison for all classifiers used in this section.

One reason for this difference in the relative performance of the two metrics may be that we are using binarised data values (where City-Block becomes equivalent to Hamming distance). City-Block distances will be more different from Euclidean distances with grey-level values than with binary, as the greater numerical range of values will be accentuated by the squaring done by the Euclidean measure. This might give some extra generalisation whilst using grey-level information in fairly constrained situations, but it is not clear that this would carry over into the more profound variation encountered in real-life images, where preprocessing can be helpful.

3.2.5 Learning Pose

Next, we investigated whether *pose* classes can be found within the data. We train the RBF network as before, with 5 examples of each class, and test with the other 5 images.

The 50/50 selection of alternate images in database means that the RBF network is trained on all 10 poses from 5 of the people in Sussex database, and tested on the other 5 people, who were therefore unseen during training.

Results

Figure A.11 shows that the intra- and inter-class Euclidean distances are much less well defined for pose than for identity (for instance, Figure A.8). We can predict from this that specifying an exact 10° pose position will be harder than specifying an exact identity. However, since the poses are linked, we might be able to improve recognition by asking for a less exact pose estimation, such as ±15°, which would require the identification of the pose within a 3-step range. Table 3.3 shows that the RBF network performs rather poorly when asked to give the exact 10° pose position, but this improves greatly for the lower precision task (±1 pose position).

Classifier	Distance Metric	Initial %	% Discarded	% After Discard
NN (WTA)	City-Block	43	-	-
	Euclidean	43	-	-
NN (Class-based)	City-Block	43	-	-
	Euclidean	43	-	-
PNN	City-Block	32	-	-
	Euclidean	39	-	-
RBF	City-Block	39	83	80
	Euclidean	39	90	100

Table 3.4: Test generalisation for 1-example classifiers using City-Block and Euclidean distance measures trained with DoG preprocessed 25×25 faces images from Sussex database.

In summary, it was not as easy to distinguish pose classes as it was for identity classes. This suggests that, for this database at least, images of different identity are further apart in Euclidean space than images of different pose. We can say that the use of learning by examples distinguished by Euclidean distances is therefore especially appropriate for face recognition in the presence of large pose changes, as the distances are affected more by identity than pose.

3.2.6 Discussion

This section has presented generalisation performance from a variety of kernel-based classifiers trained with the Sussex database. These show that it is possible to distinguish face classes using simple classifiers moderately well even under fairly large pose ranges, but that the confidence measure from the RBF network allows it to outperform the simpler methods.

Figure 3.3 clearly indicates that there is little advantage to using the higher resolution representations for recognition, as good generalisation performance can be achieved with a fraction of the data size of the 100×100 images (for example 5% for 25×25, 1% for 12×12, see Table C.1, Appendix C). In addition, data storage and computational load can both be greatly reduced using one of the lower resolutions.

We did not find the expected incremental improvement in performance through the three classifiers, indeed, if one ignores the values after discard, the simplest classifier of all, WTA NN, performed best. However, although the NN classifier and the PNN are both able to discriminate the face classes fairly efficiently, the RBF network has the great advantage of providing a confidence measure through a combination of the graded response from its hidden units and the learnt hidden to output layer weights, which fulfills Task Requirement 3e (see Section 2.1). The success of the WTA NN classifier shows that if your measure of similarity cannot be linked contextually with evidence from the same and other classes, it is better to ignore it altogether. However, the success of the RBF discard measure illustrates that such short-term approaches will not ultimately be able to provide the best classification or generalisation.

Figure 3.4 shows a comparison of the ratios between highest and second highest output value for each test image for PNNs and RBF networks. The PNN sums activations of only those hidden units for the class, and so does not allow 'contextual' influence from other classes. In contrast, the negative influence allowed in the RBF network weight layer gives it better performance than the PNN. This mechanism is illustrated by Figure 3.5, which shows exactly how the weights from the hidden units of one particular class vary according to which output unit they are connected to. Obviously, the largest values occur when connected to the output unit for their particular class, but besides that, there is quite a variety of values, even some positive ones, for the output units of other classes. This demonstrates how the RBF network provides contextual 'negative' or non-class information, unlike the NN classifier or PNN.

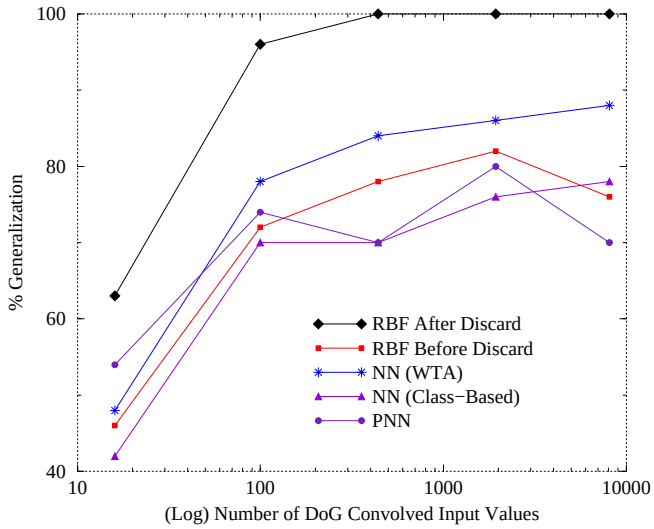


Figure 3.3: Effect of changing the number of input data values on test generalisation for 50/50 classifiers with five training examples per class. This number is varied via the original image resolution of face images from Sussex database before DoG preprocessing, see Table C.1, Appendix C, for details.
Discarding is only shown for RBF, as the nearest neighbour (NN) classifiers and probabilistic neural network (PNN) do not provide enough differentiation between output units to enable a discard measure, see Figure 3.4.

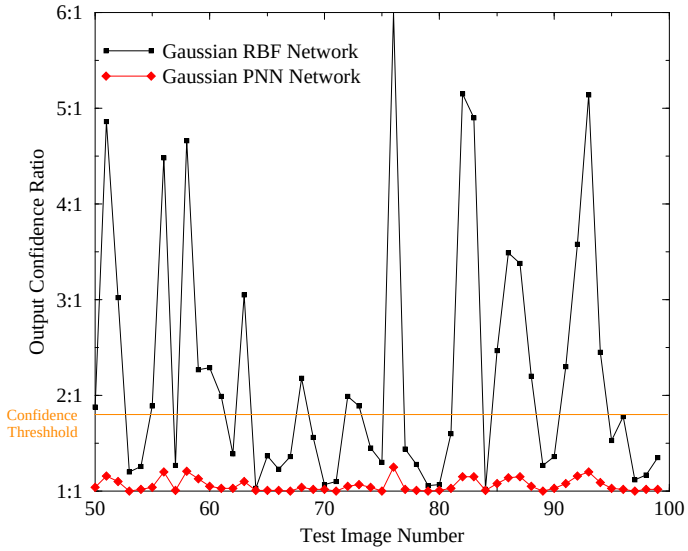


Figure 3.4: Confidence of network output represented as the ratio between highest and next highest output units for specific test images from the Sussex database for 50/50 Gaussian probabilistic neural networks (PNNs) and Gaussian radial basis function (RBF) networks.
The 'Confidence Threshold' level of 1.8:1 represents the discard threshold below which output classifications are deemed to be low-confidence. In this case, 70% of all PNN classifications were correct (none were high-confidence) and 78% of all RBF classifications were correct (100% of the high-confidence classifications were correct).

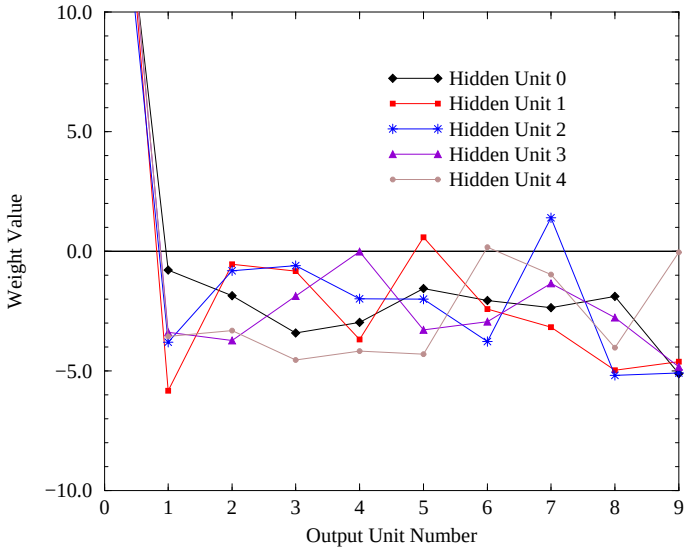


Figure 3.5: Values for hidden-to-output layer weights for a 50/50 RBF network, for hidden units 0-4 (corresponding to the 5 assigned training examples for class 0).
It can be seen that some non-class weights (those not connecting to output unit 0, in this example) are positive, which is unexpected. It is this flexibility in RBF network weight values that gives it superior generalisation performance over Gaussian probabilistic neural networks.

Worse generalisation performance (after discard) and lower discard efficiency was found using the City-Block metric together with the RBF network, for example, see Table 3.4. This, combined with little or no improvement in generalisation over the Euclidean metric, led us to retain the latter for our tests, despite its greater computational load.

We have only used DoG preprocessing up to this point in this chapter. The next section will present more detail about how different types of preprocessing affect representation and distinctiveness of face-classes.

3.3 Receptive Field Functions for Face Recognition

This section investigates types of image preprocessing that mimic the effects of receptive field functions found at various stages of the human vision system. We compare how the face representations they create affect learning and generalisation for the RBF network.

One of the main problems in computer vision, especially in face recognition, is dimensionality reduction to remove much of the redundant information in the original images. Simple mechanisms, such as sub-sampling, may give a rough reduction, but use of more specific prior knowledge to apply more sophisticated preprocessing techniques to an image is still required for the best results. Specifically, appropriate preprocessing of input representations for a face recognition scheme can overcome some of the problems of lighting and scale variations. Performance results here can be assessed alongside the number of sampled values used per image to give a measure of the usefulness, in the context of representation dimensionality, of a particular preprocessing technique.

One way of thinking about these input representations and mapping them onto our RBF networks is to use the analogy with visual neurons. The receptive field of such a neuron is the area of the visual field (image) where the stimulus can influence its response. For the different classes of these neurons, a receptive field function $f(x,y)$ can be defined. For example, retinal ganglion cells and lateral geniculate cells early in the visual processing have receptive fields which can be implemented as Difference of Gaussian filters (Marr & Hildreth, 1980). Later, the receptive fields of the simple cells in the primary visual cortex are oriented and have characteristic spatial frequencies. Daugman (1988) proposed that these could be modelled as complex 2-D Gabor filters. Lades et al. (1993) and Petkov et al. (1993) successfully implemented face recognition schemes based on Gabor wavelet input representations to imitate the human vision system, although they were extremely computationally expensive.

This section contrasts the use of Difference of Gaussian (DoG) filtering and Gabor wavelet analysis at a range of scales for our face recognition task. The question we want to ask here is whether these later stages of processing make more information explicit than the earlier DoG filters for our face recognition task.

3.3.1 Difference of Gaussians (DoG) Preprocessing

This section presents experimental results using RBF networks with the Sussex database of pose-varying faces with Difference of Gaussians (DoG) preprocessing. Section C.1 in Appendix C gives specific details on the DoG filters preprocessing technique.

Varying DoG Scale

The DoG scale parameter has a profound effect on the extracted image information, small values focussing on high-frequency details with little blurring, large values concentrating on the low-frequency features left after a high level of blurring, see Figure C.2. For example, the DoG scale of 0.4 was used in Section 3.1 for the 25×25 data as a rough, mid-way value, having a 5×5 mask which does not blur detail too much (see Figure C.2(b)). To explore how varying this scale value affected the learnability of the Sussex face data, we trained 50/50 standard RBF networks with several sets of data preprocessed at a range of DoG scales.

Results

Figure 3.6 shows that the scale has a clear effect on test generalisation, and confirms that the original value of 0.4 is a good choice for future experiments, especially after discard. Task Requirement 3e specifies that performance *after* discard will be of more interest than that before, as it reflects the success (or otherwise) of removing false positive classifications.

Other DoG Parameters

Other aspects of DoG preprocessing besides the scale can be altered. For instance, the sampled values can thresholded, or binarised, to give zero-crossing information, or left to give continuous gradient values, see Section C.1.3, Appendix C, for more details.

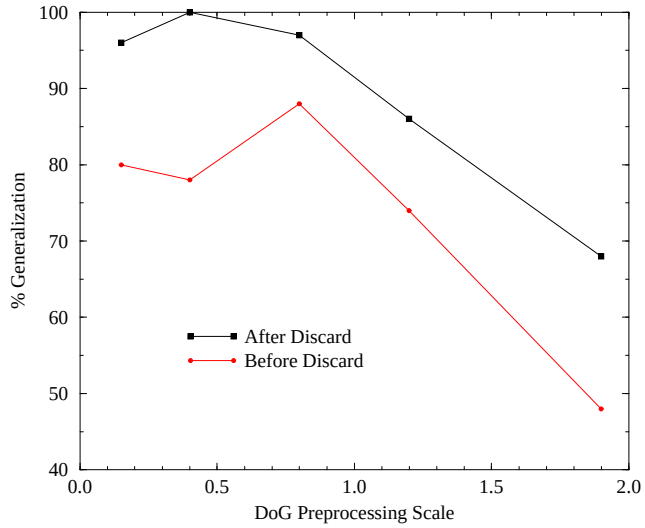


Figure 3.6: Effect on test generalisation for 50/50 RBF networks of changing DoG scale for the preprocessing of 25×25 faces images from Sussex database, before and after discard.
Changes in DoG scale will affect the mask size and, therefore, the amount of data remaining after convolution, see Figure C.2.

Number of Scales	Samples per Image	Thres-holding	Grey-Level Range	Initial %	% Discarded	% After Discard
1	441	No	Full	52	66	71
		Yes	Full	78	52	100
			Reduced	90	22	100
4	1556	Yes	Full	86	40	100

Table 3.5: Test generalisation for 5-example 50/50 RBF networks using non-thresholded (gradient) and thresholded (zero-crossings) DoG preprocessing, with one and four DoG scales.
The single-scale DoG preprocessing used a scale value of 0.4, the four-scale preprocessing used scale values of 0.15, 0.4, 0.8 and 1.3.

Results

Table 3.5 shows the results with all these variations in the preprocessing stage. Training with 'zero-crossings' thresholded data gave better generalisation compared to the non-thresholded 'gradient' data. The use of multiple DoG scales gave a modest improvement in performance, but required four times as much data than for one scale.

The use of data with a reduced range of grey-levels gave a great increased generalisation compared to tests using the full range of grey-levels, but it is an *ad hoc* heuristic at present, taking advantage of the constrained conditions of the Sussex database, and it is unclear how to generalise such a technique to all lighting conditions.

In summary, varying a wide range of parameters in the DoG preprocessing did not seem to affect the results very much, and even using a multi-scale DoG representation did not significantly improve performance. The improvement demonstrated by the use of a reduced range of grey-levels (see Section C.1.2 for details) in the images prior to preprocessing indicates that further benefits may be found by compressing the lower and upper ranges of the grey-level in order to emphasise the central range. More generally, it may be that such incorporation of 'prior knowledge', in this case highlighting detail in skin tones, is an important way of improving the face representation.

3.3.2 Gabor Filter Preprocessing

The second preprocessing technique we are looking at is based on Gabor wavelet analysis (Daugman, 1988). This differs from DoG preprocessing, which only had a scale parameter, by being able to specify both scales and orientations of interest. The face representation we use is then made up of the combination of all the filter coefficients, which means that the number of sampled values used for each image will be greater than for DoG preprocessing. We will reduce the number of sampled values by using a sparse sampling scheme, similar to Gabor 'jets' (see Section 2.4.4), instead of performing a full convolution with the filters.

Section C.2 in Appendix C gives specific details on the mathematical basis of Gabor filters preprocessing, and Section C.2.2 describes the sampling arrangements used for the different schemes, although the main details are presented in Table C.2 and Figure C.5.

Gabor wavelets have been used previously for face recognition (Petkov et al., 1993; Würtz, 1994; Wiskott et al., 1997; McKenna et al., 1997b) but their use can be extremely computationally expensive for sequential machines. The use of sparse sampling with Gabor filter masks offers a simpler and less expensive method of preprocessing.

The Effect of Gabor Orientations

The oriented nature of the Gabor filters is what distinguishes them most from DoG filters. Figure 3.7 shows how the angle used for different single orientation Gabor preprocessing schemes affects test generalisation. It can be seen that performance varies according to angle. Previously , the 90° orientation has been found to give some y-axis pose invariance (McKenna et al., 1996). However, if one anticipates more general pose variation (in other words, over other axes), it may not be appropriate to use Gabor preprocessing with only one orientation.

Results

Figure 3.8 shows how varying the number of orientations for the representation affected test generalisation and discarding of low-confidence classifications. Although the anticipated improvement in performance was found moving from one to multiple orientations, little advantage was found for using more than three orientation angles for the training data.

Table 3.6 shows the generalisation performance for RBF networks for the different 3-orientation Gabor sampling schemes. The 'A' scheme proved to be the most successful arrangement, but unfortunately the 3×3 masks are too small to be considered proper Gabor masks (see Section 3.3.3 for details). Binarisation (thresholding) of the coefficients was found to increase test generalisation for all sampling schemes. All schemes used a full range of grey-levels in the images before preprocessing

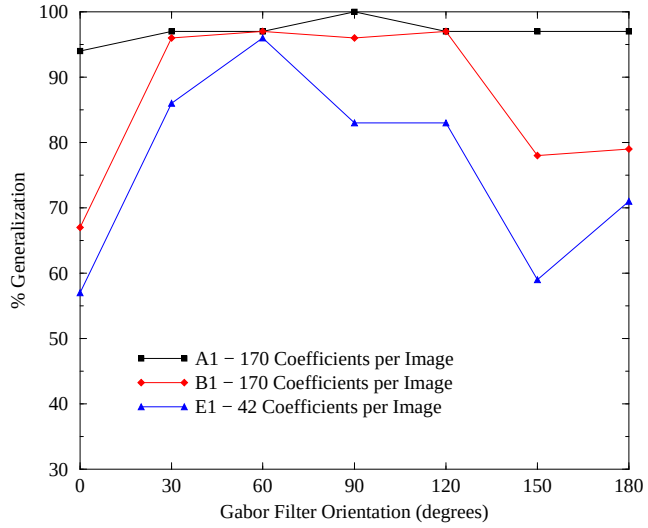


Figure 3.7: Effect of changing the angle of orientation in single orientation Gabor preprocessing on test generalisation after discard for 50/50 RBF networks using 25×25 face images from the Sussex database (see Section C.2, Appendix C, for details of sampling schemes).

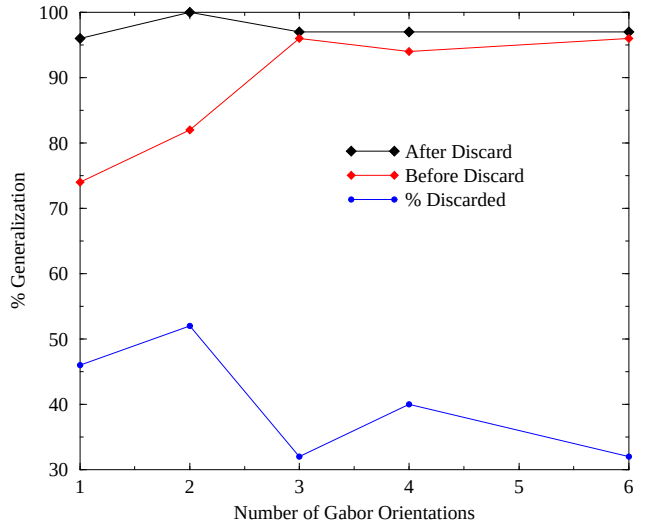


Figure 3.8: Effect of changing number of orientations on test generalisation and discard rates, using Gabor 'B*x*' preprocessing of 25×25 faces images from the Sussex database.

Scheme	Coefficients per Image	Thresholding	Initial %	% Discarded	% After Discard
A3	510	No	88	30	97
		Yes	96	20	98
A3R	510	Yes	94	16	98
A3 (Sine only)	255	Yes	94	20	98
A3 (Cosine only)	255	Yes	48	42	72
B3	510	No	84	44	94
		Yes	96	32	97
B3 (Sine only)	255	Yes	96	28	97
B3 (Cosine only)	255	Yes	66	56	77
C3	510	No	82	40	97
		Yes	90	28	97
D3	420	No	70	46	89
		Yes	82	42	97
E3	126	Yes	92	48	100

Table 3.6: Effect on test generalisation for standard 50/50 RBF networks of different 3-orientation Gabor preprocessing schemes (described in Table C.2 in Appendix C).

except A3R. Interestingly, the advantage found using a reduced range of grey-levels in images for DoG preprocessing (see Table 3.5) was *not* found in tests with Gabor preprocessing.

The coarse nature of the masks at the 3×3 resolution of mask for the A scheme is illustrated by Figure C.4(c), where the real (cosine) masks at different orientations look very similar. Table 3.6 shows the effect of that when the individual masks were separated: the data set with only sine coefficients performs as well as the joint dataset, whilst the data set with only cosine coefficients does not perform well. This is discussed further in Section 3.3.3.

The Effect of Gabor Scales

Tests with individual scales (see Table C.4 in Appendix C for specific details) were made to investigate the effect of individual scales on the overall performance of the network. Identification of redundant scales could significantly reduce the number of coefficients, and therefore the computation required.

Results

Figure 3.9 shows that quite dramatic savings can be made in the amount of information sampled from the images without a large loss of test generalisation or impractical increase in epochs for training convergence. For example, the A3-421 (E3) scheme uses only 126 coefficients, just a quarter of the 510 used for the standard A3, and shows minimal loss of performance. In addition, the effect of the individual Gabor scales was not shown to be additive, so the the A3-8 and A3-421 schemes, for instance, perform similarly even though they contain no common scales or sampling points.

The predicted advantage of overlapping receptive fields (such as for the 'B' and 'C' sampling schemes) over non-overlapping ('A' and 'E') was not demonstrated. The circular sampling scheme 'D' was also found not to be useful, probably because the resolution of the data was not sufficient to give a useful range of sampling positions.

3.3.3 Preprocessing of Low Resolution Images

Task Requirement 3b specifies that the face recognition process must be able to work with low resolution data, and we have chosen to use 25×25 as our main experimental image size. This

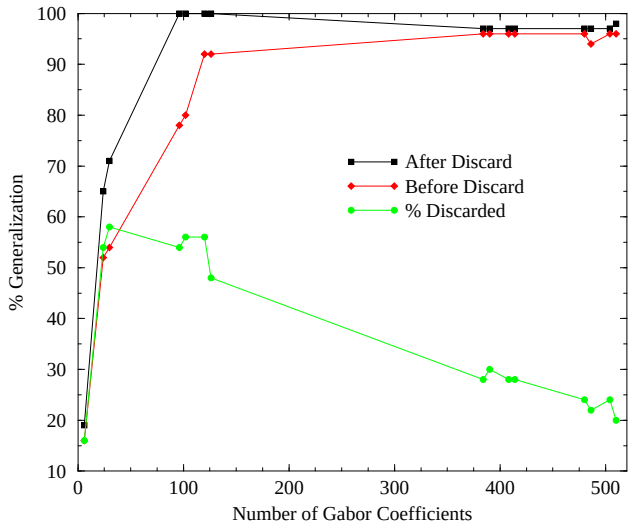


Figure 3.9: Effect on test generalisation and discard rates of changing number of Gabor coefficients through selection of specific scales (see Table C.4, Appendix C, for details) for A3 preprocessing of 25×25 faces images from Sussex database.

decision puts constraints on the nature of the filter-based preprocessing that can be performed, as the number and extent (determined by the filter mask size) of sampling positions within the image will be restricted. This range will be reduced further if samples are required to be collected at discrete, non-overlapping positions, such as in the Gabor 'A' sampling scheme. This limitation led to our initial use of very small filter masks, such as the 3×3, which, as can be seen in Figure C.4(c), are very similar even when the orientation is varied and do not look like conventional Gabor filters (such as in Figure C.4(a)).

There are two reason for this problem: Firstly, the 3×3 mask is made from a Gabor function of period 1. When sampling at unit intervals, the cosine curve with period 1 is everywhere equal to 1, and the sine curve is everywhere 0, and so neither can be made into a significant filter mask. The sine part also has a similar problem with a period of 2. A second problem is that for sampled data, a period of T is the same as a period of $1/(1-1/T)$ by the aliasing theorem, so that periods less than 2 look like periods greater than 2.

The result of this is the approximation of a Gabor function in a 3×3 array is too crude to be called a meaningful Gabor function, and that 7×7 masks (from a period of 3) are the smallest size mask from a proper Gabor function.

A similar problem exists for the smaller DoG masks, though not with the cutoff point encountered with the Gabor filters. The smaller the DoG sigma value becomes, the poorer the approximation. This is because it becomes harder to represent the smooth theoretical shape of a DoG. Figure 3.6 shows that for our task, we can get good performance with the 0.4 and 0.8 scales.

In summary, the low resolution of data has constrained the range of filtering schemes we can use (for Gabor preprocessing, in particular), because only a small range of filter sizes will fit over the 25×25 data. Although our Gabor 'A' scheme was theoretically flawed (though producing useful results in practice), we have been able to show that the substitute 'E' sampling scheme can be used as an ultra-compact replacement.

3.3.4 Discussion

This section has presented results with DoG and Gabor preprocessing of images from the Sussex database using RBF networks.

The Gabor filter preprocessing has been shown to be a more effective representation for generalisation using RBF networks with the Sussex database than the DoG preprocessing, despite limitations brought about through the sparse sampling of the low resolution data. The oriented nature of the Gabor filters has been shown to be more important than the scales for generalisation performance.

The separation of image preprocessing and network training does not have to be as obvious as we have presented it. Both filter-based preprocessing schemes can be visualised alternatively as an extra layer to our network arrangement, similar to a convolutional or a weight-sharing network (Le Cun & Bengio, 1995), sharing between input and receptive field layers (Edelman, 1995).

Figure A.13 in Appendix A shows Euclidean distance graphs using Gabor preprocessing for the same reference images as used for the the DoG preprocessing graphs in Figure A.12(ii). These show a much greater level of class separation for the Gabor graphs than for the DoG graphs, and may explain the superior performance of the RBF networks using Gabor preprocessing (Table 3.6) over DoG (Table 3.2). It can be seen that the 'same class' line is generally well below the majority of the 'other class' lines, which indicates that the Gabor filtering has been able to represent identity through the varying poses more effectively than the DoG, and will therefore be a better choice for our face recognition task.

3.4 General Discussion

This chapter has shown how different factors in the input representation for our face recognition task affects generalisation.

The first section, 3.1, presented an analysis of the Sussex face database in terms of Euclidean distance comparisons, showing how resolution and face pose affected intra- and inter-class distances. This showed that the image resolution did not affect class distinction greatly, but that the extreme pose angles (0° and 90°) were much less distinct than the mid-range images (around 45°).

Figure 3.1 shows the differences for Euclidean distances for images compared with other within its class and with those from other classes. The distances to all the other classes are averaged and also when distances from five different images (rather than just one) are averaged together. This latter graph shows that extra examples of the class will widen the Euclidean distance distinction (on average, at least) between intra- and inter-class images.

Section 3.2 showed how individual elements of the RBF network contribute to the recognition process. The simplest classification arrangement, the winner-takes-all nearest neighbour (WTA NN) classifier, unexpectedly had the highest performance of all the classifiers (if discard for the RBF network is ignored). However, because it throws away information about the activities of all but the winning hidden unit, it is not able to offer a level of confidence that the RBF network can, nor does it have the trainable weight layer, which can give both positive and negative contextual class information, as shown in Figure 3.5. These features in the RBF network combine to provide more valuable classification performance in terms of our Task Requirements, and the RBF network will therefore always be more attractive than the NN as a classifier for our task.

The results using the various classifiers confirms that little advantage is gained through using the higher resolution data, as good generalisation performance can be achieved with a fraction of the data size of 100×100 images. In addition to the benefits of dimensionality reduction in terms of lowered complexity, data storage and computational load are both greatly reduced through using one of the lower resolutions. Because of this, we use the 25×25 resolution as standard for the following work.

We investigated the suitability of the City-Block distance metric for our face recognition task. Although it is simpler computationally, our results using it with the RBF network showed poorer generalisation and a lower discard efficiency than found with the Euclidean measure. In practice, the extra computation has not been found to be excessive, and we have used the latter metric for all following experiments.

Reassigning the Sussex database image classes in order to classify them in terms of specific pose classes rather than identity classes met with less success than the other way around, though a lower precision pose could be extracted ($\pm 15°$).

Finally, Section 3.3 investigated how two preprocessing techniques, Difference of Gaussian (DoG) filtering and Gabor wavelet analysis at a range of scales, affected the face representation and generalisation performance for the RBF network. Although the DoG preprocessing did provide a reasonable level of performance, it was not as good as the Gabor filtering preprocessing. The Gabor representation has been shown to provide good identity class separation, even over a wide range of scales and sampling schemes, with good performance being provided with three equally-spaced orientations. Thus, we are able to replace the flawed 4-scale 'A' scheme (which gave the best generalisation) with the 3-scale 'E' scheme for little penalty in performance.

This chapter has been able to find solutions to Task Requirements 3b (use of low resolution images), 3(c)ii (moderate tolerance to lighting variations), 3(d)ii (recognition of pose-varying images) and 3e (provision of output confidence level). We have touched on the issue of invariance in this chapter in Sections 3.1 and 3.2. The next chapter will look at invariance more closely, focussing on shift, scale and pose invariance in particular.

Chapter 4

Invariance Properties of the RBF Network

This chapter explores the invariance characteristics of the RBF network, looking at how tolerant it is to particular forms of image variation, and how this is affected by the preprocessing of the input data. It is important to know how robust our system is to the variation anticipated for the main task, as this will determine the accuracy of face segmentation and preprocessing computational load required for data to be learnt or recognised.

The experiments in the first half of the chapter are designed to show how well the RBF network can learn identity and generalise to novel images with data where the pose varies. For instance, can profile images, where eye information from the far side of the face is occluded, be generalised to front views? This will determine over how wide a pose range the system will be effective, and the optimal paving of pose examples in the face 'view sphere' (discussed in Section 2.2). The second half of the chapter investigates how 2-D shift and scale variations in the image affect this process.

The property of 'invariance' can be seen at different processing stages. Not only can the data representation be thought of as being invariant to various forms of image variation, but the processing and reasoning performed on that representation can also give further invariance. For instance, a 'foveal' space-variant representation, discussed in Section 2.3.3, can give rotational and scale invariance simply by the nature of the representation. In addition, a preprocessing stage, such as the Gabor filtering stage, will give scale invariance, and a reasoning process, such as a weight-sharing or convolutional network (see Section 2.5.6), can give shift (translation) and deformation invariance. Each stage can be seen to be contributing different aspects of invariance, but it is not easy to isolate the characteristics of these stages, as they need to be considered together to give a coherent view of the entire scheme.

Two basic types of invariance to a particular parameter, such as illumination or head pose, can be distinguished: an *inherent invariance* which is present in any representation, processing or reasoning stage, and a *learnt invariance* which can be obtained during the learning stage by the use of training with suitably varying example images. Inherent invariance can be observed if the network is trained with images which do not exhibit variation in the parameter, whereas learnt invariance requires training images with examples of the variation.

The basic ability gained from the RBF network is that of interpolation between examples. Naturally, this technique will not be utilised in the inherent invariance tests that follow, as these use unvarying data (one example per class) and so there is nothing to interpolate. Therefore, we do not expect to see high performance from the network in these tests, as they will be extrapolating from the single training example to the test images. From the learnt invariance tests, we will be determining what intervals through the varying data being tested provide optimal interpolation from the hidden units and therefore give the best invariance performance.

4.1 Test Details

All the experiments in this chapter use the 100 image, 10 person 'Sussex Database', for details see Section A.2, Appendix A. This database has been designed to test recognition abilities for faces over a 90° range of poses from frontal to profile, see Figure A.5 for example. A pixel-based representation of the 2-D image, used as a 1-D vector for input to the network, will not provide any particular invariance to image variation by itself. It will be the preprocessing and reasoning stages that provide the necessary invariance. To compare and contrast the effects of preprocessing without a large number of results, most of the tests will concentrate on two applications of the DoG and Gabor techniques discussed in Section 3.3 of the previous chapter:

Single-scale Difference of Gaussians (DoG) filtering This is performed as a convolution of the image with a 2-D DoG filter mask of a single scale factor (0.4), with thresholding to give binary zero-crossing information. Each processed image has 441 samples, corresponding to a 21×21 convolution of the original 25×25 image.

Gabor filtering This is 2-D Gabor wavelet analysis at four scales and three orientations (termed 'A3' in Section 3.3.2). Each processed image has 510 coefficients, corresponding to the outputs of the different scaled and oriented filters at different positions.

4.2 Pose Invariance

Task Requirement 3(d)ii specifies an invariance to pose, and so it is important to test our system to determine what limits it has in this respect. In our potential environment, the subjects are allowed unrestricted movement around the room, and therefore will be visible at any pose angle towards the camera that is physiologically possible for the head around the vertical (y-) axis. Obviously, views such as the back of the head are not learnable, in terms of identity, especially as the requirements specify an invariance to hair style (Task Requirement 3(d)iii).

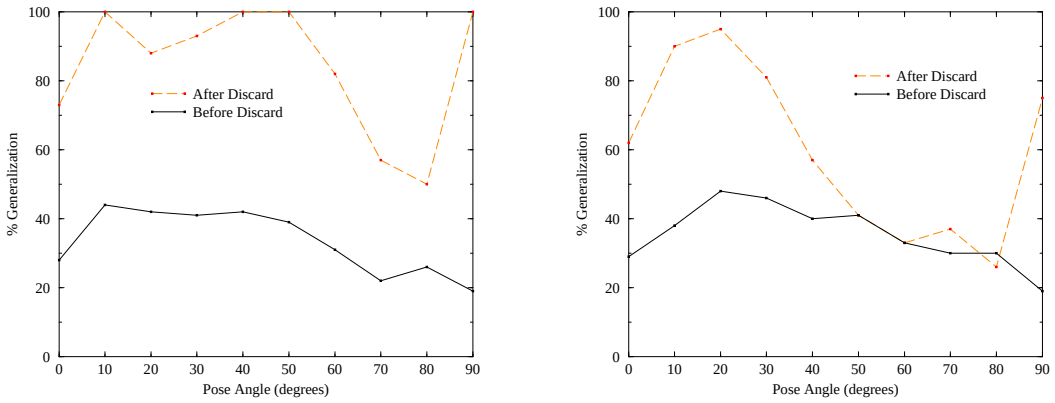
A useful system in an unrestricted environment should be expected to cope with the full range of views that contain facial information, which is roughly $\pm120°$, where $0°$ is the frontal view. Such a wide pose range is in contrast to many face recognition systems which do not explicitly dealt with pose, preferring to restrict data to face images with very slight pose variation (typically $\pm15°$), which can be approximated as linear. RBF networks, in view of their interpolation properties, should allow some pose invariance (given sufficiently close examples for effective interpolation), but the extent of this will need to be determined empirically.

In this section, we will be testing the RBF network for two types of pose invariance by training with two different arrangements of the data examples: the first searches for *inherent invariance* by training with unvaried images (in other words, one fixed pose for all classes) and testing with varied images only (all the other poses not seen during training), the second is looking for *learnt invariance* by training with explicit examples of pose variation.

4.2.1 Inherent Pose Invariance

The pose invariance that we have termed 'inherent' in this section is the generalisation obtained when the RBF network has been trained with images that have no pose variation (that is, they all come from one fixed pose position), and is then tested with images of different pose to that used for training.

When testing for inherent pose invariance with the Sussex database, where all images for each class have a different pose angle, there can only be one image per class available for training. This type of training will produce a 10/90 RBF network, with 10 training examples (one per class) and 90 test images.



(a) DoG preprocessing, average discard 88%

(b) Gabor preprocessing, average discard 65%

Figure 4.1: Inherent Pose Invariance: Test generalisation over face pose view with 10/90 (trained with one image per class) extrapolating RBF networks, the training view varying over pose.

Results

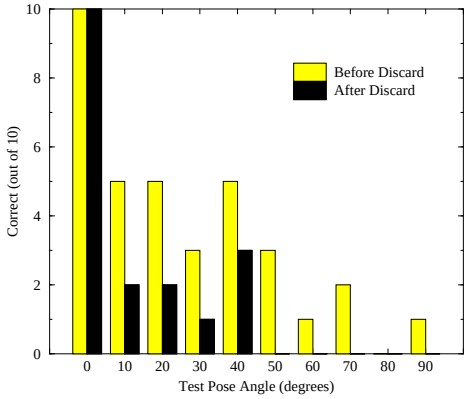
Figure 4.1 shows how the specific pose angle used to train the 10/90 network affects the test generalisation, and it can be seen that network performance is rather poor whatever the pose angle used to train it. As discussed above, this network arrangement is effectively being tested for pose *extrapolation*, as only one example of each class is available for learning. This kind of extrapolation from data is not a particular strength of RBF networks, as they are much better suited to interpolation. We expect performance for the interpolating RBF networks in the next section to be significantly better than these extrapolating networks.

The chaotic nature of the graph lines for the 'after discard' values in Figure 4.1 reflects how little information remains after extremely high discard levels. It is more informative here to compare the 'before discard' performances, which are roughly similar in shape. The front to middle range of poses, around 15–45°, appear best for generalisation. This corresponds to all the major facial features being visible on the face, without being flattened or foreshortened as they are at more oblique angles. This may have a similar basis to psychological findings on a 45° or '3/4 view advantage' in face recognition (Bruce et al., 1987), see Section 3.1.1 for further discussion.

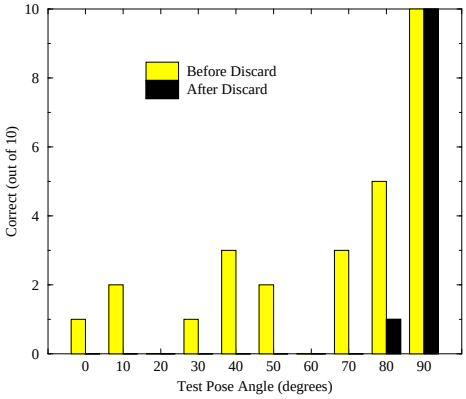
A different view of the RBF network's ability to extrapolate can be seen in Figure 4.2, which shows how the numbers of correct classifications vary by pose angle of the test images for varying pose angle in the images used to train the network. As could be expected, the 40° trained network (c), where the test images span a ±45° range relative to the trained image, does better than the 0° (a) and 90° (b) trained networks, which have to generalise to test images up to 90° from the trained image. The graphs clearly show how generalisation tails off as the angular difference between the train and test image increases.

In addition, it can be seen in Figure 4.2 that the frontal views (0–40°) have an advantage, in terms of generalisation, over the profile end of the pose range (50–90°) regardless of the training pose angle. As mentioned above, such an advantage may be due to images in the former pose range containing a greater area of facial information (more pixels representing part of the face) than those in the latter range.

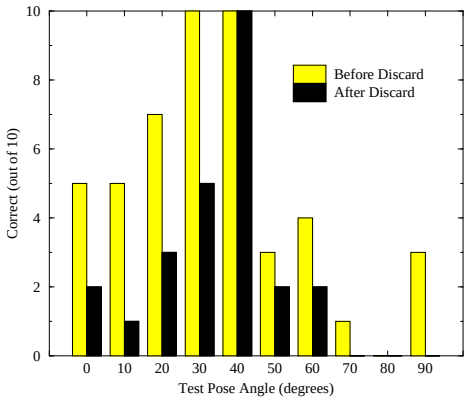
Finally, Figure 4.2 also shows how the discard measure removes a large proportion of the correct classifications in order to eliminate false classifications, especially in (b) (compare these to the generalisation rates after discard in Figure 4.1). The lower the black bars are in comparison to the shaded bars, the less of the original correct output is being retained. As discussed in Section 3.2.3 and shown in Figure 3.2, the ratio of lost true positives to discarded false positives can be adjusted according to requirements. In general, we expect large amounts of data and can justify a fairly high discard rate, but in this case, it is not clear there is enough useful data remaining after discard.



(a) Trained with pose angle 0°



(b) Trained with pose angle 90°



(c) Trained with pose angle 40°

Figure 4.2: Inherent Pose Invariance: Number of correct classifications (out of 10) of test images at specific training pose angles for 10/90 extrapolating RBF networks with DoG preprocessing.

Network	Number of Training examples	Number per Class	Number of Test images
20/80	20	2	80
30/70	30	3	70
40/60	40	4	60
50/50	50	5	50

Table 4.1: The four different types of interpolating RBF networks, used to test learnt pose invariance.

Pre-processing	Network	Training Examples	Training Pose Angles (°)	Initial %	% Discarded	% After Discard
DoG	20/80	2	20,70	48	73	91
	30/70	3	20,50,70	61	56	94
	40/60	4	10,30,60,80	67	67	100
	50/50	5	10,30,50,70,90	78	52	100
Gabor	20/80	2	20,70	71	51	95
	30/70	3	20,50,70	80	39	98
	40/60	4	10,30,60,80	88	35	97
	50/50	5	10,30,50,70,90	96	20	98

Table 4.2: Learnt Pose Invariance: Effect on varying number of training examples on test generalisation for interpolating RBF networks with DoG and Gabor preprocessing, both before and after discarding of low-confidence classifications.

Summary

Although there does seem to be some inherent pose invariance in the RBF network, it does not seem to be very controllable due to the low number of example training views. This means that the extrapolating RBF networks we used in this section are only usefully invariant over a pose range of about $\pm 20°$.

It should be noted that greater pose invariance is expected if interpolation between trained views is used. The next section will be testing this type of interpolating RBF network to confirm this expectation.

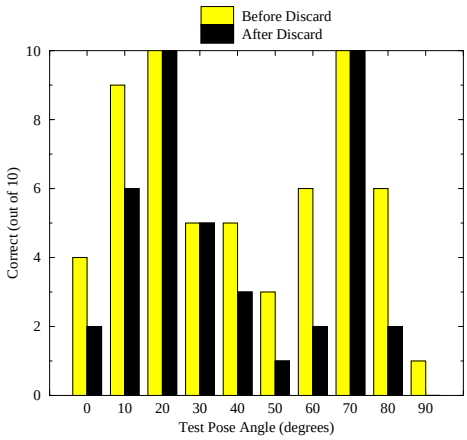
4.2.2 Learnt Pose Invariance

This section presents experiments where the RBF network learns face-class information from more than one example for each class. This will allow better generalisation than for the extrapolating networks above, as these networks can interpolate between training views of the same person at different pose angles.

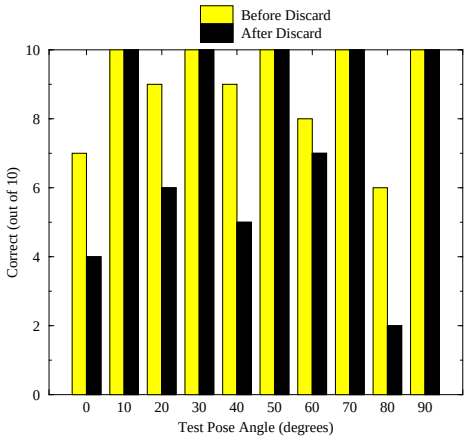
The network configuration allowed for this type of test is much less constrained than for the extrapolating networks in the previous section (they were confined to one network size due to the nature of the Sussex database), and we are able to perform experiments with four types of interpolating RBF networks, ranging from 2 to 5 training examples per class. The specific details for these are in Table 4.1.

Results

Table 4.2 shows the results for the four types of interpolating RBF networks with fixed selections of training pose angles. All configurations provided good levels of generalisation performance, especially after discard. This confirmed the expectation that interpolation between training examples is crucial for effective use of the RBF hidden units for pose invariance.



(a) 20/80 networks trained with pose angles 20° and 70°.



(b) 50/50 network trained with pose angles 10°, 30°, 50°, 70°, and 90°.

Figure 4.3: Learnt Pose Invariance: Number of correct classifications (out of 10) of test images at specific pose angles for interpolating RBF networks with DoG preprocessing.

This table also shows that the main advantage gained by adding more training examples per class was in the reduction in number of classifications discarded through low confidence, rather than an improvement in the generalisation rate (although there were more correct classifications made).

Although the 50/50 RBF network using DoG preprocessing has a slightly higher generalisation rate after discard than the network using Gabor preprocessing, it has a much higher discard rate. The result of this is that it gets 24 correct out of only 24 left after discard, rather than 39 out of 40 for the Gabor preprocessed data, so there was less useful information overall. A better, more controllable, confidence measure is developed using these 50/50 networks in Section 5.2 of the next chapter.

Figure 4.3 shows how the number of correct classifications for specific 20/80 and 50/50 networks vary according to the pose angle of the test images. It can be seen that there is slightly better generalisation performance for images from the frontal views (0–40°) than for the profile end of the pose range (50–90°).

This higher level of generalisation for the interpolating RBF networks with the frontal, rather than profile, views is a similar response to that seen in the previous section (see Figure 4.2) with the extrapolating RBF networks. However, the interpolating networks here are different in that they were able to maintain some generalisation performance over all pose angles for test images, whilst the extrapolating networks completely failed to recognise test images at some pose angles, for instance 80° in Figures 4.2(a) and (c).

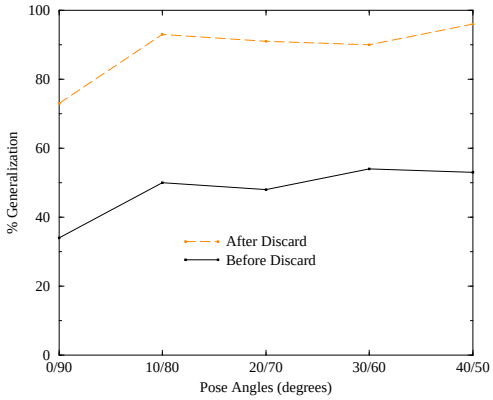
Influence of Training Selection in Learning Pose

To compare the interpolating RBF networks with the earlier extrapolating networks in a more direct way, a range of two-example per class 20/80 and three-example 30/70 networks were tested, having been trained with differing pose selections, similar to the tests done in Section 4.2.1 for inherent pose invariance, from very widely spaced intervals to very close intervals between the training examples.

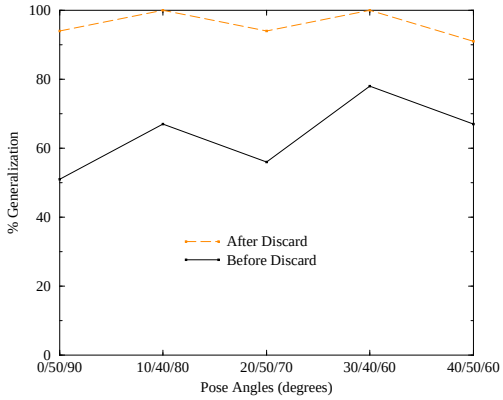
It was immediately clear that the behaviour of these two networks, in Figure 4.4, with two and three examples per class, was less erratic than the 10/90 (see Figure 4.1), which only had one per class. In addition, these networks were able to give a more useful level of generalisation, especially after discard.

Summary

We have been able to show learnt pose invariance in interpolating RBF networks, which have several training examples (each of different pose) for each class. These networks were able to

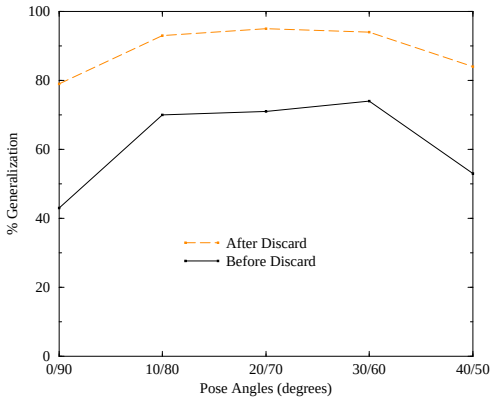


(i) 20/80, average discard 77%.

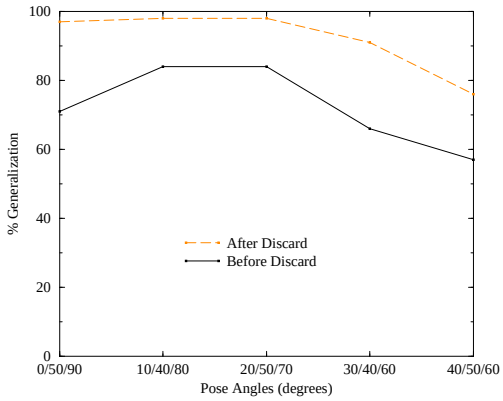


(ii) 30/70, average discard 68%.

(a) DoG preprocessing



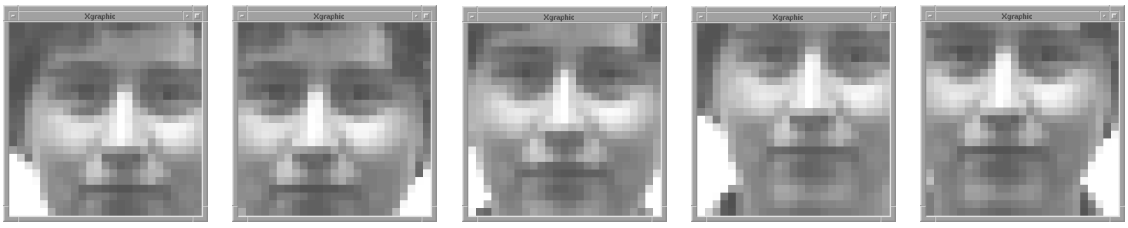
(i) 20/80, average discard 63%.



(ii) 30/70, average discard 48%.

(b) Gabor preprocessing

Figure 4.4: Learnt Pose Invariance: Test generalisation with 20/80 (trained with two images per class) and 30/70 (three per class) interpolating RBF networks, varying over selections of pose angles: from left to right, widely to closely space intervals.



(a) top left (b) top right (c) normal view (d) bottom left (e) bottom right

Figure 4.5: Example shifted versions of the original front view of one individual from the Sussex database, used to test for shift invariance.

generalise effectively when tested with images with head pose that lies between that of at least two of the training examples.

The increase in generalisation performance from the two- (20/80) to three-example (30/70) networks indicates that a good level of pose invariance is provided by each RBF hidden unit over a pose range of $\pm 20°$, with $\pm 15°$ providing a high level of generalisation.

4.2.3 Discussion

The RBF network has a limited inherent pose invariance, due to its poor extrapolation ability, although this can be improved via preprocessing (see Figure 4.1). The key to enhanced performance, both in higher generalisation and lower discard rates, is in providing training examples within the anticipated test pose range, using a suitably close interval between training pose angles, so that the RBF network can interpolate effectively.

4.3 Shift and Scale Invariance

This section tests the RBF network with some image variations that are likely to be encountered in real-life data, where automatic face localisation will not always be exact. Two type of errors may occur in an automatic localisation stage of the processing of face images:

1. The face may be incorrectly centred.
2. The face size may be incorrectly determined.

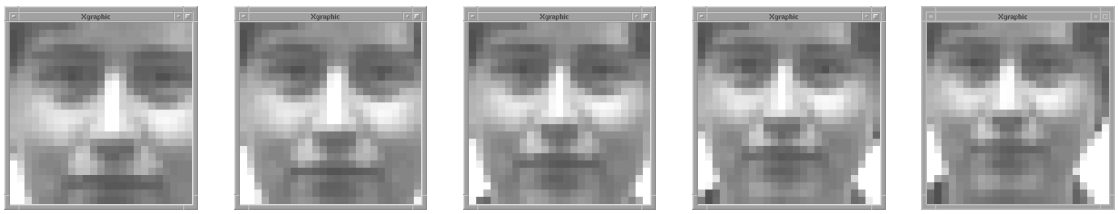
These errors correspond to the two specific modes of image variation under which we will be testing generalisation:

1. A translational shift of the face, so that the face is no longer centred compared to the standard 'nose-centred' face position determined for the Sussex database (we assume this original position was correctly registered, see Section 3.1.3 for further details).
2. A scale variation of a normally centred face, so that the face is no longer the same size compared to the standard face size determined for the Sussex database.

It is important to know, in each case. how much invariance can be expected from the RBF network, so that trade-offs between explicit processing stages for specific types of invariance can be made, and the bounds on localisation accuracy determined. Minor tolerance to these two variations is task requirement 3(c)i from Chapter 2.

4.3.1 Shift- and Scale-Varying Data

In order that each specific mode of 2-D variation in the images could be studied separately, each were isolated by creating two new data sets of 500 images each from the original 100-image Sussex dataset:



(a) +25%, 111×111 (b) +12.5%, 107×107 (c) normal, 100×100 (d) –12.5%, 94×94 (e) –25%, 87×87

Figure 4.6: Example scaled versions of the original front view of one individual from the Sussex database, used to test for scale invariance, with relative size to the normal sampling area, and size of window grabbed from (in pixels).

Variation	Network	Pre-processing	Initial %	% Discarded	% After Discard
Shift	100/400	DoG	12	89	7
		Gabor	35	82	47
Scale	100/400	DoG	48	76	76
		Gabor	83	36	88

Table 4.3: Inherent Shift and Scale Invariance: Effect on test generalisation for the RBF network of different variations in the dataset, both before and after discarding of low-confidence classifications: networks trained with all ten non-varied versions of poses for each person and testing with varied versions (100 training and 400 test images).

- A *shift-varying* data set with five copies of each image: one at the standard sampling 'window' position, and four others at the corners of a box where all x,y positions were ± 10 pixels from the centre (see Figure 4.5).
- A *scale-varying* data set with five copies of each image: one at the standard sampling 'window' size of 100×100, and four re-scaled at $\pm 12.5\%$ and $\pm 25\%$ of its surface area, ranging from 87×87 to 111×111 (see Figure 4.6).

Similarly to the previous section, where we dealt with pose invariance, we experiment with the RBF network with two different types of data examples: the first searches for *inherent invariance* by training with original images only, the second is looking for *learnt invariance* by training with shift and scale varying images.

4.3.2 Inherent Shift and Scale Invariance

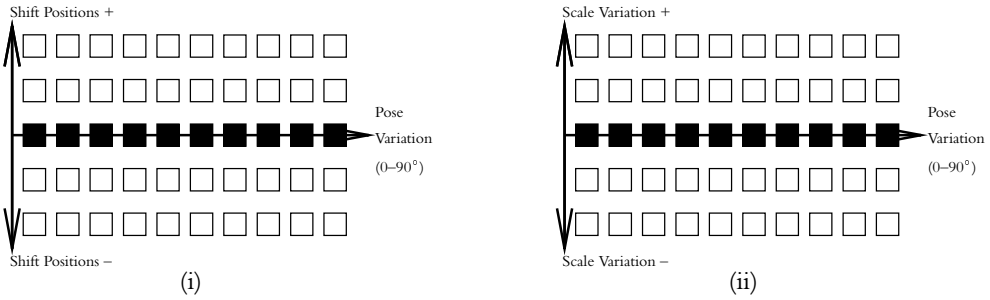
The experiments in this section are testing for the inherent shift and scale invariance in the RBF network using the Sussex database. Inherent invariance is the generalisation exhibited by the network to test images of a particular type of variation, in the absence of exposure to that variation through explicit training examples.

To test for this intrinsic shift or scale invariance, only the original image from each group of five (see Figures 4.5 and 4.6) is used for training, the four varied ones being reserved for testing; see Figures 4.7(a)(i) and (ii) for a diagram of how this is done. This means that for all experiments in this section, there are 100 training and 400 test images.

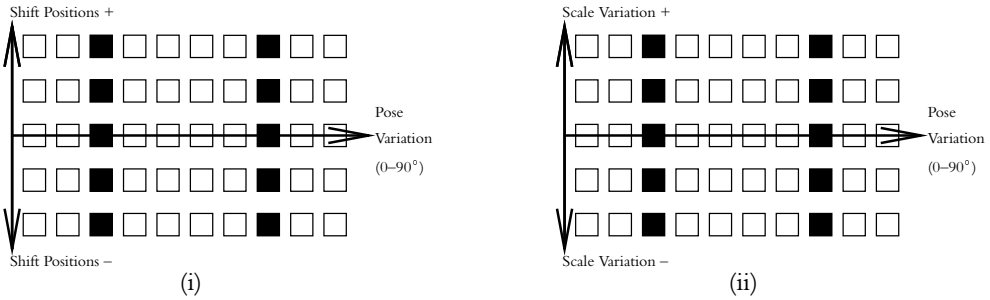
Results

From the results in Table 4.3, it is immediately obvious that the RBF networks trained with no shift or scale variations performed very differently when tested with the shift rather than with the scale varying data.

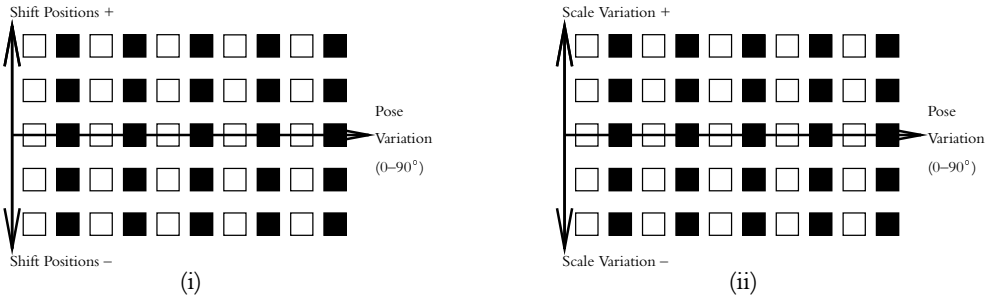
In the absence of explicit training examples, there was a complete failure of generalisation in the network using the shift-varying test data with DoG preprocessing, as performance after discard



(a) *Inherent invariance* tests using original face images from all 10 pose angles: 10 training (black) and 40 test (white) images.



(b) *Learnt invariance* tests with images from two pose angles: 10 training and 40 test images.



(c) *Learnt invariance* tests with images from five pose angles: 25 training and 25 test images.

Figure 4.7: Selection of training and test data from the 50 images available for each person with the (i) shift and (ii) scale varying data.

Variation	Network	Pre-processing	Initial %	% Discarded	% After Discard
Shift	100/400	DoG	38	85	85
		Gabor	69	60	89
	250/250	DoG	52	71	92
		Gabor	85	35	98
Scale	100/400	DoG	44	77	78
		Gabor	64	55	88
	250/250	DoG	66	57	94
		Gabor	90	26	97

Table 4.4: Learnt Shift and Scale Invariance: Effect on test generalisation for the RBF network of different variations in the dataset, both before and after discarding of low-confidence classifications: networks trained with all five shift or scale-varied versions of two (100/400) or five (250/250) equally spaced poses for each person.

(7%) being even lower than random (10%). The network using Gabor preprocessing was able to give a low level of useful generalisation.

In contrast, the scale-varying test data appears to be much easier for the network to generalise to, even without explicit training examples, and a useful level of performance was obtained with both types of preprocessing. As before, the Gabor preprocessed training data was easier to learn and generalise with compared to the DoG preprocessed data, and networks using the former were able to give a high level of generalisation performance, even without discard.

Summary

This section has shown that the RBF network has a significant inherent invariance to scale differences with the Gabor preprocessed face data from the Sussex database, and a moderate invariance with the DoG preprocessed data. In marked contrast, the shifted images were very much harder for the network to generalise to with both preprocessing techniques.

Figures 4.8(a) and (b) show that these differences primarily arise out of the choice of preprocessing, although the scale transformation also seems to alter the image vector less than the shift transformation. This is shown by the 'other class' line for the Gabor scaled images, (b)(ii), being noticeably further away from the 'same class' line than for the other combinations of transformation and preprocessing.

4.3.3 Learnt Shift and Scale Invariance

The experiments in this section test for learnt shift and scale invariance. As before, they use a fixed selection of pose positions for training examples, but this time use all five versions (4 varied, 1 unvaried) of each original image. This helps the network to learn about the shift and scale image variation during training and thus develop a learnt invariance. The difference between the generalisation performance found in the previous section (with inherent invariance) and in the tests in this section will be due to this learnt invariance.

Two levels of training are used in this section, corresponding to the two- (Figures 4.7(b)(i) and (ii)) and five-example (Figures 4.7(c)(i) and (ii)) networks for pose invariance in Section 4.2.2. These use 10 and 25 training images for each class, creating 100/400 and 250/250 networks respectively.

The first level of training is used to allow a direct comparison with the results from the previous section, as it uses the same number of training examples. The second is used to establish whether the same level of performance improvement seen in Section 4.2.2 for five-example pose invariance networks over two-example networks (see Table 4.2) would be repeated for the shift and scale invariance networks.

Results

The results in Table 4.4 shows that not only can the RBF network learn identity in spite of pose variations, but it can continue to be invariant to pose in the presence of other variations. In addition, it can learn an invariance to scale variation more easily than shift as shown by both the better classification performance and the lower discard rates.

Interestingly, the *inherent invariance* 100/400 network using Gabor preprocessed data, with a generalisation rate of 83% before discard (shown in Table 4.3), was able to perform better than the *learnt invariance* 100/400 network using Gabor preprocessed data, with only 69% before discard (shown in Table 4.4).

The 5-pose example 250/250 networks gave high levels of generalisation with both types of preprocessing.

Summary

These results suggest once more that the use of only two learnt pose views for training the RBF network is not the most efficient arrangement for good pose generalisation, especially when this is required in addition to scale invariance, and that three or more are required for robust performance. This is backed up by the superior performance exhibited by the 5-pose example 250/250 networks, which corresponds to the high performance obtained from the learnt pose invariance 50/50 network (Table 4.2). This indicates that there are sufficient pose examples in the 250/250 networks to interpolate efficiently.

It should be noted that this discussion of an absolute number of examples required for good generalisation is only relevant within the specific context of the 90° pose range encountered in the Sussex database, as other ranges in different data will obviously require differing amounts of training views.

4.3.4 The Contribution of Multi-Scale Preprocessing

The Euclidean distances graphs in Figures 4.8(a) and (b) indicate that the Gabor preprocessing does separate the within-class images from the other-class images more than the DoG preprocessing for the shift and scale varying data, although the effects appear quite small.

To investigate further why there is such a big difference between the two preprocessing techniques when using the shift and scale-varying data, further tests were made to determine if it was the multi-scale nature of the Gabor preprocessed representation that gave its advantage, rather than the design of the filters themselves. These used two variants on the DoG and Gabor preprocessing schemes used previously:

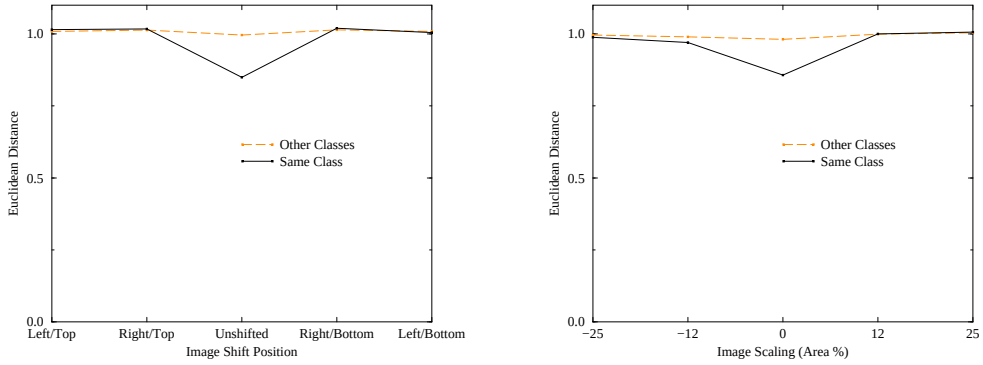
Multi-scale DoG Preprocessing This used 4 scale factors (0.15, 0.4, 0.8 and 1.3) to give 1556 samples per image (compared to 441 for the normal single scale representation).

Gabor A6 Preprocessing This is the same as used in Chapter 3, having six orientations and four scales, and has 1020 coefficients per image (the standard A3 Gabor preprocessing having 510, see Table C.2). This was used simply as a control to compare to the multi-scale DoG method, as it had similar numbers of data values per image.

Results

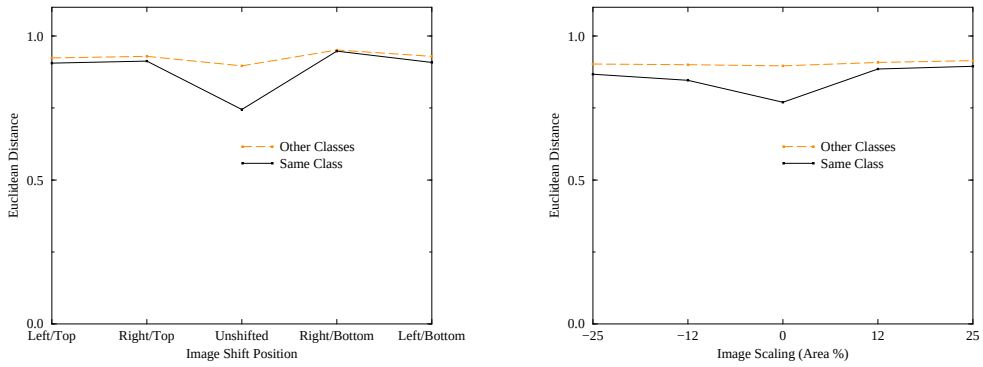
The results using the multi-scale version of the DoG preprocessing for the shift and scale-varying data (Tables 4.5 and 4.6) are very similar to the original tests using the normal DoG and Gabor schemes (Tables 4.3 and 4.4), bearing in mind that more data samples were provided, which should have improved the learnability of the task. Of course, no amount of extra data will help if the representation is not appropriate to the task.

If it is not the different scales in the representation which are improving generalisation, it must be the oriented nature of the Gabor preprocessed data which is significant. There is a finite limit to its usefulness, as shown by the results in Figure 3.8 in Section 3.3.2 and here. These indicate that performance is not enhanced through the use of more than three orientations.



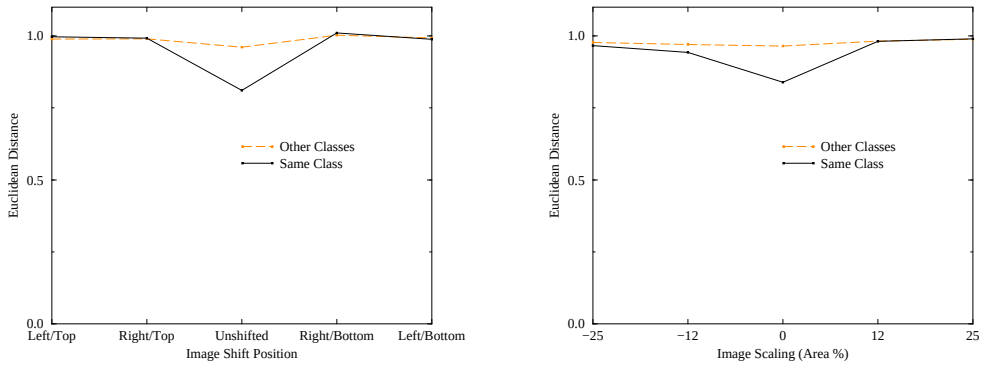
(i) (ii)

(a) Single scale (normal) DoG preprocessing



(i) (ii)

(b) Multi-scale (normal) Gabor preprocessing



(i) (ii)

(c) Multi-scale DoG preprocessing

Figure 4.8: Euclidean distances for images from the Sussex database to same- and other-class images, averaged over all classes, varying over specific shift and scale variations, (i) shift, (ii) scale, and different types of preprocessing.

Variation	Network	Pre-processing	Initial %	% Discarded	% After Discard
Shift	100/400	Multi-DoG	9	88	4
		Gabor A6	37	82	58
Scale	100/400	Multi-DoG	57	66	77
		Gabor A6	78	48	94

Table 4.5: Inherent Shift and Scale Invariance with extra data values: Generalisation rates for RBF networks trained with all ten non-varied versions of poses for each person and testing with varied versions (100 training and 400 test images).

Variation	Network	Pre-processing	Initial %	% Discarded	% After Discard
Shift	250/250	Multi-DoG	67	62	97
		Gabor A6	85	40	95
Scale	250/250	Multi-DoG	80	44	99
		Gabor A6	87	31	99

Table 4.6: Learnt Shift and Scale Invariance with extra data values: Generalisation rates for RBF networks trained with five shift or scale-varied versions of five equally spaced poses for each person (250 training and 250 test images).

Summary

It was concluded that providing a range of scales in the DoG preprocessing stage will not, in itself, provide the representation with the same generalisation power that the Gabor process did. This is backed up by the similarity of the Euclidean distance values between the single and multi-scale DoG preprocessed images in Figures 4.8(a) and (c).

In comparison, adding extra orientations to the Gabor preprocessing for the A6 data also did not greatly improve performance over that already provided by three orientations in the A3 data. This matches the results shown in Figure 3.8, where varying the number of Gabor orientations did not dramatically affect the results.

4.3.5 Discussion

The RBF networks seem to have quite powerful inherent scale invariance, but no inherent shift invariance. Both image variations could be learnt to give a high level of generalisation if given suitable training examples.

Gabor preprocessing allows greater invariance to these transformations than the DoG preprocessing. This effect is most pronounced with the scale variations, which is not surprising considering the multi-scale nature of the data representation. However, we have discounted that it is this aspect alone of the Gabor preprocessing which gives it its power, as the tests using a multi-scale DoG representation showed that the explicit extra information at different scales gave no improvement in network performance over that obtained with training without these data values. Indeed, the RBF networks using the multi-scale DoG preprocessing were unable to improve on those using single scale DoG preprocessed data.

It might be expected that it was the oriented nature of the Gabor preprocessed data which gave it an advantage in learnability over the DoG preprocessed data. However, such an advantage can not be controlled through varying the number of orientations, as Figure 3.8 shows that performance remains fairly constant in such circumstances. This does not mean it is cannot be the oriented nature of the data that gives the advantage, but that there is a limit to how much useful information can be extracted with a single technique.

4.4 General Discussion

Chapter 3 showed how the representation used for input data can have a profound effect on the ability of the RBF network to generalise from a learnt task. This chapter has developed these ideas to analyse how specific variations in the image will affect such generalisation.

The experiments in Section 4.2 looked at invariance to head pose, a complex 3-D image transformation. The results show that this is not inherent to the RBF network over any wide pose angle range, but that it can be effectively learnt if appropriate training examples are available to give high levels of test generalisation. What this implies is that even if we cannot create an image representation that is (inherently) pose-invariant over large angles, we can produce an RBF hidden layer representation with useful (learnt) pose invariance.

Section 4.3 dealt with shift and scale transformations, both simple, 2-D image-plane variations. The experiments show that both can be learnt and give high levels of generalisation performance. The Gabor preprocessing appears to be more effective, as discard rates for networks using this are much lower than those with DoG preprocessed data.

In marked contrast to its invariance to pose and shift variation, the RBF network has been shown to have a useful inherent scale invariance for this kind of data, as the Euclidean distances between intra-class and inter-class images are more distinct (Figures 4.8 and 3.1). This is clearly indicated by the good generalisation (only slightly lower than the 5-example (250/250) learnt rates) in conditions where the network completely failed with the shift-varying data. This means that fewer training examples need to be explicitly used to teach the network about image scale variation. In addition, although a rough approximation may be sufficient when localizing the face during preprocessing, it is important to correctly register it so that the face region is accurately centralised within the image.

The experiments in this chapter have been addressing the specific Task Requirements 1b, 3(c)i and 3(d)ii from Chapter 2. It can been seen that for moderate levels of the variations expected in the unconstrained identification task, such as in pose, scale and shift, the RBF networks can be trained to provide very high levels of generalisation performance (sufficient to be capable of supporting the main task). The Sussex database only has a 90° range of pose variation, but Chapter 6 will show how greater pose ranges can be dealt with by the RBF network, in tests using real-life image sequences. These image sequences will also contain some of the other variations not tested here, such as lighting and expression.

The next chapter will be concerned with how variations in the structure of the RBF network itself can enhance its performance and allow greater control over the discarding procedure.

Chapter 5

Face Unit RBF Networks

This chapter introduces a different way of learning the face recognition task through the reorganisation of the standard RBF networks into a group of smaller 'face recognition units', each trained to recognise a single person. This type of system organisation allows flexible scaling up which could be used either by itself or in conjunction with a standard RBF network trained on all classes where the combined decisions might give greater reliability.

The concept of *face recognition units* was suggested in the perceptual frameworks for human face processing proposed by Hay and Young (1982) and Bruce and Young (1986). We are adopting this face unit concept as a useful way of developing a modular, scalable architecture, creating fast small RBF networks trained with examples of views of the person to be recognised. The face unit network uses these views of the person to be recognised as positive evidence together with selected confusable views of other people as the negative evidence, which are linked to just 2 outputs corresponding to 'yes' or 'no' decisions for the individual. This training using explicit negative examples is in contrast to the HyperBF network scheme used by Edelman et al. (1992), who preferred to use implicit negative evidence in their study (see Section 2.5.5).

For each individual, an RBF network is trained to discriminate between that person and others selected from the data set. Rather than using all the data available from the other classes to train the network against an individual, the strategy adopted was to use only negative data that was most similar (using an Euclidean distance metric) to the positive data. This strategy is based on the assumption that similarity leads to confusion, so the inclusion of this type of negative evidence in the training should improve discrimination. This data would be the hardest to learn to discriminate 'for' and 'against' the individual, since it would be the most ambiguous.

The reduction in the size of the network using the face unit organisation plus the use of negative knowledge should allow a more efficient coding of the information. Furthermore, people can be added to the data set of a trained set of networks by the creation of a new 'face unit' network for each new individual to be added without retraining the entire database, as the reorganised scheme is completely modular. In the standard RBF network, a new individual means a complete retraining with the expanded dataset.

5.1 The Face Unit Network Model

The face unit network is essentially a normal RBF network with two output units, see Figure 5.1, which produces a positive signal only for the particular person it is trained to recognise. It differs from the RBF networks used in previous chapters only in the selection of training data, the data for the face unit network being manipulated to present a many-class problem as a two-class problem: 1) a particular class and 2) all others.

Unlike the standard RBF network used in Chapters 3 and 4, with positive output signals (one

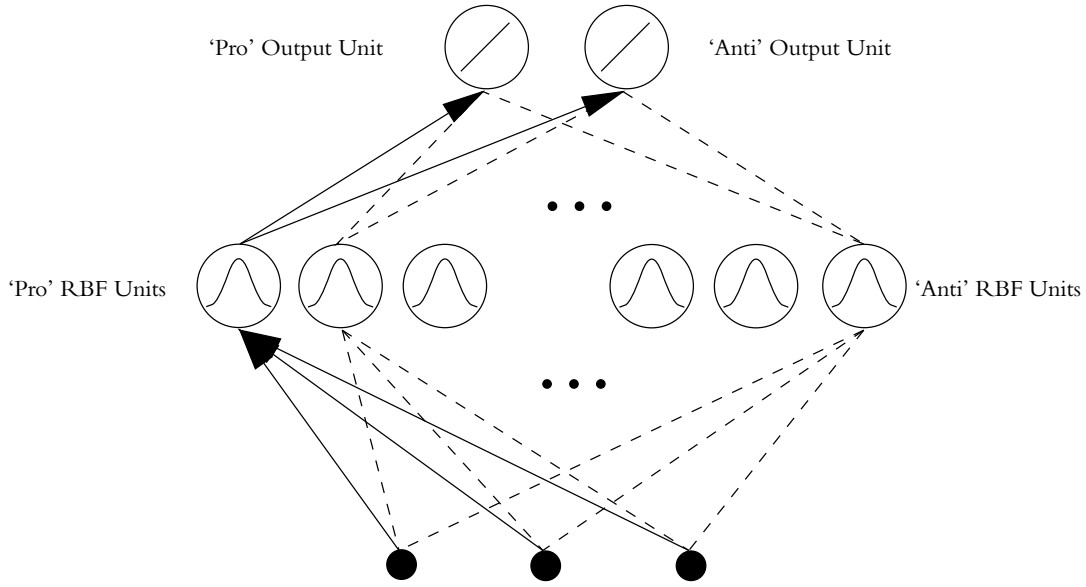


Figure 5.1: General structure for a 'face unit' RBF network. Although there can be a varying number of and ratio between pro and anti hidden units, there are always two output units (for and against the class learnt by the network). All hidden units are fully connected to both output units. This can be compared with the standard RBF network model shown in Figure B.1, Appendix B.

per class) only, the face unit network has two output units, one positive, denoting 'yes' for the current class and, and one negative, ('no') for all other classes. We use the term *pro* to denote hidden units or evidence for the class, and *anti* for that against the class, the negative evidence. For each individual, a face unit RBF network can be trained to discriminate between that person and others selected from the data set, using this pro (supporting) and anti (differentiating) evidence for and against the individual. The ratio between the two can be varied.

Although this approach increases complexity, as more networks need to be trained and and the training data needs to be manipulated differently for each face unit, the splitting of the training for individual classes into separate networks gives a modular structure that can potentially support large numbers of classes, since network size and computational load for weight calculations for the 'standard' RBF model may become impractical as the number of classes increases.

5.1.1 Selection of Negative Evidence

The fundamental process in the face unit network is the splitting of the training data into two halves: class and non-class. The small size of the network is due to the limited amount of non-class data used for training, only those that are seen as hardest to distinguish with the class are included. This selection of negative evidence was based on Euclidean vector distance comparisons of the class face image with images of the same pose angle of non-class faces. In order to make the most efficient arrangement of training examples, the 'anti' evidence was taken from the class that was the closest (in Euclidean distance terms) to the 'pro' class. As the RBF network's hidden units response is based on the same Euclidean distance comparison, it is important to distinguish the closest non-class examples, as these will be the most 'confusable' for the network, and any other other non-class images further away will then be automatically excluded.

5.1.2 Types of Face Unit Networks

To investigate the characteristics of the face unit network model, several different network configurations are devised. To assess how varying the pro/anti balance affected performance, two general types of network layout are used:

Network	Training Examples per Class	
	1	5
Standard RBF network	10	50
Single anti Face Unit Network	2	10
Double anti Face Unit Network	3	15

Table 5.1: Numbers of hidden units used by different RBF networks for same task (when using the Sussex database).

'Single anti' face unit network This uses equal numbers of pro and anti hidden units.

'Double anti' face unit network This uses two anti hidden units for every one pro.

The double anti face unit network is closer than the single anti arrangement to the full standard RBF model, in that it uses more negative than positive evidence. It is included in the tests to show whether this additional information would give the network better discrimination from the negative classes than the single anti arrangement. This characteristic will be more important as the number of classes in the dataset increases, as the number of negative classes will become proportionately greater.

We can compare the relative sizes of the face unit network and the standard RBF network. The standard RBF network uses cn hidden units, where c is the number of identity classes and n is the number of training examples per class. This gives $10n$ hidden units in total when using the Sussex database, as shown in Figure 5.1. The single anti face unit network has only two classes for training (for and against a single person) and a single anti hidden unit for every pro unit, and therefore has $2n$ hidden units in total (however many identity classes there are). The double anti face unit network uses two anti hidden units for every one pro, and therefore has $3n$ hidden units in all. The outcome of this is that as c, the number of identity classes, increases, the face unit network required for a particular task will becomes much smaller relative to the standard RBF network needed for the same task.

Once the number of examples is chosen, we then use two different strategies for the selection of the anti evidence. This gives two further types of network:

'Single best negative' (sbn) face unit networks These use an average of all vector distances between the pro image and all anti images, within each pose angle, averaged over all pose angles to compare whole classes rather than individual images from classes. The lowest overall average value was used to select one anti class, which then represented all negative evidence at all pose angles.

'Multiple best negative' (mbn) face unit networks These use the closest anti image to the pro image for each pose angle, so that several anti classes may be used for a face unit network with more than one training example.

It was anticipated that *sbn* face unit networks would be superior to *mbn* face unit networks, as a more coherent 3-D class boundary would be given by a single negative person-class for all pose angles. On the other hand, the *mbn* approach may utilise local class differences to learn a more efficient solution.

5.1.3 Face Unit Network Terminology

As the face unit networks are arranged differently to the standard RBF networks, they are labelled slightly differently. The face unit network size is denoted here by '$p+a$', where p is the number of pro hidden units, and a is the number of anti hidden units. Tests were made on a range of

Figure 5.2: Example of the range of negative classes that can be selected during the training of a 5+10 double anti, multiple best negative (*mbn*) face unit RBF network.
The top line shows the supporting, 'pro' evidence, the middle and bottom lines the differentiating, 'anti' evidence (middle line is the closest to the pro class, bottom line the second closest).

network sizes from 1+1 to 6+12 on the standard 100-image Sussex database (if these networks had been labelled in the standard 'train/test' form, this would correspond to a range between 2/98 and 18/82 networks). To give an optimal spread of the image data for training, fixed selections of pose angle were used for each size of network, as used in Chapter 4 (see Table 4.2). For instance, the 5+5 and 5+10 networks used poses 10°, 30°, 50°, 70° and 90°, where the pose range was 0° (frontal)–90° (profile).

Figure 5.2 shows how the images used for training were selected for a 5+10 *mbn* face unit network in the experiment. This illustrates not only how several anti classes are used in the *mbn* scheme, but also how they are ranked for the double anti arrangement.

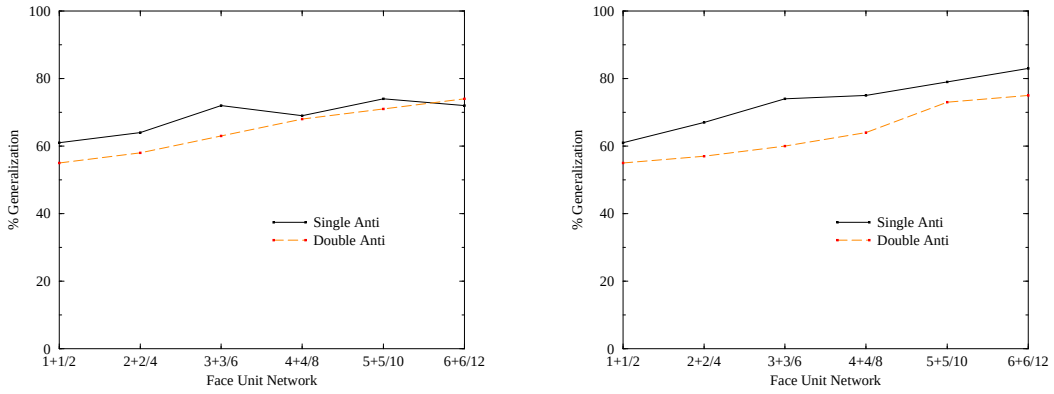
5.1.4 Results

As in the previous chapter, for clarity, our tests use two standard preprocessing methods only: the single-scale DoG and the Gabor A3 with four scales and three orientations (details in Section 3.3 and Appendix C).

Figure 5.3 summarises the overall results for the various types of face unit networks, with different pro/anti ratios and different strategies for selection of anti images. To simplify the information, these graphs do not show the rates after discard, but these gave a consistent improvement of about 7–15% over rates before discard for all networks.

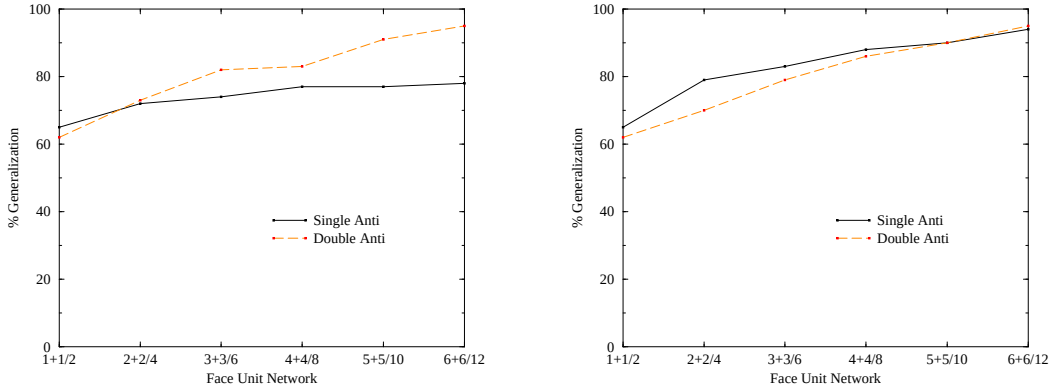
The face unit networks are essentially working in a two-class classification problem, so a random level of generalisation would be 50%. Interestingly, the double anti network arrangement did not appear to give radically better performance than the single anti, except for the 5- and 6-example networks using Gabor preprocessed data. This indicates that the selection of appropriate anti images is efficient enough by itself to create a division in image space between the class and all others without requiring additional negative examples.

Table 5.2 shows specific generalisation rates for the 5-example (5+5 and 5+10) face unit networks before and after discard. It can be seen here that the Gabor preprocessed data allowed the



(i) (ii)

(a) DoG preprocessing



(i) (ii)

(b) Gabor preprocessing

Figure 5.3: Comparing single and double anti training for face unit networks, with average generalisation for all face units shown, but no discard results: (i) Single best negative (*sbn*) networks (ii) Multiple best negative (*mbn*) networks.

Pre-processing	Network	Initial %	% Discarded	% After Discard
DoG	5+5	74	75	87
	5+10	71	49	81
Gabor	5+5	77	50	84
	5+10	91	39	98

(a) Single best negative (*sbn*) networks

Pre-processing	Network	Initial %	% Discarded	% After Discard
DoG	5+5	79	73	97
	5+10	73	42	75
Gabor	5+5	90	47	97
	5+10	90	40	99

(b) Multiple best negative (*mbn*) networks

Table 5.2: Test generalisation for 5-example face unit networks (5+5 and 5+10) using the Sussex database.

Variation	Pre-processing	Network	Initial %	% Discarded	% After Discard
Shift	DoG	25+25	70	67	82
		25+50	69	46	69
	Gabor	25+25	83	35	92
		25+50	86	27	92
Scale	DoG	25+25	78	50	89
		25+50	79	40	84
	Gabor	25+25	88	29	94
		25+50	90	23	96

Table 5.3: Generalisation for 5-pose-example multiple best negative (*mbn*) face unit networks (25+25 and 25+50) with shift and scale varying data.

RBF network to perform more efficiently than the DoG preprocessed data, both in lower discard rates and generalisation before and after discard.

Summary
The *mbn* strategy for selecting anti evidence seemed slightly better than the *sbn*, indicating that dealing with local (at a pose level) confusions was more efficient that trying to identify one global class with which the main class should be contrasted.

5.1.5 Shift and Scale-Varying Data

To assess learnt invariance to the shift and scale-varying Sussex data, introduced in Section 4.3, tests were made using 5 pose example face unit networks. Single and double anti networks were tested to check which reacted best to the more demanding datasets.

Table 5.3 shows that the networks were able to learn shift and scale invariance very similarly to the standard RBF network, in that the scale-varying data was learnt more easily than the shift-varying data, and the Gabor preprocessing allowed both higher generalisation and lower discard rates than the DoG preprocessing. The double anti networks did not give higher generalisation overall, but did give lower discard rates on all tests.

5.1.6 Discussion

From the results, the most useful configuration of face unit RBF network should have:

- more than one training example for both pro and anti data. This is a similar conclusion to that arising from the interpolation test in Chapter 4.
- use equal numbers for pro and anti, although exceptions for particular conditions can be seen.
- use the multiple best negative (*mbn*) strategy to identity the most useful anti evidence to match each pro example on a pose-by-pose basis.

This section has shown that the face unit network can operate to a high level of performance when used in isolation. The next section will show how cooperation with a standard RBF network can be accomplished, and assess the usefulness of such combined information.

5.2 Face Unit Networks as Adjudicators

One potential drawback to using face unit networks is that the processing required to input the test image to every network may become excessive for large number of classes. It would be possible

Multi-Class RBF Network Confidence	Face Unit Network		Confidence to Accept
	Output	Confidence	
High	Yes	High	1
		Low	2
	No	High	6
		Low	5
Low	Yes	High	3
		Low	4
	No	High	8
		Low	7

Table 5.4: Possible outcomes given a particular classification from a standard RBF network when used to index into one specific face unit network, based on the outputs of the two networks. These can be combined to give levels of cooperative confidence in accepting the initial classification, ranging from 1 (the highest) to 8 (lowest).

to take advantage of the specialised training characterised by each individual face unit network by using them in cooperation with other networks.

For instance, a single face unit network could be used to confirm or dispute a classification from a standard RBF network trained on all individuals. The initial output from the multi-class network would be used to index into the group of face unit networks to identify which one was needed, and the outputs from the two networks could then be used in conjunction. It is anticipated that this will give a more reliable result.

5.2.1 Confidence Measures

The standard confidence measure, which has been used in all tests so far for both face unit and standard RBF networks, is based on the difference between the highest and second highest output values. Classifications with a large difference (generally a ratio of 1.8:1 or above) are labelled as high confidence, all the rest as low confidence. See Section 3.2.3 and Figure 3.2 for more details.

The outputs of face unit networks and standard RBF networks can be combined, using this standard confidence measure for both networks. Several levels of classification confidence are then possible, shown in Table 5.4. These range from 1, the highest, where both networks have high confidence, to 8, the lowest, where the standard RBF network has low confidence and the face unit network has high confidence *against* the classification.

Confidence Rating Thresholds

A threshold based on these cooperative ratings can be used to control which classifications are thought of as high confidence. All classifications rated above the threshold are accepted, all below are discarded.

If the multi-class and face unit network concur, even if both are low confidence, then we might say that it is reasonable evidence for a correct classification. According to how confident we want our networks to be in this agreement, we can set a threshold on the confidence ratings between 1 and 4.

It is harder to decide on heuristics for where the two networks disagree. Thresholds set at levels 5 and 6 might still leave useful classifications undiscarded. Some conflicts could be decided on the basis of accepting the decision of which ever network had the higher confidence.

5.2.2 Results

The face unit and standard RBF networks were tested together, the face unit chosen to test each image according to the output of the standard RBF network, and the results were arranged accord-

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		78	52	100	24/24
Cooperative Threshold	1	78	66	100	17/17
	2	78	52	100	24/24
	3	78	46	93	25/27
	4	78	6	79	37/47
	5	78	6	79	37/47
	6	78	6	79	37/47
	7	78	0	78	39/50
	8	78	0	78	39/50

(a) DoG preprocessing

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		96	20	98	39/40
Cooperative Threshold	1	96	38	97	30/31
	2	96	20	98	39/40
	3	96	20	98	39/40
	4	96	2	96	47/49
	5	96	2	96	47/49
	6	96	2	96	47/49
	7	96	0	96	48/50
	8	96	0	96	48/50

(b) Gabor preprocessing

Table 5.5: Generalisation and discard rates for different discard measures: 'Standard RBF Network Only' is the result using a simple discard measure applied to the output of a standard 50/50 multi-class RBF network by itself, the 'Cooperative Threshold' is a threshold value applied to the confidence rating arising from cooperating 50/50 multi-class standard RBF networks and 5+5 single anti multiple best negative (*mbn*) face unit RBF networks.

ing to the cooperating confidence rating thresholds from 1 to 8. Tests were made with both single and double anti networks.

Tables 5.5 and 5.6 show that the cooperating networks are able to give a much finer gradation of confidence levels than the normal confidence measure based on the standard RBF network only.

Although the single anti face unit networks discarded less when in combination with the standard RBF networks than with the double anti networks, their performance was worse on the whole. The double anti networks gave more useful results, giving a good increase in performance compared to no discard at all, and generally equivalent generalisation performance to the conventional, one network discard, with lower discard rates.

Summary

Threshold levels of 1 and 6 on the cooperative confidence rating scale were found to be useful in practice. The highest confidence rating threshold discard level, **1** requires both standard and face unit network to have high confidence for the same class. This threshold value can be used to give better, or at least as good, generalisation performance, after discard, as that provided by the original confidence measure used with the standard RBF network alone. This superior performance is at

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		78	52	100	24/24
Cooperative Threshold	1	78	86	100	7/7
	2	78	58	100	21/21
	3	78	58	100	21/21
	4	78	54	91	21/23
	5	78	48	92	24/26
	6	78	48	92	24/26
	7	78	6	81	38/47
	8	78	0	78	39/50

(a) DoG preprocessing

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		96	20	98	39/40
Cooperative Threshold	1	96	64	100	18/18
	2	96	20	98	39/40
	3	96	20	98	39/40
	4	96	14	98	42/43
	5	96	14	98	42/43
	6	96	14	98	42/43
	7	96	6	96	45/47
	8	96	0	96	48/50

(b) Gabor preprocessing

Table 5.6: As Table 5.6, except the 'Cooperative Threshold' measure uses a 5+10 double anti multiple best negative (*mbn*) face unit RBF network.

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		85	35	98	159/163
Cooperative Threshold	1	85	49	98	126/128
	6	85	13	91	197/217

(a) Shift-varying data

Discard Measure		Initial %	% Discarded	% after Discard	Ratio after Discard
Standard RBF Network Only		90	26	97	178/184
Cooperative Threshold	1	90	42	98	141/144
	6	90	10	94	212/226

(b) Scale-varying data

Table 5.7: Generalisation and discard rates for different discard measures with shift and scale varying data: 'Standard RBF Network Only' is the result using a simple discard measure applied to the output of a standard 50/50 RBF network by itself, the 'Cooperative Threshold' is a threshold value applied to the confidence rating arising from cooperating 250/250 multi-class standard RBF networks and 25+50 double anti multiple best negative (*mbn*) face unit RBF networks, using Gabor preprocessing.

the cost of higher proportion of discarded classifications.

The confidence rating threshold level **6**, which ignores low-confidence output from the face unit network, gives roughly equivalent generalisation to that given using the normal confidence measure with output from the standard network only, but with much lower discard rates. This could well be the most useful configuration for general use.

Stages 4 and 8 on the confidence rating threshold scale were not found to be useful, but are worth mentioning to clarify the coordinating confidences threshold process. The confidence rating threshold level **4** is equivalent to not using the normal confidence measure for either network, relying on the values of the two 'raw' network classifications. This does not appear to be a useful arrangement, giving no advantage over the use of the standard confidence measure with standard RBF network alone.

A confidence rating threshold of **8** is not useful, as it allows no discard at all. This is because all face unit network output and the confidence rating of the standard RBF network are both effectively ignored. This is demonstrated by the 0% discard levels shown for the threshold set to 8 in Tables 5.5 and 5.6.

5.2.3 Shift and Scale-Varying Data

As in Section 5.1.5, tests were made to assess learnt invariance to the shift and scale-varying Sussex data, this time using 5-pose-example face unit networks in cooperation with standard RBF networks, as in the previous section.

Table 5.7 shows a similar gradation of performance, controlled by confidence rating threshold, to that found in the previous section.

5.2.4 Discussion

The cooperative use of the face unit network with the standard multi-class RBF network shown in this section can be seen as a more subtle approach to assessing classification confidence than the

simple, one-network threshold used previously.

Different rating threshold levels can be used with the cooperative scheme to give either *high confidence with high discard* (using a rating threshold of 1), or *moderate confidence with low discard* (rating threshold 6), compared to the *moderate confidence with moderate discard* provided by the original confidence measure using the standard RBF network alone. Intermediate threshold levels (from 2 to 7) provide other combinations of confidence and discard ratios. This ability to vary the system behaviour via the threshold level would be useful for real-life applications, as it allows the user to engineer the solution required.

5.3 Updating Face Units

This section is about how face unit networks can be retrained during use. Face unit networks allow a flexible approach to learning in dynamic environments compared to other neural networks models which have to be completely retrained if the training data is altered in any way.

The face unit network only uses a few of the total number of classes in a problem to train, so operations on any of the other classes not used for training will leave it unaffected. As the number of classes increases, the chance of each face unit network needing retraining due to an operation on another class will become less.

5.3.1 Adding Face Units

To add a new person-class, vector differences need to be compared for all training images, just as for the initial training. Distance calculations for all classes each time a change is made can be avoided, however, by saving the Euclidean distance information, so that only the values for the new class need to be calculated.

Any face unit where an image from the new class is closer than its existing anti evidence would need to be re-trained. All other face units would not require further training. In the worse case, this would mean the entire system of face unit networks being re-trained, but this is less likely as the number of classes increases.

5.3.2 Removing Face Units

Removal of a particular face unit is simpler, as it just requires a check for other face units currently using that face class as anti evidence. Only those that did use the removed face unit would require retraining.

To update an old face unit would require two steps, as it would first need to be removed and then the new data added.

5.3.3 Discussion

This section has tried to address the issue of long-term use of face recognition systems. Task Requirements 3(d)iii and 3(d)iv specify a tolerance to middle-term (makeup, facial hair, etc) and long-term (ageing, etc) changes in appearance. This implies that potential systems will need to be flexible enough to update their training data for such changes.

Although our standard RBF network is fairly fast in retraining completely, this might not be the case if many more examples for each class are used (representing x- and z-axis movement, for example).

It is an open issue how an automatic system would determine that a known individual required retraining due to change in their appearance - could the system itself monitor how confident it was recognising that person and retrain when this fell below some limit, or would it require manual intervention from the user to initiate the process? A system that could be aware of these changes automatically would be more useful than one which simply failed to recognise a previously known person.

5.4 General Discussion

This chapter has presented experimental work using a novel variant of the RBF network model, the face unit network, which learns to distinguish a single class from a range of other classes. This can be used either in groups, one for each class, or singly in conjunction with a multi-class network to give greater reliability to classification.

The most useful configuration of face unit RBF network overall seems to be the single anti multiple best negative (*mbn*) face unit network, which selects the most useful anti evidence to match each pro example on a pose-by-pose basis.

The standard RBF network will give similar positive and negative information about classes, because of the fully interconnected hidden to output unit layer, but the face unit network, by concentrating only on distinguishing one class at a time, allows the negative influences of such non-class connections to be more specialised, indeed optimised, to give the most effective 'one class against all others' partitioning in image space.

The modular approach presented in this chapter using face unit RBF networks to learn identity is especially attractive for the unconstrained recognition task, as it allows the modification of the learned element of the system during use, and can give a secondary classification decision which can either confirm or dispute the primary RBF network output.

Although the face unit network allows finer control in the recognition process with the standard RBF network than can be provided by the latter alone, it is not used for the next two chapters, which deal with image sequences and the recognition of temporal patterns. This is because it is felt that the results will be more understandable if the common baseline of system configuration from chapters 3 and 4, comprising of the standard RBF network with simple discard measure, is maintained for this later experimental work.

Chapters 3 and 4 and this chapter have explored the behaviour of the RBF network in the narrow context of training with the Sussex database. The next chapter applies a more realistic test to the network, using image sequences from a less tightly constrained environment.

Chapter 6

Face Recognition using Image Sequences

This chapter presents experiments using the Radial Basis Function (RBF) network to tackle a more unconstrained face recognition problem using low resolution video information. The work on static images presented in previous chapters is extended here to the time varying case, where an individual is to be recognised in an image sequence. We first consider training and testing on images from a seated subject where scale and shift are quite constrained. We then go on to consider the case where the subject is free to walk about and is tracked (imperfectly). Training and testing in this unconstrained situation is much more problematic, as the data contains a much greater degree of variability in scale, shift, pose, and expression. As in previous chapters, the simple confidence measure, based on relative output magnitudes, is used to discard low-confidence classifications. Because each frame in an image sequence is related to preceding and succeeding frames, the final section of the chapter considers schemes to enforce temporal consistency through the use of time 'windows'.

Initial research often requires restrictions on the variability of test data in order that fundamental principles can be investigated in isolation. However, this means that such applications are far removed from real-world environments, where data is noisy and unpredictable. Besides the theoretical desire to remove such constraints, there is also a real commercial demand for a system that can rapidly identify a person from a small group of users.

Face recognition is a computationally expensive process and to obtain real-time performance requires certain trade-offs. A police record application would require access to enormously large amounts of data but accuracy would have priority over speed, as instantaneous recognition would not be the primary factor. To cope with hundreds of thousands of individuals, views may be limited to face-on or profile only with the face at a specific region of the image, allowing precise pin-pointing and measurement of feature points. In our application, however, we are considering a less constrained environment, where people can move around freely, and so we need to recognise the person from the full range of views where the face is visible.

In addition, the example police application would require extremely low error rate, and only very few (maybe even just one) image of each individual would be available. We have opted for a lower accuracy method that is considerably faster, and provides a reasonable discarding of low-confidence output. Looking at an image sequence, such as in Figure A.14 (Appendix A), it can be seen that there is an abundance of data, and if the current image is ambiguous, it can be discarded and the next considered. The temporal coherence of human faces allows the matching of series of frames linked by movement information with the use of 'time windows' to combine information from several frames.

Previous chapters have shown that the RBF network can learn to be invariant to certain types of variations that can be expected in real-life face images. The experiments presented in this chapter use these abilities in a less constrained environment using image sequences. As mentioned before,

we are computationally constrained to the inter-frame period (of the order of tens of milliseconds) determined by the frame grabber and the localisation software. Offsetting this limitation is the vast quantity of data from image sequences, which suits any technique that can discard low-confidence output to leave a high ratio of correct classifications. In the context of videos of people moving around a room, where large numbers of images of each person in the environment will be produced and changes in the identities present will not be abrupt (from one frame to the next), even quite high discard ratios of 80-90% may be acceptable if the remaining output is of sufficiently high quality.

6.1 Specification for Image Sequences

The image sequences used in the tests reported here are the result of collaboration with Stephen McKenna and Shaogang Gong at Queen Mary and Westfield College (QMW), London, who are researching real-time face detection and tracking (McKenna et al., 1996). This work is still at a preliminary stage, and many issues are still unresolved, such as the nature of appropriate training data: how constrained does it need to be, and how automatic its original collection from the data source should be in a useful real-world application.

The experiments presented here address the situation where the training data is more constrained than the test data, as this is assumed to be the most efficient method of learning identity-specific information. It might be argued that training data should not be constrained, but contain any anticipated test variation, in the light of results from Chapter 4 where it was shown that most variations need to be trained for explicitly. However, it is not obvious how examples of the whole range of variations can be automatically collected from data where free movement is allowed and, therefore, it is not guaranteed that all possible variations will be seen during any particular period.

The face images in the sequences used here are different from those in the Sussex database used in previous chapters, in that they are centred on the head (head-centered), rather than nose-centered (compare Figures A.5 and A.7 in Appendix A). This is because the face pose would need to be known before segmentation if the nose is to remain centered for all face views. Currently, the QMW head tracker does not extract such information (though such an ability is a priority for the future). Tests in Sections 3.1.3 and 3.2.3 showed that the choice of centering algorithm has a marked effect on training and generalization.

We have two types of sequences to simulate a typical unconstrained environment, termed 'Primary' and 'Secondary', see Section A.3 in Appendix A for specific details. The intention is to train the network with a controlled set of data – the Primary image sequences – known to include the types of variability which we want our trained system to generalize over (including 180° range of pose angles, taken against a blank background), and to test on totally unrelated data – the Secondary image sequences. As mentioned in Section 2.3.1, such total separation of training and test data insures against the system using spurious environmental details, such as lighting or background features, to classify individuals. Such problems can always appear in databases where test and training data are collected at the same time and each taken in the same manner. The lack of differentiation between training and test data will also appear in any approach that selects both arbitrarily from a central database. As these studies are still at an initial stage, however, we will start by training and testing just with the Primary sequences in order to get a clearer idea of how the selection of training images from sequences affects subsequent test generalisation.

6.1.1 Preprocessing of Segmented Data

Two main techniques are again used for the preprocessing of the images: Difference of Gaussian (DoG) filtering and Gabor wavelet analysis at a range of scales. In contrast to methods using warping based on registration of features, such as Craw et al. (1995), our approach uses simpler preprocessing, but learns to discriminate using the RBF networks to overcome self-occlusion arising out of head rotation (Task Requirement 3(d)ii).

Training Selection Interval	Train/Test Frames	Initial %	% Discarded	% After Discard
2	278/276	96	12	98
5	114/440	88	30	99
10	60/494	75	50	90
20	33/521	58	68	90
30	24/530	48	81	93
50	16/538	40	81	86

(a) DoG preprocessing

Training Selection Interval	Train/Test Frames	Initial %	% Discarded	% After Discard
2	278/276	99	2	100
5	114/440	98	7	100
10	60/494	95	16	98
20	33/521	87	35	94
30	24/530	73	55	94
50	16/538	67	62	94

(b) Gabor preprocessing

Table 6.1: Effect of preprocessing methods on test generalisation before and after discard of low-confidence output for a standard RBF Network trained with images taken at differing selection intervals from eight Primary image sequences, and tested on those frames from the Primary sequences not used for training.

The experiments presented here again concentrate on two specific applications of these techniques:

DoG convolution with a scale factor of 0.4, with a reduced range of grey-levels, with thresholding to give *zero-crossings* information. A 25×25 image gave 441 samples per image.

Gabor 'A3' sampling with a full range of grey-levels. Four non-overlapping scales were used with three orientations including sine and cosine components. A 25×25 image gave 510 coefficients per image.

6.2 Single Frame Tests

This section presents experiments that treat all the frames in the image sequences as separate points in time, unrelated to frames before and after. This should provide a comparison with the earlier tests on the Sussex database, where the static images were much more constrained in pose and lighting.

6.2.1 Tests with Primary Sequences

To test the ability of the RBF network to classify test images after training with the Primary sequences, experiments were done initially by dividing the Primary sequences into training and test groups (see Table 6.1). To allow the training images to be automatically selected, images at set intervals were extracted from the Primary sequences, all the others were used for testing.

Figures A.14 and A.15 in Appendix A show that for the case of our Primary sequences, this arrangement gives reasonable spread of poses, although the first frame is sometimes not correctly localized. The ratio of frames in each of these two groups is shown in the Train/Test column in the tables.

Training Selection Interval	Train/Test Frames	Initial %	% Discarded	% After Discard
2	278/169	43	69	42
5	114/169	32	76	19
10	60/169	44	75	35
20	33/169	23	76	21

(a) DoG preprocessing

Training Selection Interval	Train/Test Frames	Initial %	% Discarded	% After Discard
2	278/169	61	41	77
5	114/169	56	45	77
10	60/169	60	43	81
20	33/169	54	42	66

(b) Gabor preprocessing

Table 6.2: Effect of preprocessing methods on test generalisation before and after discard of low-confidence output for a standard RBF Network trained with images taken at differing selection intervals from eight Primary image sequences, and tested with a separate Secondary sequence.

The RBF Network trained on eight Primary image sequences was able to generalise very well when tested with the remaining images from the Primary sequences not used for training (Table 6.1). Although the initial results tailed off as the sampling interval increases, the confidence discard allowed a high standard of performance to be maintained, even at the sparser sampling intervals. The network could still recognise 94% of the images with a pose angle range of 180°, having been trained with only four of each class (at a sampling interval of 20). From Figure A.14(a) in Appendix A, it can be seen that this corresponded to the first, third, fifth and seventh images with white boxes, and as noted before, the first one was incorrectly localized. The network was performing well above random performance, and yet each RBF hidden unit was then generalizing over a range of at least $\pm 23°$, given that some examples were at the extreme end of the pose range.

It is significant, in terms of the task requirements detailed in Chapter 2, that the discard measure gave such high levels of generalization after discard, even if the actual proportion discarded were quite high. This is less important than where static images are used, as data from image sequences is cheap and plentiful.

6.2.2 Tests with Secondary Sequences

The Secondary sequences are still in development, but encouraging results have been collected from preliminary experiments, see Table 6.2, where the RBF network is trained with the Primary sequences and tested with a provisional Secondary sequence. What was immediately apparent was that the Gabor preprocessing was essential for this inherently more variable data in the Secondary sequence.

6.2.3 Discussion

The blind selection of training examples at fixed intervals was used, because no prior knowledge about pose or localization accuracy was available. Figures A.14 and A.15 in Appendix A show that a reasonable spread of poses is provided because of the constrained nature of the Primary sequences. Obviously, labelled data would allow a more systematic structuring of the training group required for each individual: how many images in total are needed, what pose angles and whether expression and/or lighting are also represented. However, the labelling of frames requires analysis modules,

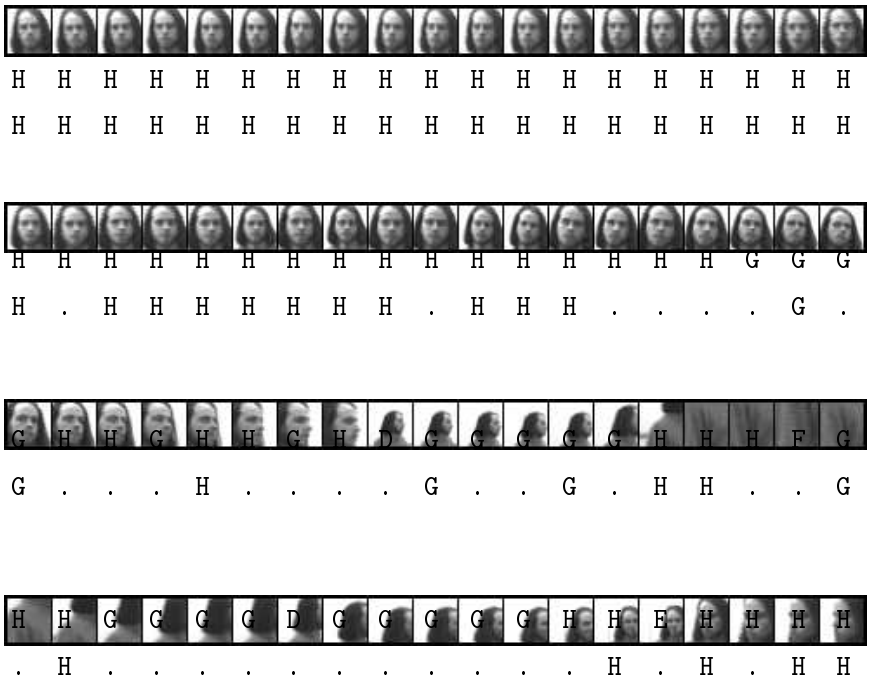


Figure 6.1: Output values for a typical section of the Gabor preprocessed Secondary sequence containing class *steve* (H). Top row of letters shows initial output, lower row output after discard of low-confidence values ('.' indicating where a value has been discarded).

such as pose estimation, at an earlier stage in the face data extraction process.

These results can be used to estimate the likely number of training images needed for a more general system under similar conditions, bearing in mind that no attempt has been made here to address other variations, such as x-axis movement or lighting and expression changes, and it is expected that extra training examples would be required to accommodate such image variations. With the maximum number of classes at around 50, the use of around 20 training examples per class would give a network of about 1000 hidden units with 1000×50 trainable parameters, which would not be prohibitively slow to train (no longer than a minute on current systems).

6.3 Temporal Integration

The next stage of development in our use of image sequences was with a 'time window' integration level using the raw output. When using image sequences from real life, where individuals will be present for significant periods, it is appropriate to use techniques which take advantage of temporal coherence to improve performance. The idea here is that periods of low confidence output can be "patched" into a coherent stream by some kind of moving average rather than a full belief-based mechanism (for example, Buxton and Gong (1995a)).

An on-line source of data cannot be evaluated like a conventional database, as not all of the data is yet present, nor is it known who will be present in the environment or for how long. If an assumption of *temporal coherence* is made, in other words that one person will not transform into another instantly, high-confidence information in a 'time window' can be utilised in periods of low-confidence output to lend support to the current output (assuming it is the same classification, though conflicting output could also be useful). In this environment, a running total of output values for a time window can be kept, and a expression for the individual currently in view given. For this to work, fairly low (above random) success rates will suffice.

To illustrate this technique, consider Figure 6.1, which shows a section of the test Secondary

Time Window	Initial %	% After Discard
1	66	86
3	72	86
5	68	89
10	68	100

Table 6.3: Results for the TDRBF network using temporal integration through varying time windows in output for the sequence shown in Figure 6.1.

sequence. Table 6.3 shows how use of time windows to re-assess the output value can affect performance. Periods of correct output followed by random values can be interpreted as all correct, using the last stable output as a type of memory.

6.3.1 Discussion

Although the testing here was by no means exhaustive, it can be seen that such techniques which take advantage of temporal coherence can be used to improve performance.

6.4 General Discussion

It must be emphasised that this research is at a preliminary stage but that the technique shows promise for scaling up to large databases. Although only a few individuals are shown in our image sequences, this type of network has been shown to work well with larger numbers of classes. For example, the Olivetti Research Laboratory database of faces (see Section 2.6 for details), with 400 images of 40 people can be distinguished with a high level of performance.

Several points can seen from the results:

1. The RBF network is shown to generalise well from samples in classifying faces in real-time sequences.

2. Gabor preprocessing is shown to give a more generally useful input representation than the DoG preprocessing, especially for the more difficult Secondary sequence.

3. The confidence measure used in discarding uncertain classifications is shown to be important for handling sequences especially where a small training set is used.

If it can be assumed that only one person is present in an image sequence throughout the period of tracking, a simple total for each class can be used to determine identity. This cumulative evidence can be gathered during the duration of the sequence. The results shown in Table 6.2b would certainly be sufficient for such a scheme, as the correct class was identified in more than 50% of the frames.

In conclusion, the locally-tuned RBF networks showed excellent performance in the simpler face recognition task when trained and tested on images from Primary sequences. This is a promising result for the RBF techniques, considering the high degree of variability due to the varying views (mostly rotations) of a person's face in these data sets. The results so far from the Secondary sequences also show considerable promise, especially with the additional use of temporal coherence to improve performance. In these sequences, the face detection scheme (McKenna et al., 1996) currently selects and re-scales faces in near face-on views but does not discard the others. Gong and McKenna are currently working on closer segmentation over view and the provision of face detection confidence for each frame, which should be able, for instance, to remove many of the more anomalous frames in our current Secondary sequence. It is expected that further development of this scheme will allow improvements in the reliable and consistent labelling of faces in

unconstrained image sequences. It is clear that the ability of the RBF networks to give a measure of confidence, which allows temporal integration over image frames where the visual evidence is poor, is essential for this development.

Work is progressing together with colleagues at QMW in refining the face detection scheme and automated on-line learning of new classes of individual. The next stage of development will integrate this refined on-line face detection and localisation with the trained RBF networks to cope with real-time image sequences including the usual variations in illumination as well as position, scale, view and facial expression. It is clear from the work of Bishop (1995) and others that using statistically based techniques is the key to good performance. The RBF techniques are mathematically well-founded, which gives a clear advantage in engineering a solution to our application problems.

Chapter 7

Recognition of Simple Behaviours using Time-Delay RBF Networks

This chapter presents experiments using a Radial Basis Function (RBF) variant of the Time-Delay neural network (TDNN) with image sequences of human faces. The main purpose of this is to show that the network is able to learn simple behaviours based on *y*-axis head rotation and generalise on different data.

The recognition of simple behaviours is an important capability for many computer vision applications, such as visual surveillance (Buxton & Gong, 1995a) or biomedical sequence understanding (Psarrou & Buxton, 1993). The behaviour in the experiments presented here was simple head rotation to the left or right or the head held still. However, the work raises important issues for connectionist techniques: 1) time, 2) representation, and 3) learning with generalization. Multi-layer perceptrons with supervised learning are very popular for applications which use static representations, but time is important in many domains such as vision, speech and motor control. Dynamic neural networks can be constructed by adding recurrent connections to form a contextual memory for prediction in time (Jordan, 1989; Elman, 1990; Mozer, 1994). These partially recurrent neural networks can be trained using back-propagation but there may be problems with stability and very long training sequences when using dynamic representations. Instead, we use a simple Time-Delay mechanism in conjunction with an RBF network, which we term a TDRBF network, to allow fast, robust solutions to our problem of recognising head turning behaviour.

In learning to recognise behaviour with a TDRBF network, it is again important to use an input representation (now ordered in time) that allows generalisation over variations in lighting, scale and translation (shift). The results presented in Chapters 3–6 indicate that complex 2-D Gabor filters (Daugman, 1988), which approximate the receptive fields of simple cells in the primary visual cortex, provide just such a representation. In this chapter we show how this work can be adapted from face recognition from a single image frame to the problem of behaviour recognition in extended video sequences. With our approach, images containing pre-segmented faces in a typical motion sequence can be analyzed to obtain the appropriate Gabor representation for each time frame in the motion sequence.

7.1 The Time-Delay RBF Model

The Time-Delay Neural Network (TDNN) model (for an introduction, see Hertz et al. (1991)), incorporates the concept of time-delays in order to process temporal context, and has been successfully applied to speech and handwriting recognition tasks (Waibel et al., 1989). Its structured design allows it to specialise on spatio-temporal tasks, but, as in weight-sharing network, the reduction of trainable parameters can increase generalisation (Le Cun et al., 1989).

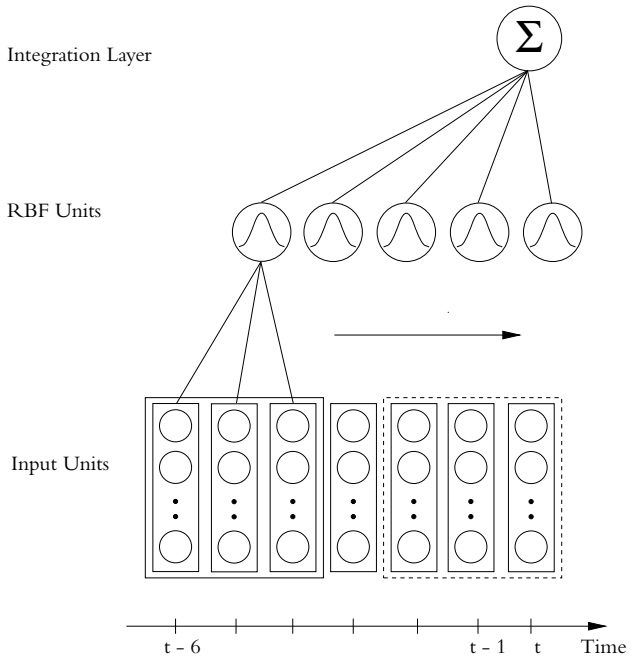


Figure 7.1: Structure of a single class for a TDRBF network with time window of 3 and a integration window of 5 (after Berthold (1994)).

We adapted our RBF network model to use time-delays in order to process temporal context. This Time-Delay version of the RBF network, the TDRBF network, is similar to that used by Berthold (1994) and combines data from a fixed time 'window' into a single vector as input, see Figure 7.1. Berthold, however took a constructive approach, combining the idea of a sliding input window from the standard TDNN network with a training procedure for adding and adjusting RBF units when required. We have used a simpler technique, successful in previous work with RBF networks, which uses a fixed number of units, one for each example, and the pseudo-inverse process to calculate weights, see Appendix B for details.

The integration layer used in Berthold (1994) is not used for the tests presented in the first section, as the images in the training and test sequences are not situated in a real sequence. For these experiments, we treat each time window as separate moments in time because of this synthetic nature of the temporal version of the Sussex database. However, Section 7.3 tackles this issue by using real image sequences, where classification over time is significant, and investigates the use of an integration layer.

7.2 Learning Actions Through Time

Simple experiments were made with the TDRBF network using image sequences to train it to identify types of *y*-axis head rotation. The data used was from the Sussex database of 10 people each in 10 different poses at 10° intervals from face-on to profile, see Appendix A for specific details. The identity contained in each frame was discarded and classes were reassigned to groups of frames in terms of the direction of pose change within the frame sequence, effectively treating the database as 10 image sequences of a person rotating their head from side to side. As each image for a particular person in the Sussex database varies by 10°, the effect of head rotation can be synthesized by concatenating several adjacent frames. The TDRBF network can then be trained to distinguish the presence and direction of movement in these simple fixed sequences.

For all experiments presented here, half of the database was used to train the network, and the other half used to test it, and each sequence was taken from a fixed 'window' from within the ten

frames 0–9 for each person. Two schemes were devised to split the data up:

- the *Alternate Frames* (AF) tests, illustrated in Figure 7.2, used alternate frames from each person, so that training and test data contained all ten people and the window size range was 2–5 frames, and
- the *Alternate Person* (AP) tests, illustrated in Figure 7.3, which used all the frames from 5 people for training, and the other 5 for testing. The window size range for the AP tests was 2-9 frames.

It can be seen that the two types of selection process and the varying window sizes gave a wide range of numbers of sequences that could be used as data. In addition to this, the variety of data will be increased by the type of head rotation.

Head Rotation Classes
Three classes were used for training the TDRBF network, corresponding to three types of rotation present in the image sequence:

LR sequences These simulate a left to right head rotation in a 'window' within the ten frames 0–9 for each person, such as shown in Figure 7.2(a). Sequences were interleaved with each other to use all the frames for each person. For example, if the window size was 4, the sequences used would be, for the AF tests, 0,2,4,6 and 2,4,6,8, and for the AP tests, 0,1,2,3, 1,2,3,4, 2,3,4,5 up to 6,7,8,9.

RL sequences These are identical to LR, except that the rotation is in the opposite direction (from right to left), as shown in Figure 7.2(b), so the frame numbers go from 9 to 0. For example, if the window size was 4, the sequences used would be, for the AF tests, 8,6,4,2 and 6,4,2,0, and for the AP tests, 9,8,7,6, 8,7,6,5, 7,6,5,4 down to 3,2,1,0.

Static sequences These simulate a fixed head position through time, illustrated by Figure 7.2(c). They were created by repeating the middle frame of the time window.

Preprocessing
Gabor wavelet analysis at a range of scales was used for preprocessing of the images. The 'A3' scheme was used, where data was sampled at four non-overlapping scales from 8×8 to 1×1 and three orientations (0°, 120°, 240°) with sine and cosine components, see Section 3.3 and Section C.2 in Appendix C for specific details. The Samples column in the tables show the total number of Gabor coefficients contained in each input vector. A discard measure was used on some of the tests to exclude low-confidence output; the proportion discarded and the subsequent generalization rate are shown for these tests.

7.2.1 Alternate Frame Tests

The Alternate Frame (AF) tests used alternate frames from all ten people for training and testing (see Figure 7.2). Three types of network training were used, using either two or three classes:

Static/LR Here, two classes were trained and tested for: left to right movement and static, illustrated by Figures 7.2(a) and (c). The network was trained with a window from frames 0, 2, 4, 6 and 8 of all ten people, and tested on a window from frames 1, 3, 5, 7 and 9, thus using 20° intervals between frames. Static sequences were simulated by repeating the middle frame of the time window.

Static/RL This was similar to LR, except that the rotation was in the opposite direction, illustrated by Figures 7.2(b) and (c). This meant that the network was trained with frames 8, 6, 4, 2 and 0, and tested on 9, 7, 5, 3 and 1.

(a) (b) (c) (d)

Figure 7.2: Example data sequences for Alternate Frame (AF) tests with a time window of three frames: (a) LR Training - Frames 2, 4 and 6 (b) RL training (frames 6, 4 and 2) (c) Static training (frame 4 repeated) (d) LR Test - Frames 3, 5 and 7 of same person. Data from all 10 people in the Sussex database was used for both training and testing, each taken from alternate frames.

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
5	2550	20/20	100	5	100
4	2040	40/40	95	5	100
3	1530	60/60	100	8	100
2	1020	80/80	90	8	92

(a) 2 Classes, distinguishing LR and static sequences.

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
5	2550	30/30	100	7	100
4	2040	60/60	97	8	100
3	1530	90/90	93	8	100
2	1020	120/120	83	25	96

(b) 3 Classes, distinguishing LR, RL and static sequences.

Table 7.1: Effect of time window size on generalization rates for TDRBF network trained and tested on image sequences from alternate frames (AF testing). The test sequences contain alternate frames from those seen during training.

Static/LR/RL This was a mixture of the above two schemes, as the network was trained and tested with three classes: left to right movement, right to left movement and static, illustrated by Figures 7.2(a), (b) and (c).

Note that the LR sequence vectors are mirror-images of the RL in Euclidean space, so the Euclidean distance of LR sequence 3-5-7 to 2-4-6 is the same as the RL sequence 7-5-3 to 6-4-2. For this reason, the results for the Static/RL tests are identical to those for the Static/LR tests, and therefore not shown here.

Table 7.1 shows that the TDRBF network can learn the different types of head rotation and generalize with sequences from alternate frames to those used for training. The longer time windows gave the best generalization performance and lowest discard proportions.

Summary

In light of the findings of Section 4.2.2, where the level of pose invariance provided by each RBF hidden unit was at least $\pm 15^\circ$, it was not surprising that these AF TDRBF networks could perform well. This was because each test sequence frame had a training sequence frame available that was only 10° away in pose angle.

7.2.2 Alternate Person Tests

For the Alternate Person (AP) tests, the TDRBF network was trained with sequences of images from five people and tested with sequences of images from the other five people, so that generalisation will reflect learning of the temporal task rather than identity. This is a harder test for the network, as it is tested with images of people not seen during training (see Figure 7.3). Three types of AP network training were used, as with the AF tests, using either two or three classes:

Static/LR As for the AF tests, using two classes, but trains with a window from all ten frames from 0 to 9 of five people, and tests on a window from all ten frames from 0 to 9 of the other five, illustrated by Figures 7.3(a) and (c). As all the frames available are used, the sequences use 10° intervals between frames.

Static/RL As for the AF tests, using two classes, but trains and tests with all ten frames from 9 to 0, illustrated by Figures 7.3(b) and (c).

Static/LR/RL As for the AF tests, using three classes, but trains and tests with all ten frames from five people each, illustrated by Figures 7.3(a), (b) and (c).

Table 7.2 shows that the TDRBF network can learn the different types of movement and generalise with sequences from individuals not encountered during training. As in the alternate frame (AF) tests shown earlier, the longer time windows gave the best generalization performance and lowest discard proportions.

Summary

The AP tests were similar to the pose learning tests done in Chapter 3, in that the data from a particular individual is used to train the network about non-identity information. The TDRBF network was able to learn the movement more easily than the RBF network could learn the pose differences. This could be due to different levels of subtlety in the two tasks, as movement over several frames characterized in Euclidean space will be more distinctive than solitary pose examples.

7.2.3 Discussion

The TDRBF network has been shown to be able to learn the simple behaviour recognition task, distinguishing both single- and two-direction movement sequences from static sequences constructed from the Sussex database.

The alternate frame (AF) and alternate people (AP) tests were different in terms of the rotation speed of the head. The AF test sequences, using every other frame, are effectively moving twice

(a) (b) (c) (d)

Figure 7.3: Example data sequences for Alternate Person (AP) tests with a time window of 3 frames: (a) LR training (frames 2, 3 and 4) (b) RL training (frames 4, 3 and 2) (c) Static training (frame 3 repeated) (d) LR test (frames 2, 3 and 4 of different person).
Data from five of the 10 people in the Sussex database was used for training and from the other five for testing.

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
10	5100	10/10	100	0	100
8	4080	30/30	100	3	100
6	3060	50/50	96	10	100
4	2040	70/70	94	13	97
2	1020	90/90	89	13	92

(a) 2 Classes, distinguishing LR and static sequences.

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
10	5100	15/15	100	7	100
8	4080	45/45	100	11	100
6	3060	75/75	93	19	100
4	2040	105/105	90	31	100
2	1020	135/135	67	48	97

(b) 3 Classes, distinguishing LR, RL and static sequences.

Table 7.2: Effect of time window size on generalization rates for TDRBF network trained and tested on image sequences from alternate people (AP testing). The test sequences contain people *not* seen during training.

Figure 7.4: The real image sequence from QMW, used to test TDRBF networks trained on sequences from the Sussex database. Note the wide variation in head position and gaze direction.

Time	Samples Window	Training/ Test	Integration Layer Size (% Correct)				
			1	3	5	7	9
6	3060	75/57	54	53	53	53	54
5	2550	90/58	62	62	67	64	69
4	2040	105/59	64	61	76	83	75
3	1530	120/60	63	60	73	80	78
2	1020	135/61	56	56	52	57	48

Table 7.3: Test generalization with TDRBF networks trained with Static/LR/RL (3 class) sequences from alternate people (AP) in the Sussex database, tested on the QMW real image sequence shown in Figure 7.4, varying both the input time window and the output integration layer size.

as fast as the sequences for the AP tests. In this case, this did not matter, as the networks were not required to learn any time base invariance for these two task.

7.3 Use of Real Image Sequences

To investigate the TDRBF network model further, networks trained with 3 alternate people (AP) classes from the Sussex database were tested on previously unseen image sequences containing a variety of head movement (see Figure 7.4). These image sequences are the result of collaboration with Stephen McKenna and Shaogang Gong at Queen Mary and Westfield College (QMW), London, who are researching real-time face detection and tracking. Chapter 6 has shown how the standard RBF network can work well with this type of data.

Although the amount of test data from the real image sequence remained constant, there were more test sequences as the window size became smaller. A simple majority decision was used to apply rotation classes to windows where more than one type of movement was present, for instance where the head reached full profile and started to move back towards frontal position.

As the network was being tested on a real image sequence, an integration layer was introduced for these tests. The integration layer takes the classifications from a range of time steps and sums output for each class, giving an overal classification on a 'winner-take-all' basis. This is effectively an extra time window over the TDRBF output layer (the input to which was from an original time window).

7.3.1 Results

Table 7.3 shows that the TDRBF network trained with sequences from the Sussex database can generalize to test data taken under very different conditions and containing people not seen in the training data. The integration layer did have an effect on performance. The optimum size for this layer seems to be around 7 time steps (with a input time window of 4 steps), reflecting the slow speed of head rotation present in the data.

7.3.2 Discussion

The issue of the *'time base'* of actions, that is, how fast or slow actions occur, would have to be taken into account in any real-life image sequences, as any movement would occur at a variety of speeds quite naturally. Although Berthold (1994) used the integration layer to cope with shifts of event position in time, the scale of events was not discussed. As mentioned before, the rotation present in the alternate frame (AF) test sequences was twice as fast as that in the alternate people (AP) tests.

Variations in time scale and pauses within the overall gesture can be handled by a recurrent network, or training data which explicitly demonstrated the different classes of motion-based behaviour at a variety of speeds, for example, using time-warping. Taking this into account, the original image sequence was subsampled to match the rotation speed of the original data, which was 10° per time step.

Figure 7.4 highlights the problems encountered when trying to assign fixed behaviours to real-life data. Not only will some time windows contain the end of one and the beginning of other movements, but the acceleration of rotation will be changing. This shows that left to right head movement is much more complex than a simple, fixed speed change in pose.

7.4 General Discussion

Several points can seen from the results:

- The TDRBF network is shown to be able to learn certain simple behaviours based on y-axis head rotation.
- The TDRBF network maintained a high level of performance even on data containing individuals not seen during training (the alternate person (AP) tests).
- An integration layer in a TDRBF network can allow the extraction of behaviour information even with quite markedly different data to that with which the network was trained.

The main points here are 1) the simple, deterministic 'training' of the TDRBF networks means that they are highly suited to on-line learning, 2) the shift invariance and ability to recognise features in time means they are capable of recognising simple behaviours, and 3) high levels of performance on the generalisation to new datasets that behave in similar ways means they are very useful for such practical dynamic vision tasks. The limitations of this technique are 1) the problem of the time-base which was not fully overcome even with the addition of an integration layer, and 2) the problem of defining the simple behaviours. The TDRBF networks are capable of distinguishing a 'quick turn' from a 'slow turn' as well as distinguishing whether the turn was to the right or the left, but it seems that more qualitative definitions of behaviour would best be tackled using more general recurrent networks. This issue is discussed further by Mozer (1994) and by Psarrou and Buxton (1994). In addition, Cleeremans (1989) shows that partially recurrent networks together with a qualitative input representation can be successfully used even for the demanding task of predicting state to state transitions in finite state automata. It is clear, however, that the TDRBF networks are able to perform extremely well where there is a straightforward quantitative relationship between the data and the simple behaviour pattern to be learnt.

The temporal task we used is very simplistic, as each image sequence only contains one standard-length movement. We have tested our trained network on a less constrained sequence, similar to those used in previous section, with some success. It is difficult to assess the performance using such data at present, as it is not yet clear how real-life behaviours will need to be segmented for meaningful analysis. Most human gestures consist of combinations of movements of several part of the body. To make things harder, these are not necessarily in unison, but choreographed with each other at specific times. To signal that a specific gesture has occurred recognition will have to occur over time, linking output from multiple recognisers.

The tests presented in this chapter have shown that the TDRBF network has useful temporal recognition properties, which could be of use in real-life applications, due to its rapid learning and operation, in comparison to more general, but slower, recurrent networks.

Chapter 8

Conclusion

The aim of this thesis has been to explore the practicalities of computer-based face recognition in everyday environments, such as living-rooms or offices. This chapter summarises the main results and contributions from the thesis and outlines directions for future work.

Chapter 2 described the task of face recognition in unconstrained environments in detail and drew up specific requirements to fulfill it. A review of general theories of object recognition and psychological evidence was then followed by a more detailed discussion of current approaches to face recognition, with specific emphasis on three aspects of the recognition process – acquisition, representation and reasoning. This allowed us to establish a suitable approach, using filter-based preprocessing with a radial basis function (RBF) network, to fulfill our task requirements. This scheme was then tested on a standard database in order that it could be compared to other published results using the same data.

Chapter 3 introduced our pose-varying Sussex database, and used it to discuss methods for face representation, normalisation and preprocessing. Resolution and pose variations in face images were studied to analyse how they affect recognition performance in term of Euclidean distance measurements in image comparisons. The contributions of particular elements of the RBF network towards its overall generalisation behaviour are also analysed and compared with related classifiers, in order to assess how the RBF network was able to generalise in the presence of complex 3-D transformations, such as head pose.

Chapter 4 explored the generalisation properties of the RBF network, looking specifically at pose, scale and shift invariance. We were able to establish significant advantages of using Gabor filter preprocessing over the Difference of Gaussians preprocessing, both for generalisation and confidence of classification. The invariance provided by the preprocessing determined the accuracy of face segmentations and amount of training data required for effective learning or recognition. This, together with the results from Chapter 3, laid the experimental foundation for work in later chapters.

Chapter 5 presented a novel variant of the RBF network, the 'Face Unit' network, which learnt to distinguish one particular individual only from the known group. We then were able to demonstrate how it can be used to give an alternative method of learning tasks which can then provide additional evidence for identity.

Chapter 6 explained how the RBF network can be applied to image sequences. The data used here was much taken from a much less 'constrained' environment than the other face recognition databases. We were able to establish the suitability of the filter-based preprocessing and RBF network approach for classifying individuals in uncertain and ambiguous test data, even in the presence of large lighting and pose variation.

Chapter 7 explored the temporal learning abilities of the RBF network. We used a TDRBF

network, previously used for speech recognition, to recognise simple image-based behaviours based on head rotation. We were able to show that such actions can be easily learnt and generalised to even with simple training methods.

8.1 Contributions of the Thesis

There are four main contributions made by the thesis:

1. *Gabor filter representation.* We have developed an efficient sparse-sampled Gabor filter representation suitable for extraction from low resolution face images. We have been able to establish that this can, in contrast to a Difference of Gaussians representation, provide some inherent scale invariance (that is, without the provision of explicit scale varying images) which is not present for other variations, such as translation (lateral shift).

2. *Radial basis function (RBF) network scheme for image sequences.* We have developed a fast RBF network scheme which has been shown to provide robust generalisation when used with pose-varying face data in image sequences.

3. *Face Unit RBF network scheme.* We have developed our own novel Face Unit RBF network model that can be used either alone to classify individuals in a known group, one for each person, or to accompany standard RBF network output for a cooperating classification.

4. *Image-based Time-Delay RBF network scheme.* We have established the suitability of the Time-Delay RBF model, previously only used for speech recognition, for image analysis. We were able to show that the network could recognise simple head-turning behaviours in image sequences in an extension to our previous, static frame training methods.

8.2 Discussion

Section 2.1 devised four main areas of requirements for our target task: Group 1) general requirements that need to be satisfied by all parts of the system, Group 2) acquisition requirements concerned with monitoring and extraction of useful information, Group 3) face recognition requirements for the recognition stage and Group 4) identity requirements which are concerned with how the recognition information is used. As mentioned in that section, those from Group 2 are assumed to have been previously fulfilled via existing technology and those from Group 4 are the subject for future work.

We believe the Group 1 General Requirements, 1a and 1b, are addressed appropriately in our RBF network scheme with filter-based preprocessing. We have been able to show rapid preprocessing, training and classification (in Section 2.6) and robust generalisation of trained RBF networks to test image sequences containing significantly different examples of everyday lighting and pose variation (in Section 6.2.2).

For Group 3, the Face Recognition Requirements, the following requirements have been fulfilled in the thesis (with the sections where this was demonstrated shown in brackets): 3a – Fast learning and real-time recognition of up to 50 individuals (Section 2.6); 3b – ability to work with low-resolution face images (Chapters 3 and 4); 3(c)i – minor translation (shift) and scale invariance (Section 4.3); 3(c)ii – moderate illumination invariance (Sections 3.3 and 6.2.2); 3(c)iv – background invariance (Section 6.2.2); 3(d)ii – head pose invariance, including self-occlusion (Chapter 3, Section 4.2 and Chapter 6); 3e – confidence available for output (Section 3.2 and Chapter 5).

The following Group 3 Requirements were *not* demonstrated in the thesis and would need to be addressed in future work: 3(c)iii – tolerance to occlusion by other objects (although the suitability of similar approaches to tackling both this and self-occlusion (3(d)ii) has been shown by results in Edelman and Poggio (1992) and Ahmad and Tresp (1993)); 3(d)i – moderate expression variation; 3(d)iii – tolerance to everyday changes, such as hair-styles and glasses; 3(d)iv – tolerance

to longer-term changes, such as aging and weight change; 3f – ability to detect (not classify) a person who is outside the known group.

8.2.1 Limitations of Approach

The approach we have taken is extremely robust in terms of handling noisy, low-resolution images. In such conditions, there is always going to be ambiguity and uncertainty, and the RBF network will be able to make the most of this. This is not to say that it is the perfect solution for all types of face recognition, indeed, we cannot make any claims for tasks outside our current requirements. In particular, any neural network scheme may not be suited to classifying particularly large numbers of people or using high-resolution images, because of an exponential computational load related to the dimensionality of the problem. The low resolution of the images may well create problems if used to discriminate large numbers of people, due to a limit being reached on the number of different patterns actually possible with the pixels available.

We do not have a method of detecting people outside our known group at present, which makes it hard to give robust classification when using real-life image sequences. A further complication is that such detection would ideally need to distinguish between genuine unknown people and non-face images, such as shown in the Secondary test sequence used in Section 6.2.2. However, this may be easier in the future, as McKenna and Gong (1997) are now able to provide a level of confidence of face being present in a segmented frame, which would immediately allow non-faces and strangers to be kept separate.

We cannot use the approach as it stands for more general testing of image sequences, as it is not known how to train the network effectively to recognise each individual. The only strategies we can use currently are 1) use all frames from a sequence, 2) use a selection taken at a fixed interval, and 3) take a random selection. The ideal solution would be to provide a range of views of each person, regularly spaced with respect to head pose (y-axis variation), as this can be expected even due to camera angle alone. In addition, we have not addressed the other other axes of head movement, which are also present if the head is nodded or inclined or if the camera is subject to roll or yaw (rather than being passive). Although the y-axis head pose will vary most, as it is linked to body orientation most closely, there will always be some x- and z-axis variation, so explicit examples of this will be required for efficient recognition.

8.3 Future Work

We have left the Group 2 Acquisition Requirements to be fulfilled by others, particularly the QMW Vision Group (McKenna & Gong, 1996, 1997), who have been making rapid progress in real-time face tracking and localisation. Significant recent work has allowed more robust segmentation through colour detection, and confidence levels can now be provided with each segmented frame to denote the likelihood of a face being present in the image. Research to incorporate such advances in our own work will obviously lead to improvements in our system, especially as our results in Section 6.2.2 included classifications for spurious frames which contained no face and could therefore never obtain high performance. In spite of this, if the criterion for successful recognition in such sequences is the correct class being identified for the majority of frames, then the system is already working. However, the serious lack of test data weighs against such optimism to sound a note of caution against thinking that our approach is already useful for real-life applications. Clearly, much work remains to be done.

A particular problem for the research is the lack of suitable databases of test image sequences. Standard databases, such as M2VTS and FERET, are designed for different purposes. There is an urgent need for a structured, freely available database with wide ranges of examples of both head and body movement.

We identified above those Group 3 requirements which had not been fulfilled within the thesis. Attention would need to be focussed on these areas before any useful application could be produced.

Specifically, in terms of invariance, more precise limits need to be established for the system for translation (shift) and scale variation, and research needs to be done to establish limits on expression and lighting variation.

As mentioned in Section 2.1, the Group 4 requirements were were not included in the framework for the thesis and were to be addressed in future work. This group of requirements was concerned with adapting the known group of individuals the system could recognise. This is obviously an important ability for any application that is needed to continue working for more than a day or so. A key requirement for these higher-level processes is Requirement 3f, the ability to detect a person from outside the known group of individuals. Such a capability is therefore the highest priority for future work if any progress is to then be made with the Group 4 requirements. It is not clear what the best approach would be to detecting individuals from outside the group, but techniques may include analysis of network confidence levels or developing a 'group detector' network similar to the Face Unit network, the output of which would be either 'yes, a member of the group' or 'no, a stranger'. Of course, providing negative evidence for such a task might be as difficult as providing 'non-faces' for face detection schemes.

Although we are particularly interested in using image sequences, it has not been possible to do a useful number of tests with sequences up to this point, and a lot of questions relating to their use still remain open, such as how to select canonical views for training. We want an even spread of pose views - should we start off with a few, extracting 'on the fly' and add to these as they become available? The issue of training for dynamic environments is difficult, as it is not clear what representative views of a person really are. It is clear that humans use cues such as clothes and hairstyles for everyday recognition – is this an efficient way for a computational system to train, and if so, how are suitable data determined? In addition, people change from week to week, so it may well become necessary to re-collect training examples to show their new appearance – how would this be decided and should an archive of previous hairstyles, for example, be kept as a 'memory' in case they return to a similar style later?

Our work with temporal behaviour and gestures within image sequences is also at a preliminary stage. We have been able to establish the suitability of the TDRBF network for certain limited actions, but for more complex gestures and to obtain robust performance, we will need to expand the training regimes to encompass different and variable 'time bases' for the same behaviour at different speeds, especially when the data contains lengthy actions or simultaneous combinations of learnt behaviours. Solutions to this problem may require either subdividing the behaviours into fast and slower versions and/or merging these in a second stage of behavioural analysis. Such flexibility may turn out to be an advantage in practice, as the interpretation of a fast pointing action may be different from a slower action. We also wish to develop body movement and gesture models for the user-specific interpretation of behaviour and intention. Together with the QMW group, we hope to develop real-time 'intentional tracking' techniques that exploit such behavioural models by estimating and predicting expected head and body movement and primitive gestures.

This thesis has been able to make good progress towards a solution for our task requirements. To be able to expand face recognition techniques to tackle everyday environments has been a fascinating problem, encompassing a wide range of disciplines. We look forward to taking the approach further to address more general behaviour.

Bibliography

Ahmad, S., & Tresp, V. (1993). Some solutions to the missing feature problem in vision. In Hanson, S. J., Cowan, J. D., & Giles, C. L. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 5, pp. 393–400 San Mateo, CA. Morgan Kaufmann.

Akamatsu, S., Sasaki, T., Fukamachi, H., Masui, N., & Suenaga, Y. (1992). An accurate and robust face identification scheme. In *Proceedings of International Conference on Pattern Recognition*, pp. 217–220 The Hague, Netherlands.

ATR (1996). Cognitive and representative models of visual image information: Interaction of multi-modal information relayed by face.
Web page: `http://www.hip.atr.co.jp/departments/Dept2/Reports_FP.html`.

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity?. *European Journal of Cognitive Psychology*, *3*, 87–103.

Ballard, D. H., & Rao, R. P. N. (1994). Seeing behind occlusions. In Eklundh, J. O. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 800, pp. 274–285 Stockholm, Sweden. Springer-Verlag.

Baron, R. J. (1981). Mechanisms of human facial recognition. *International Journal of Man-Machine Studies*, *15*, 137–178.

Bartlett, M. S., & Sejnowski, T. J. (1996). Unsupervised learning of invariant representations of faces through temporal association. In Bower, J. M. (Ed.), *Computational Neuroscience: Trends in Research 1995*, pp. 317–322. Academic Press, San Diego, CA.

Bartlett, M. S., & Sejnowski, T. J. (1997). Viewpoint invariant face recognition using independent component analysis and attractor networks. In Mozer, M., Jordan, M., & Petsche, T. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 9 Cambridge, MA. MIT Press.

Bennett, A., & Craw, I. (1991). Finding image features using deformable templates and detailed prior statistical knowledge. In Mowforth, P. (Ed.), *Proceedings of British Machine Vision Conference*, pp. 233–239. Springer-Verlag.

Benson, P. J., & Perrett, D. I. (1994). Visual processing of facial distinctiveness. *Perception*, *23*, 75–93.

Berthold, M. R. (1994). A time delay radial basis function network for phoneme recognition. In *Proceedings of IEEE International Conference on Neural Networks*, Vol. 7, pp. 4470–4473 Orlando, FL. IEEE Computer Society Press.

Beymer, D. J. (1994). Face recognition under varying pose. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 756–761 Seattle, WA. IEEE Computer Society Press.

Beymer, D. J. (1995). Vectorizing face images by interleaving shape and texture computations. Tech. rep. 1537, AI Lab, MIT, Cambridge, MA.

Beymer, D. J., & Poggio, T. (1995). Face recognition from one example view. In *Proceedings of International Conference on Computer Vision*, pp. 500–507 Cambridge, MA. IEEE Computer Society Press.

Beymer, D. J., & Poggio, T. (1996). Image representations for visual learning. *Science*, *272*, 1905–1909.

Bichsel, M. (Ed.). (1995). *Proceedings of International Workshop on Automatic Face & Gesture Recognition*, Zurich, Switzerland. University of Zurich.

Bichsel, M., & Pentland, A. (1994). Human face recognition and the face image set's topology. *Computer Vision, Graphics, and Image Processing: Image Understanding*, *59*, 254–261.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychological Review*, *94*, 115–147.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Bossomaier, T. R. J. (1989). Efficient image representation by Gabor functions – an information theory approach. In Kulikowski, J. J., Dickinson, C. M., & Murray, I. J. (Eds.), *Seeing Contour and Colour*, pp. 698–704. Pergamon Press, Oxford, UK.

Bouattour, H., Fogelman-Soulié, F., & Viennet, E. (1992). Solving the human face recognition task using neural nets. In Aleksander, I., & Taylor, J. (Eds.), *Proceedings of International Conference on Artificial Neural Networks*, pp. 1595–1598 Brighton, UK. Elsevier Science Publishers BV.

Breuel, T. M. (1992). Geometric aspects of visual object recognition. Tech. rep. 1374, AI Lab, MIT, Cambridge, MA.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, *2*, 321–355.

Bruce, V. (1988). *Recognising Faces*. Lawrence Erlbaum Associates, London.

Bruce, V., & Green, P. (1990). *Visual Perception*. Lawrence Erlbaum Associates, London.

Bruce, V., Hanna, E., Dench, N., Healey, P., & Burton, M. (1992). The importance of 'mass' in line-drawings of faces. *Applied Cognitive Psychology*, *6*, 619–628.

Bruce, V., & Humphreys, G. W. (1994). Recognising objects and faces. *Visual Cognition*, *1*, 141–180.

Bruce, V., Valentine, T., & Baddeley, A. D. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology*, *1*, 109–120.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77, 305–327.

Brunelli, R., & Falavigna, D. (1995). Person recognition using multiple cues. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *17*, 955–966.

Brunelli, R., & Poggio, T. (1991). HyperBF networks for real object recognition. In Myopoulos, J., & Reiter, R. (Eds.), *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1278–1284 Sydney, Australia. Morgan Kaufmann.

Brunelli, R., & Poggio, T. (1992a). Face recognition through geometrical features. In Sandini, G. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 588, pp. 792–800 Santa Margherita Ligure, Italy. Springer-Verlag.

Brunelli, R., & Poggio, T. (1992b). HyperBF networks for gender classification. In *Proceedings of DARPA Image Understanding Workshop*, pp. 311–314 San Diego, CA.

Brunelli, R., & Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *15*, 1042–1052.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, USA*, *89*, 60–64.

Bülthoff, H. H., Edelman, S., & Tarr, M. (1995). How are three-dimensional objects represented in the brain?. *Cerebral Cortex*, *5*, 247–260.

Burton, A. M. (1994). Learning new faces an interactive activation and competition model. *Visual Cognition*, *1*, 313–348.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, 361–380.

Buxton, H., & Gong, S. (1995a). Advanced visual surveillance using Bayesian nets. In Mundy, J. L., & Strat, T. (Eds.), *Proceedings of IEEE Workshop on Context-Based Vision, International Conference on Computer Vision* Cambridge, MA. IEEE Computer Society Press.

Buxton, H., & Gong, S. (1995b). Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, *78*, 431–459.

Chellappa, R., Wilson, C. L., & Sirohey, S. (1995). Human and machine recognition of faces: A survey. In *Proceedings of IEEE*, Vol. 83, pp. 705–740.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, *2*, 302–309.

Cleeremans, A. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

Cliff, D. T., & Bullock, S. G. (1993). Adding 'foveal vision' to Wilson's Animat. *Adaptive Behaviour*, *2*, 47–70.

Cootes, T. F., & Taylor, C. J. (1992). Active shape models - 'smart snakes'. In Hogg, D., & Boyle, R. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 266–275 Leeds, UK. Springer-Verlag.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1992). Training models of shape from sets of examples. In Hogg, D., & Boyle, R. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 9–18 Leeds, UK. Springer-Verlag.

Cootes, T. F., Taylor, C. J., Lanitis, A., Cooper, D. H., & Graham, J. (1993). Building and using flexible models incorporating grey-level information. In *Proceedings of International Conference on Computer Vision*, pp. 242–246 Berlin, Germany. IEEE Computer Society Press.

Cottrell, G. W., Munro, P., & Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In *Proceedings of Annual Conference of the Cognitive Science Society*, pp. 461–473 Seattle, WA. Lawrence Erlbaum Associates.

Craw, I., & Cameron, P. (1991). Parameterising images for recognition and reconstruction. In Mowforth, P. (Ed.), *Proceedings of British Machine Vision Conference*, pp. 367–370. Springer-Verlag.

Craw, I., & Cameron, P. (1992). Face recognition by computer. In Hogg, D., & Boyle, R. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 498–507 Leeds, UK. Springer-Verlag.

Craw, I., Costen, N., Kato, T., Robertson, G., & Akamatsu, S. (1995). Automatic face recognition: combining configuration and texture. In Bichsel, M. (Ed.), *Proceedings of International Workshop on Automatic Face & Gesture Recognition*, pp. 53–58 Zurich, Switzerland. University of Zurich.

Craw, I., Tock, D., & Bennett, A. (1992). Finding face features. In Sandini, G. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 588, pp. 92–96 Santa Margherita Ligure, Italy. Springer-Verlag.

Daugman, J. G. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *36*, 1169–1179.

Diamond, R., & Carey, S. (1986). Why faces are, and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*, 107–117.

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, *6*, 345–358.

Duc, B., Maître, G., Fischer, S., & Bigün, J. (1997). Person authentication by fusing face and speech information. In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 311–318 Crans-Montana, Switzerland. Springer Verlag.

Edelman, S. (1995). Receptive fields for vision: from hyperacuity to object recognition. Tech. rep. CS-TR 95-29, Weizmann Institute, Israel.

Edelman, S., & Intrator, N. (1997). Learning as extraction of low-dimensional representations. In Medin, D., Goldstone, R., & Schyns, P. (Eds.), *Mechanisms of Perceptual Learning*. Academic Press, San Diego, CA. In press.

Edelman, S., & Poggio, T. (1992). Bringing the grandmother back into the picture: a memory-based view of object recognition. *International Journal of Pattern Recognition & Artificial Intelligence*, *6*, 37–62.

Edelman, S., Reisfeld, D., & Yeshurun, Y. (1992). Learning to recognize faces from examples. In Sandini, G. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 588, pp. 787–791 Santa Margherita Ligure, Italy. Springer-Verlag.

Edwards, G. J., Lanitis, A., Taylor, C. J., & Cootes, T. F. (1996). Modelling the variability in face images. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pp. 328–333 Killington, VT. IEEE Computer Society Press.

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1997). Learning to identify and track faces in image sequences. In Clark, A. F. (Ed.), *Proceedings of British Machine Vision Conference*, pp. 130–139 Colchester, UK. BMVA Press.

Ellis, H. D., & Young, A. W. (1989). Are faces special?. In Young, A. W., & Ellis, H. D. (Eds.), *Handbook of Research on Face Processing*. North-Holland, Amsterdam, The Netherlands.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Essa, I. (Ed.). (1996). *Proceedings of International Conference on Automatic Face & Gesture Recognition*, Killington, VT. IEEE Computer Society Press.

Ezzat, T., & Poggio, T. (1996). Facial analysis and synthesis using image-based models. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pp. 116–121 Killington, VT. IEEE Computer Society Press.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*, 2379–2394.

Fleming, M. K., & Cottrell, G. W. (1990). Categorization of faces using unsupervised feature extraction. In *Proceedings of International Joint Conference on Neural Networks*, pp. 65–70 San Diego, CA.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, *20*, 121–136.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*, 119–130.

Girosi, F. (1992). Some extensions of radial basis functions and their applications in artificial intelligence. *Computers & Mathematics with Applications*, *24*(12), 61–80.

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7, 219–269.

Gong, S., McKenna, S. J., & Collins, J. J. (1996). An investigation into face pose distributions. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pp. 265–270 Killington, VT. IEEE Computer Society Press.

Gutta, S., Huang, J., Singh, D., Shah, I., Takacs, B., & Wechsler, H. (1995). Benchmark studies on face recognition. In Bichsel, M. (Ed.), *Proceedings of International Workshop on Automatic Face & Gesture Recognition*, pp. 227–231 Zurich, Switzerland. University of Zurich.

Gutta, S., & Wechsler, H. (1997). Face recognition using hybrid classifiers. *Pattern Recognition*, *30*, 539–553.

Hancock, P. J. B., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network: Computation in Neural Systems*, *3*, 61–70.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (1997). A comparison of two computer-based face recognition systems with human perceptions of faces. *Vision Research*, *(Submitted)*.

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1995). Preprocessing images of faces: correlations with human perceptions of distinctiveness and familiarity. In *Proceedings of IEE Fifth International Conference on Image Processing and its Applications* Edinburgh, Scotland.

Hay, D. C., & Young, A. (1982). The human face. In Ellis, H. D. (Ed.), *Normality and Pathology in Cognitive Functions*. Academic Press, San Diego, CA.

Hay, D. C., Young, A., & Ellis, A. W. (1991). Routes through the face recognition system. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *43*, 761–791.

Hertz, J. A., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City CA.

Howell, A. J., & Buxton, H. (1995). Invariance in radial basis function neural networks in human face classification. *Neural Processing Letters*, *2*(3), 26–30.

Jordan, M. I. (1989). Serial order: A parallel, distributed processing approach. In Elman, J. L., & Rumelhart, D. E. (Eds.), *Advances in Connectionist Theory: Speech*. Lawrence Erlbaum, Hillsdale, NJ.

Kanade, T. (1973). Picture processing by computer complex and recognition of human faces. Tech. rep., Department of Information Science, Kyoto University, Japan.

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *12*, 103–108.

Kittler, J., Li, Y. P., Matas, J., & Sánchez, M. U. R. (1997). Combining evidence in multimodal personal identity recognition systems. In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 327–334 Crans-Montana, Switzerland. Springer Verlag.

Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd edition). Springer-Verlag, Berlin, Germany.

Kohonen, T., Oja, E., & Lehtiö, P. (1981). Storage and processing of information in distributed associative memory systems. In Hinton, G. E., & Anderson, J. A. (Eds.), *Parallel Models of Associative Memory*, pp. 105–143. Lawrence Erlbaum Associates, Hillsdale, NJ.

Konen, W., & Schulze-Krüger, E. (1995). ZN-Face: a system for access control using automatic face recognition. In Bichsel, M. (Ed.), *Proceedings of International Workshop on Automatic Face & Gesture Recognition*, pp. 18–23 Zurich, Switzerland. University of Zurich.

Krüger, N., Pötzsch, M., Maurer, T., & Rinne, M. (1996). Estimation of face position and pose with labeled graphs. In Fisher, R. B., & Trucco, E. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 735–743 Edinburgh. BMVA Press.

Kruizinga, P. (1995). The Face Recognition Home Page.
Web page: `http://www.cs.rug.nl/~peterkr/FACE/face.html`.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*, 300–311.

Lando, M., & Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network: Computation in Neural Systems*, *6*, 551–576.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1995a). An automatic face identification system using flexible appearance models. *Image & Vision Computing*, *13*, 393–401.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1995b). A unified approach to coding and interpreting face images. In *Proceedings of International Conference on Computer Vision*, pp. 368–373 Cambridge, MA. IEEE Computer Society Press.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *19*, 743–756.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural network approach. *IEEE Transactions on Neural Networks*, *8*, 98–113.

Le Cun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In Arbib, M. A. (Ed.), *The Handbook of Brain Theory and Neural Networks*, pp. 255–267. MIT Press, Cambridge, MA.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551.

Lin, S.-H., Kung, S.-Y., & Lin, L.-J. (1997). Face recognition/detection by probabilistic decision-based neural network. *IEEE Transactions on Neural Networks*, *8*, 114–132.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*, 401–414.

Lucas, S. M. (1997). Face recognition with the continuous n-tuple classifier. In Clark, A. F. (Ed.), *Proceedings of British Machine Vision Conference*, pp. 222–231 Colchester, UK. BMVA Press.

Manjunath, B. S., Chellappa, R., & von der Malsburg, C. (1992). A feature based approach to face recognition. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 373–378 Champaign, IL. IEEE Computer Society Press.

Marr, D. (1982). *Vision*. Freeman, San Francisco, CA.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of Royal Society London, Series B*, *207*, 187–217.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of Royal Society London, Series B*, *200*, 269–294.

Maurer, T., & von der Malsburg, C. (1995a). Learning feature transformations to recognize faces rotated in depth. In Fogelman-Soulié, F., & Gallinari, P. (Eds.), *Proceedings of International Conference on Artificial Neural Networks*, Vol. 1, pp. 353–358 Paris, France. EC2 & Cie.

Maurer, T., & von der Malsburg, C. (1995b). Single-view based recognition of faces rotated in depth. In Bichsel, M. (Ed.), *Proceedings of International Workshop on Automatic Face & Gesture Recognition*, pp. 248–253 Zurich, Switzerland. University of Zurich.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, *88*, 375–407.

McKenna, S. J., & Gong, S. (1996). Tracking faces. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pp. 271–276 Killington, VT. IEEE Computer Society Press.

McKenna, S. J., & Gong, S. (1997). Non-intrusive person authentication for access control by visual tracking and face recognition. In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 177–184 Crans-Montana, Switzerland. Springer Verlag.

McKenna, S. J., Gong, S., & Collins, J. J. (1996). Face tracking and pose representation. In Fisher, R. B., & Trucco, E. (Eds.), *Proceedings of British Machine Vision Conference*, pp. 755–764 Edinburgh. BMVA Press.

McKenna, S. J., Gong, S., & Raja, Y. (1997a). Face recognition in dynamic scenes. In Clark, A. F. (Ed.), *Proceedings of British Machine Vision Conference*, pp. 140–151 Colchester, UK. BMVA Press.

McKenna, S. J., Gong, S., Würtz, R. P., Tanner, J., & Banin, D. (1997b). Tracking facial feature points with Gabor wavelets and shape models. In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 35–42 Crans-Montana, Switzerland. Springer Verlag.

Millward, R., & O'Toole, A. (1986). Recognition memory transfer between spatial-frequency analysed faces. In Ellis, H. D., Jeeves, M. A., Newcombe, F., & Young, A. W. (Eds.), *Aspects of Face Processing*, pp. 34–44. Nijhoff, Dordrecht, The Netherlands.

Moghaddam, B., & Pentland, A. (1995). Probabilistic visual learning for object detection. In *Proceedings of International Conference on Computer Vision*, pp. 786–793 Cambridge, MA. IEEE Computer Society Press.

Moody, J., & Darken, C. (1988). Learning with localized receptive fields. In Touretzky, D., Hinton, G., & Sejnowski, T. (Eds.), *Proceedings of 1988 Connectionist Models Summer School*, pp. 133–143 Pittsburgh, PA. Morgan Kaufmann.

Moody, J., & Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, *1*, 281–294.

Moses, Y., Adini, Y., & Ullman, S. (1994). Face recognition: the problem of compensating for illumination changes. In Eklundh, J. O. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 800, pp. 286–296 Stockholm, Sweden. Springer-Verlag.

Moses, Y., & Ullman, S. (1992). Limitations of non model-based recognition schemes. In Sandini, G. (Ed.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 588, pp. 820–828 Santa Margherita Ligure, Italy. Springer-Verlag.

Mozer, M. C. (1994). Neural net architectures for temporal sequence processing. In Weigend, A. S., & Gershenfeld, N. A. (Eds.), *Time Series Prediction: Predicting the Future and Understanding the Past*, pp. 243–264. Addison-Wesley, Redwood City, CA.

Murase, H., & Nayar, S. (1995). Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, *14*, 5–24.

Musavi, M. T., Ahmad, W., Chan, K. H., Faris, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural Networks*, *5*, 595–603.

ORL (1994). The Olivetti Research Laboratory database of faces.
Web page: `http://www.cam-orl.co.uk/facedatabase.html`.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). A low-dimensional representation of faces in high dimensions of the space. *Journal of the Optical Society of America A*, *10*, 405–410.

O'Toole, A. J., Bülthoff, H. H., & Walker, L. C. (1995). Face recognition across viewpoint. Tech. rep. 21, Max–Planck–Institut für biologische Kybernetik Arbeitsgrupp Bülthoff, Tübingen, Germany.

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In Long, J., & Baddeley, A. (Eds.), *Attention and Performance IX*, pp. 135–151. Erlbaum, Hillsdale, NJ.

Pentland, A. (1996). Smart rooms. *Scientific American*, *274*(4), 68–76.

Pentland, A., Moghaddam, B., & Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 84–91 Seattle, WA. IEEE Computer Society Press.

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1989). Visual neurons responsive to faces. *Trends In Neurosciences*, *10*, 358–364.

Perrett, D. I., & Oram, M. W. (1993). Neurophysiology of shape processing. *Image & Vision Computing*, *11*, 317–333.

Petkov, N. (1995). Biologically motivated computationally intensive approaches to image pattern recognition. *Future Generation Computer Systems*, *11*, 451–465.

Petkov, N., Kruizinga, P., & Lourens, T. (1993). Biologically motivated approach to face recognition. In Mira, J., Cabestany, J., & Prieto, A. (Eds.), *New Trends in Neural Computation, Proceedings of International Workshop on Artificial Neural Networks, Lecture Notes in Computer Science*, Vol. 686, pp. 68–77 Sitges, Spain. Springer-Verlag.

Phillips, P. J., Moon, H., Rauss, P. J., & Rizvi, S. A. (1997). The FERET September 1996 database and evaluation procedure. In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 395–402 Crans-Montana, Switzerland. Springer Verlag.

Phillips, P. J., Rauss, P. J., & Der, S. Z. (1996). FERET (face recognition technology) recognition algorithm development and test results. Tech. rep. ARL-TR-995, US Army Research Laboratory, Adelphi, MD.

Pigeon, S., & Vandendorpe, L. (1997). The M2VTS multimodal face database (release 1.00). In *Proceedings of 1st International Conference on Audio & Video-based Biometric Person Authentication, Lecture Notes in Computer Science*, Vol. 1206, pp. 403–410 Crans-Montana, Switzerland. Springer Verlag.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*, 263–266.

Poggio, T., & Girosi, F. (1990a). Networks for approximation and learning. In *Proceedings of IEEE*, Vol. 78, pp. 1481–1497.

Poggio, T., & Girosi, F. (1990b). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.

Pomerleau, D. A. (1989). ALVINN: An autonomous land vehicle in a neural network. In Touretzky, D. S. (Ed.), *Advances in Neural Information Processing Systems*, Vol. 1, pp. 305–313 San Mateo, CA. Morgan Kaufmann.

Popvision (1994). POPVISION library routines.
Web page: `http://www.cogs.susx.ac.uk/users/davidy/teachvision/vision0.html`.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical Recipes in C*. Cambridge University Press, Cambridge.

Psarrou, A., & Buxton, H. (1993). Hybrid architecture for understanding motion sequences. *Neurocomputing*, *5*, 221–241.

Psarrou, A., & Buxton, H. (1994). Motion analysis with recurrent neural nets. In *Proceedings of International Conference on Artificial Neural Networks*, pp. 54–57 Sorrento, Italy. Springer-Verlag.

Rao, R. P. N., & Ballard, D. H. (1995). Natural basis functions and topographic memory for face recognition. In Mellish, C. S. (Ed.), *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 10–17 Montréal, Canada. Morgan Kaufmann.

Rhodes, G., & McLean, I. G. (1990). Distinctiveness and expertise effects with homogeneous stimuli: Towards a model of configural coding. *Perception*, *19*, 773–794.

Robertson, G., & Craw, I. (1994). Testing face recognition systems. *Image & Vision Computing*, *12*, 609–614.

Rolls, E. T. (1994). Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes*, *33*, 113–138.

Rolls, E. T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research*, *66*, 177–185.

Rosenblum, M., & Davis, L. S. (1996). An improved radial basis function network for autonomous road-following. *IEEE Transactions on Neural Networks*, 7, 1111–1120.

Rosenblum, M., Yacoob, Y., & Davis, L. S. (1996). Human emotion recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7, 1121–1138.

Rowley, H. A., Baluja, S., & Kanade, T. (1996). Human face detection in visual scenes. In Touretzky, D. S., Mozer, M. C., & Hasselmo, M. E. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 8, pp. 875–881 Cambridge, MA. MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.

Saha, A., & Keeler, J. D. (1990). Algorithms for better representation and faster learning in radial basis function networks. In Touretzky, D. S. (Ed.), *Advances in Neural Information Processing Systems*, Vol. 2 San Mateo, CA. Morgan Kaufmann.

Samal, A., & Iyengar, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, *15*, 65–77.

Samaria, F. S. (1994). *Face Recognition using Hidden Markov Models*. Ph.D. thesis, Cambridge University, UK.

Samaria, F. S., & Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision* Sarasota, FL.

Sergent, J., Ohta, S., MacDonald, B., & Zuck, E. (1994). Segregated processing of facial identity and emotion in the human brain: a PET study. *Visual Cognition*, *1*, 349–370.

Simard, P., Victorri, B., le Cun, Y., & Denker, J. (1992). Tangent prop - a formalism for specifying selected invariances in an adaptive network. In Moody, J. E., Lippman, R. P., & Hanson, S. J. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 4, pp. 895–903 San Mateo, CA. Morgan Kaufmann.

Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, *384*, 460–463.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, *3*, 109–118.

Stokbro, K., Umberger, D. K., & Hertz, J. A. (1990). Exploiting neurons with localized receptive fields to learn chaos. *Complex Systems*, *4*, 603–622.

Stonham, T. J. (1986). Practical face recognition and verification with WISARD. In Ellis, H. D., Jeeves, M. A., Newcombe, F., & Young, A. W. (Eds.), *Aspects of Face Processing*. Martinus Nijhoff, Dordrecht.

Stryker, M. P. (1991). Temporal associations. *Nature*, *354*, 108–109.

Sung, K.-K., & Poggio, T. (1994). Example-based learning for view-based human face detection. Tech. rep. 1521/CBCL paper 112, AI Lab, MIT, Cambridge, MA.

Ting, C., & Chuang, K.-C. (1993). An adaptive algorithm for Neocognitron to recognize analog images. *Neural Networks*, *6*, 285–299.

Tistarelli, M. (1994). Recognition by using an active/space-variant sensor. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 833–837 Seattle, WA. IEEE Computer Society Press.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71–86.

Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition*, *32*, 193–254.

Ullman, S. (1996). *High-level Vision*. MIT Press, Cambridge, MA.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *13*, 992–1006.

Valentin, D., Abdi, H., O'Toole, A. J., & Cottrell, G. W. (1994). Connectionist models of face processing: A survey. *Pattern Recognition*, *27*, 1208–1230.

Vetter, T., & Poggio, T. (1996). Image synthesis from a single example image. In Buxton, B., & Cipolla, R. (Eds.), *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 1065, pp. 652–659 Cambridge, UK. Springer-Verlag.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *37*, 328–339.

Wallis, G., & Rolls, E. T. (1997). A model of invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*, 167–194.

Wallis, G., Rolls, E. T., & Földiák, P. (1993). Learning invariant responses to the natural transformations of objects. In *Proceedings of International Joint Conference on Neural Networks*, pp. 1087–1090.

Weng, J. J., Ahuja, N., & Huang, T. S. (1993). Learning recognition and segmentation of 3-D objects from 2-D images. In *Proceedings of International Conference on Computer Vision*, pp. 121–128 Berlin, Germany. IEEE Computer Society Press.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, Vol. 4, pp. 96–104. IRE, New York.

Wiskott, L., Fellous, J. M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *19*, 775–779.

Wiskott, L., & von der Malsburg, C. (1996). Face recognition by dynamic link matching. Tech. rep. IR-INI 96-05, Institut für Neuroinformatik, Ruhr-Universität Bochum, Bochum, Germany.

Würtz, R. P. (1994). *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*. Ph.D. thesis, Bochum University, Germany.

Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: everyday difficulties and errors in recognising people. *British Journal of Psychology*, *76*, 495–523.

Young, D. S. (1987). Representing images for computer vision. Tech. rep. CSRP 96, School of Cognitive and Computing Sciences, University of Sussex, UK.

Yuille, A. L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, *3*, 59–70.

Yuille, A. L., Hallinan, P. W., & Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, *8*, 99–111.

Appendix A

Face Database Information

This appendix gives specific details on the face databases used for the experimental chapters. Section A.2.2 contains Euclidean distance comparisons for the Sussex database.

A.1 The ORL Database

The Olivetti Research Laboratory (ORL) database of faces (ORL, 1994) has been used for the initial experiments in Section 2.6. It is valuable, as there are a wide range of published face recognition results based on the database which can be used for comparison. It contains 400 greyscale images of 40 people at a resolution of 92×112, see Figure A.2. Each individual is represented by 10 images, and for some, these have been taken at different times.

Variations allowed in the image included lighting, facial expressions (such as open or closed eyes and smiling or not smiling) and facial details (such as glasses or no glasses). All the images were taken against a plain background, with tilt and rotation up to 20°, and scale variation up to 10%.

A.2 The Sussex Database

The Sussex Database is designed to assess how the performance of a particular face recognition technique will be affected by significant pose variations. It only contains data for ten people, which is a relatively small number by current face database standards. However, the main purpose of the database is not to test how many individuals a recognition system can discriminate, as there are

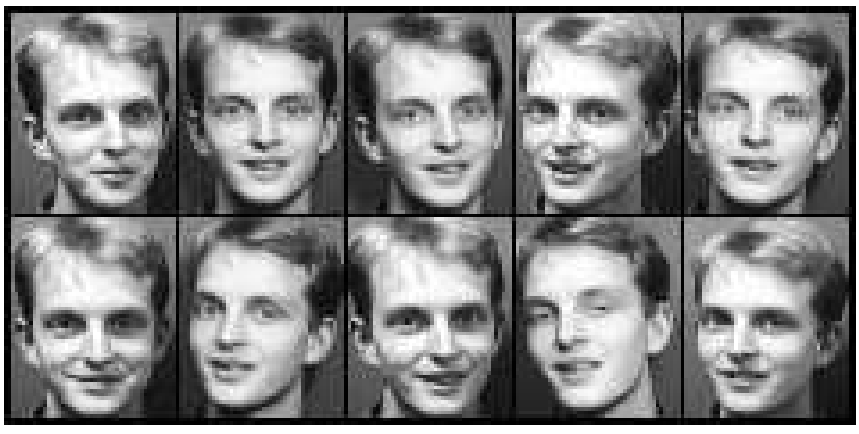


Figure A.1: Set of 10 images for one person in ORL database, illustrating moderate *x*-, *y*- and *z*-axis rotation with expression and illumination variation.

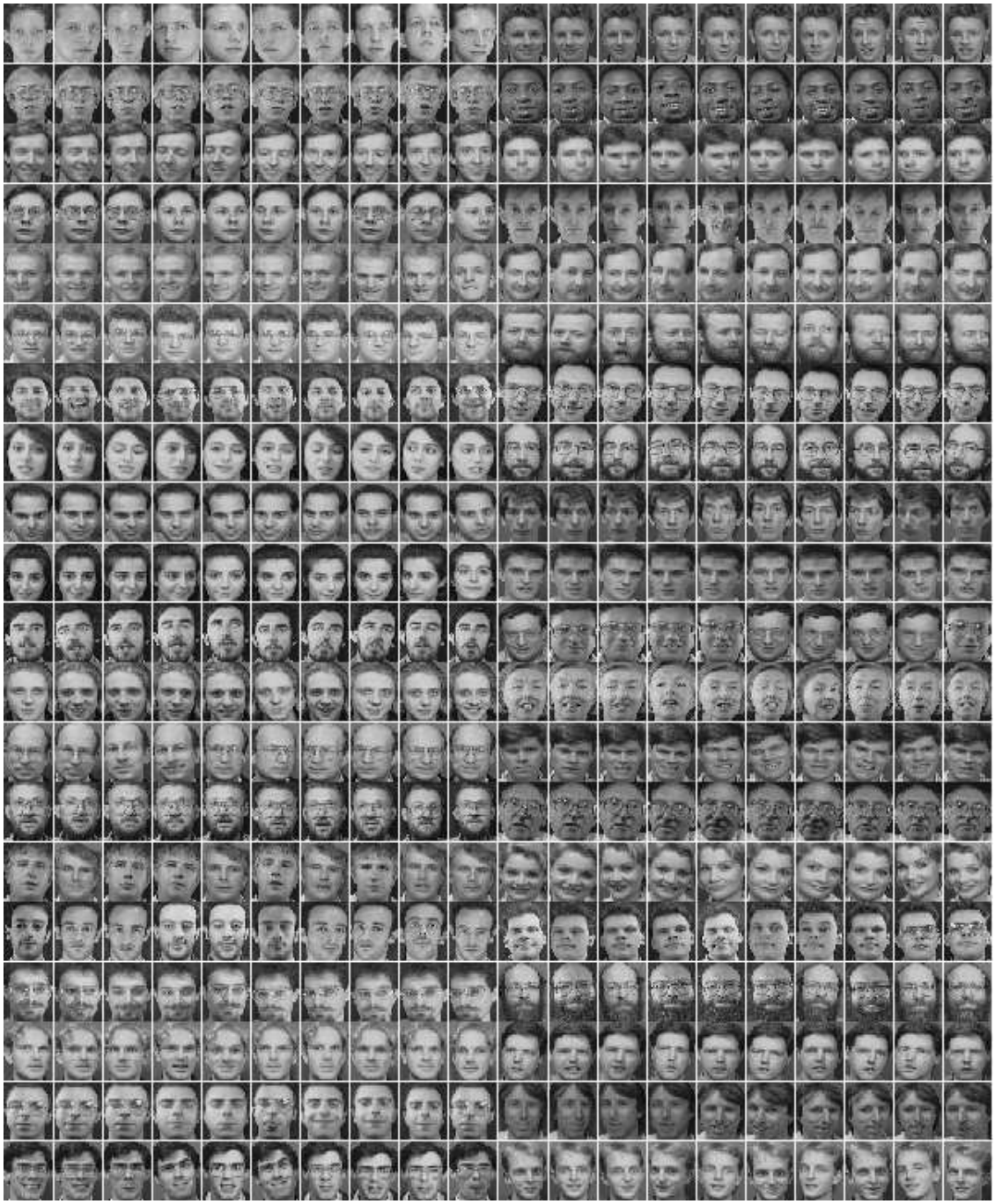


Figure A.2: The complete ORL Database.

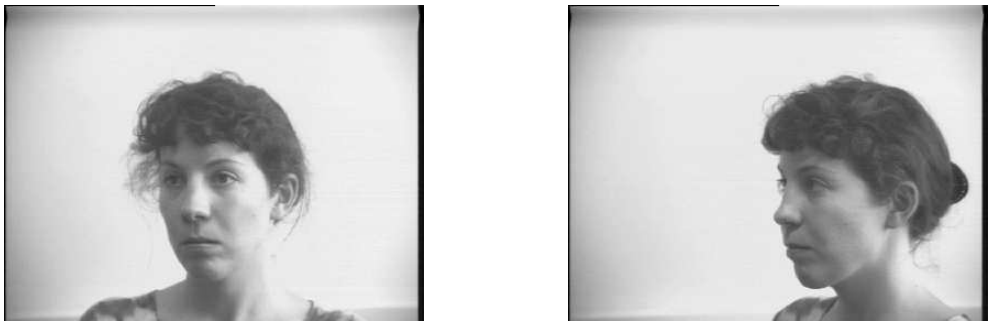
many publically-available databases that could be used for this purpose. The use of ten people is sufficient that the task is not trivial, but not so large that computation is excessive.

For each of the ten individuals in the database, ten images of the head and shoulders were taken in ten different positions from about 2 metres away in 10° steps from face-on to profile of the left side, 90° in all. This gave a data set of 100 8-bit grey-scale 384×287 images (see Figure A.3) from ten individuals. Lighting and facial expressions have been kept fairly constant to focus more clearly on pose variation.

A.2.1 Localization

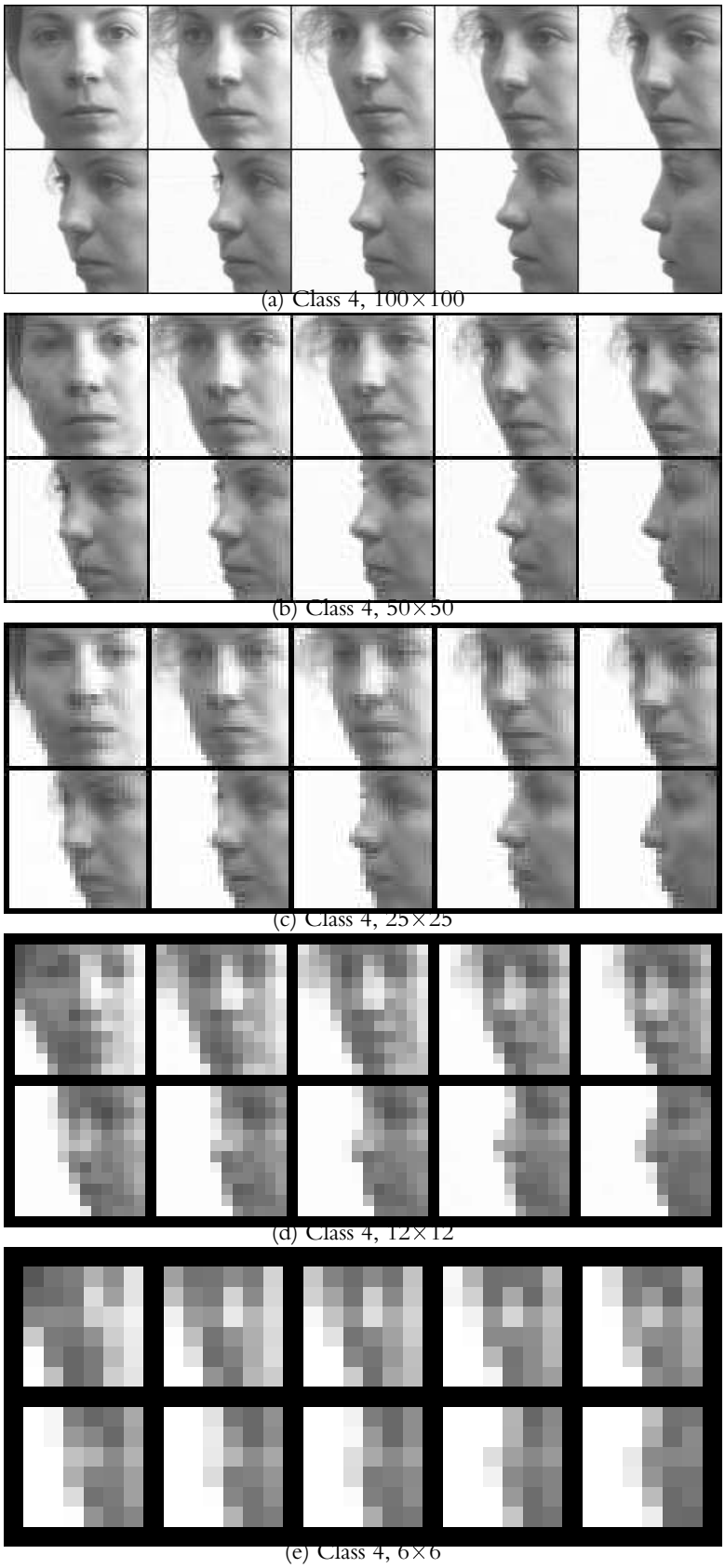
A 100×100-pixel 'window', containing all required facial information, was located manually in each image, see Figure A.4(a), and extracted for further processing. This size of window was large enough to give all of the main face features (if visible) without large areas of hair (Task Requirement 3(d)iii) or background (Task Requirement 3(c)iv), both of which are transient and may mislead recognition at a later date. Figures A.4(b-e) show this window subsampled to a range of resolutions.

Figures A.5 and A.6 show all 10 views for all 10 people in the Sussex database. The image window is centred by hand on the tip of the person's nose, so that visible features on profiles, for instance, should be in roughly similar locations to face-on. This means that some images, such as the profiles, do contain some background information. The background was a uniform plain white wall, and is anticipated to be a neutral component of the data. Figure A.7 illustrates an alternative face segmentation technique to the one used above.



(a) (b)

Figure A.3: Two examples of original 384×287 images from the Sussex database, showing pose angles (a) 10°, (b) 60°.



(a) Class 4, 100×100

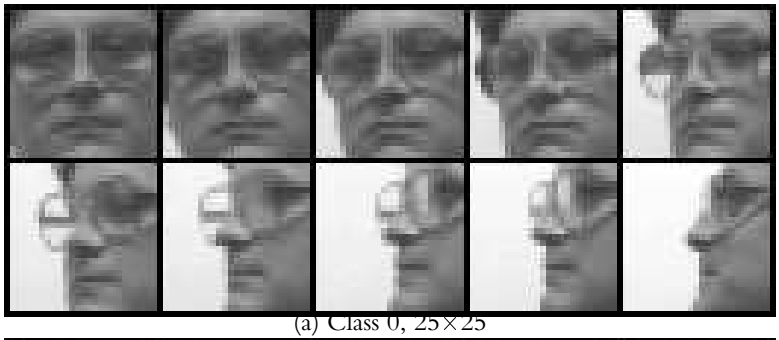
(b) Class 4, 50×50

(c) Class 4, 25×25

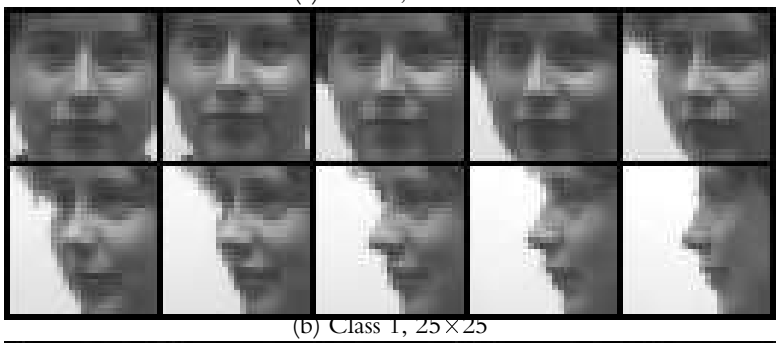
(d) Class 4, 12×12

(e) Class 4, 6×6

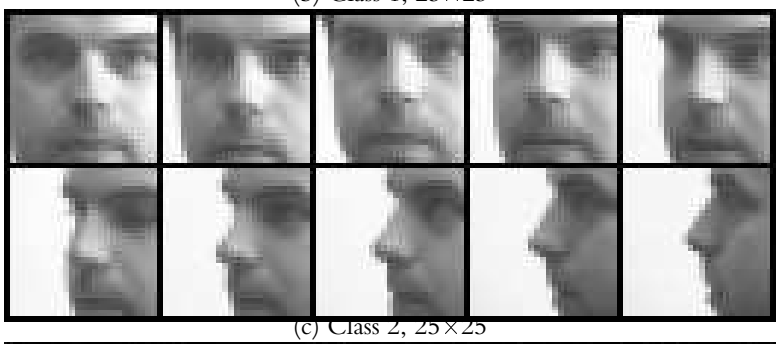
Figure A.4: Examples of all ten images for one person (class 4) from the Sussex database, nose-centred and subsampled before preprocessing, showing a *y*-axis rotation of 90°.



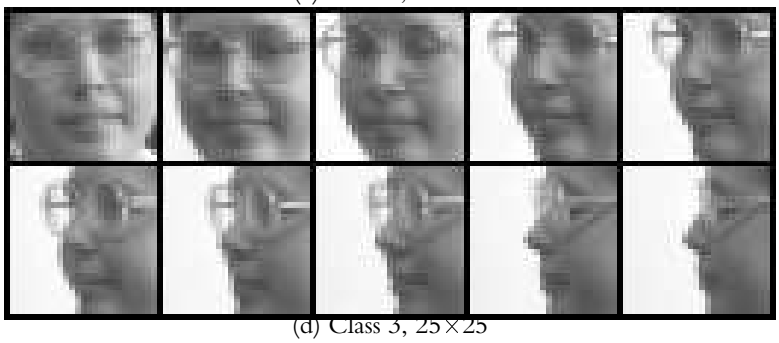
(a) Class 0, 25×25



(b) Class 1, 25×25

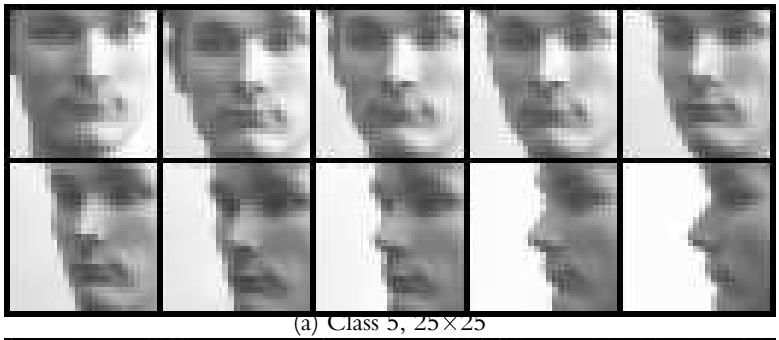


(c) Class 2, 25×25

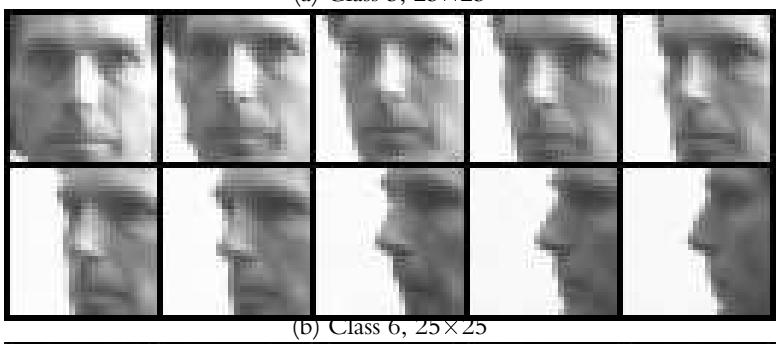


(d) Class 3, 25×25

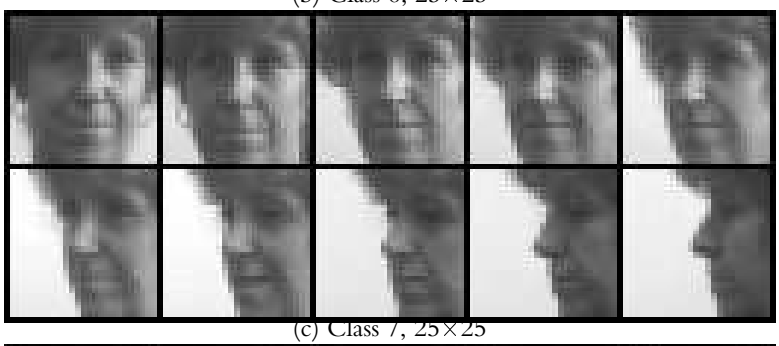
Figure A.5: All ten images for classes 0–3 from the Sussex database, nose-centred and subsampled to 25×25 before preprocessing.



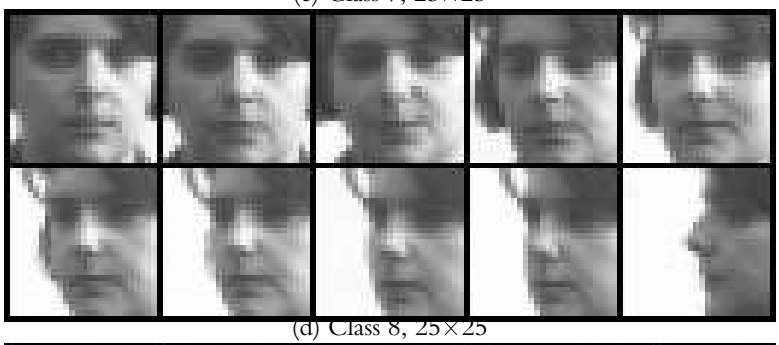
(a) Class 5, 25×25



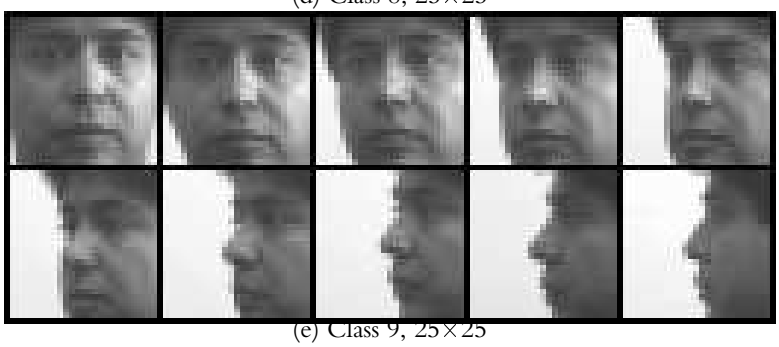
(b) Class 6, 25×25



(c) Class 7, 25×25

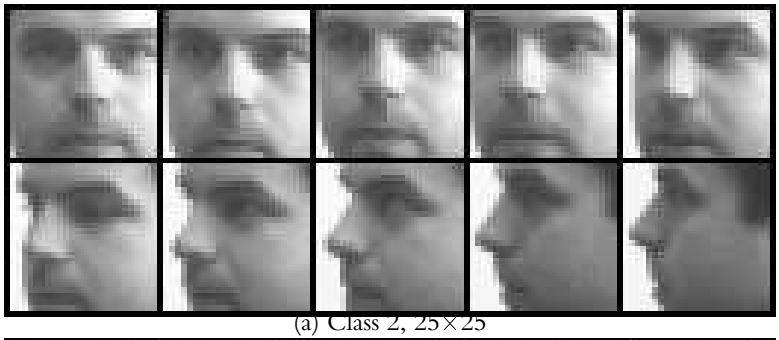


(d) Class 8, 25×25

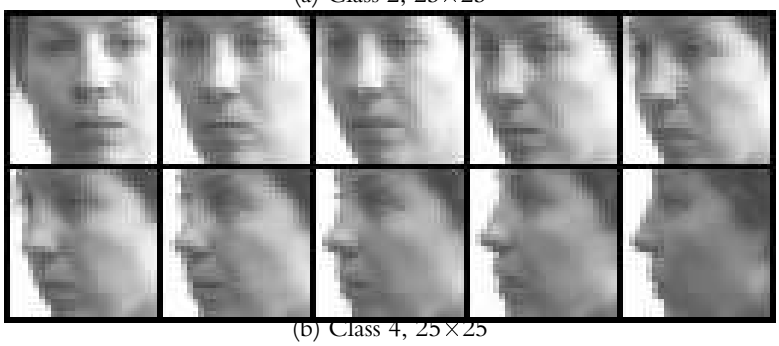


(e) Class 9, 25×25

Figure A.6: As for Figure A.5, but for classes 5–9.



(a) Class 2, 25×25



(b) Class 4, 25×25

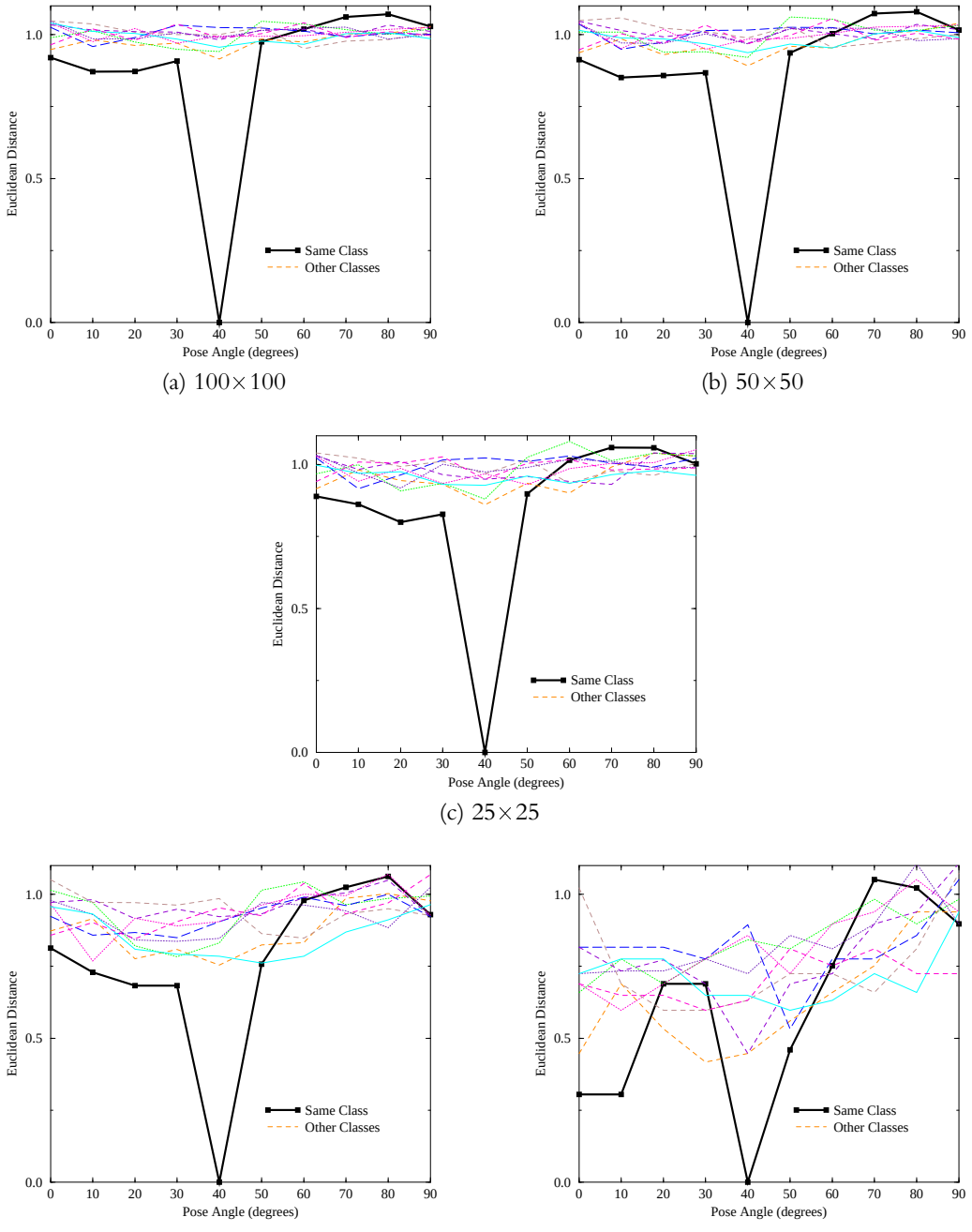
Figure A.7: As for Figure A.5, but using face-centering, rather than nose-centering, for localization of faces and only showing classes 2 and 4.
Note that this face-centering technique only attempts to fill the image with as much surface area from the face as is possible. A true, 'pose-free' centering algorithm would use head mass for localization, and the face area extracted would therefore contain the entire head.

A.2.2 Euclidean Distance Comparisons

These figures refer to experiments in Chapter 3, and illustrate Euclidean distance comparisons for the face images in the Sussex database. In all the graphs, one image is compared to all the others (including itself), and all 100 distances are shown, connected by lines according to class. The zero value can be seen where the image is compared to itself.

Figures A.8 and A.9 each show how the Euclidean distances vary through resolution for one specific image. Figures A.10 is similar, except the face-centering technique was used for segmentation, rather than nose-centering. Figures A.11 uses pose classes rather than identity classes.

Figure A.12, which shows all Euclidean distances for six individual images using DoG preprocessing at the 25×25 resolution from the Sussex Database, three each from classes 0 and 1 using pose angles of 0° (frontal), 40° and (a) 90° (profile). Figure A.13 is the same, except using Gabor preprocessing.

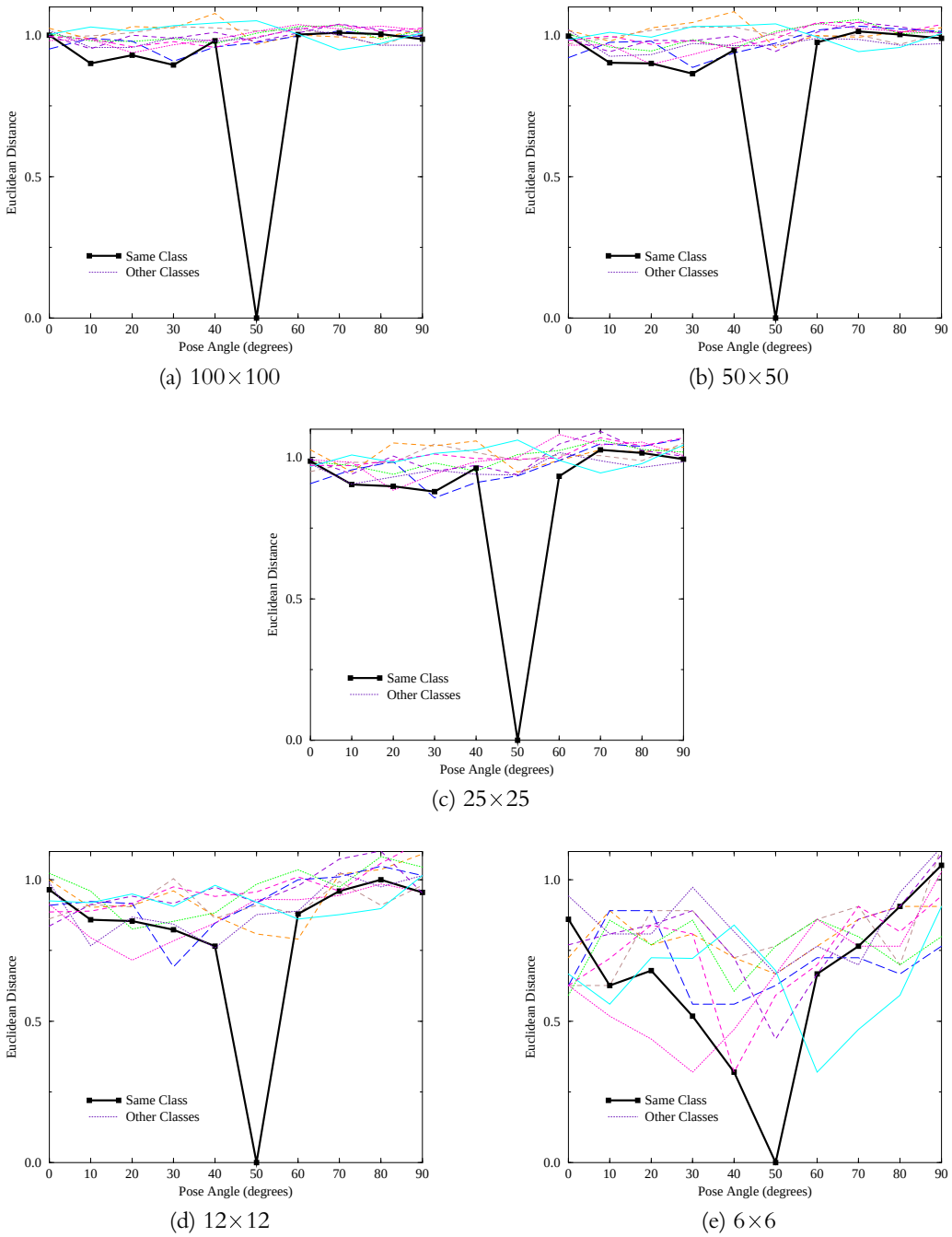


(a) 100×100 (b) 50×50

(c) 25×25

(d) 12×12 (e) 6×6

Figure A.8: Euclidean distances from one reference face image (pose 40° from class 0) to all others from the Sussex database, at varying resolutions using single scale DoG preprocessing.
Each line denotes distances from the reference image to comparison images from one class: the thin lines showing inter-class distances (where the class for the comparison image is different to that for the reference image), the thick line showing intra-class distances (same class as reference image).

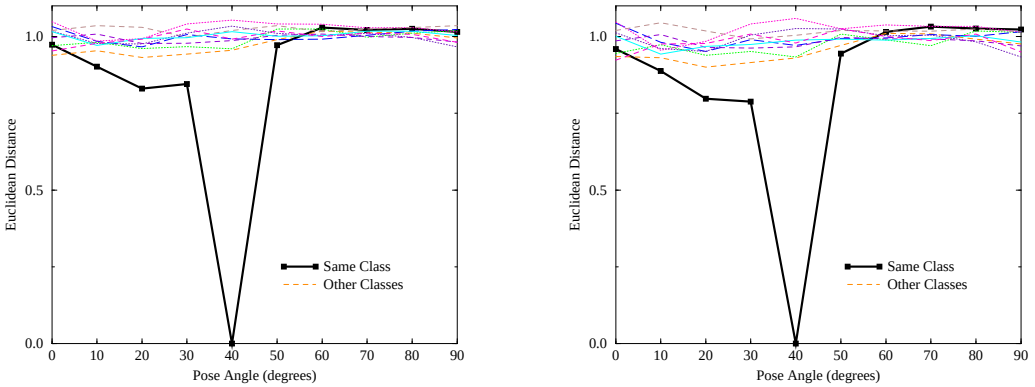


(a) 100×100 (b) 50×50

(c) 25×25

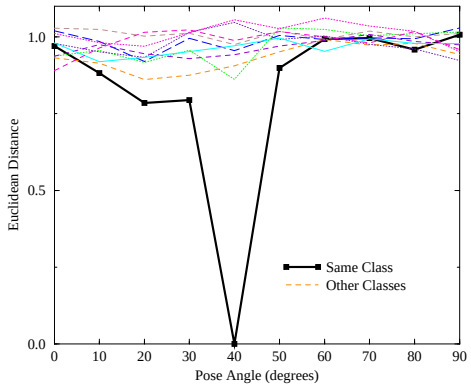
(d) 12×12 (e) 6×6

Figure A.9: As for Figure A.8, but for pose 50° from another class (5) of face images.

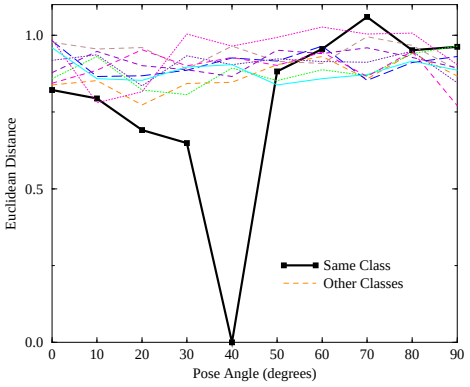


(a) 100×100

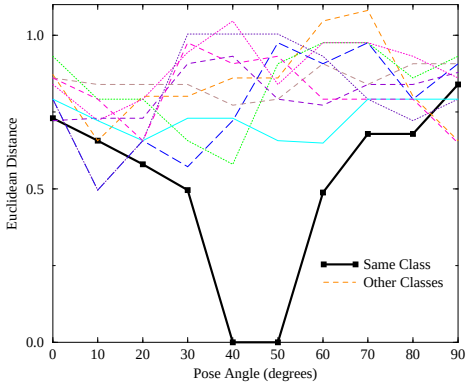
(b) 50×50



(c) 25×25

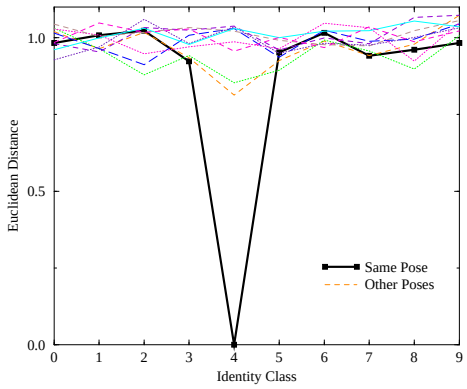


(d) 12×12



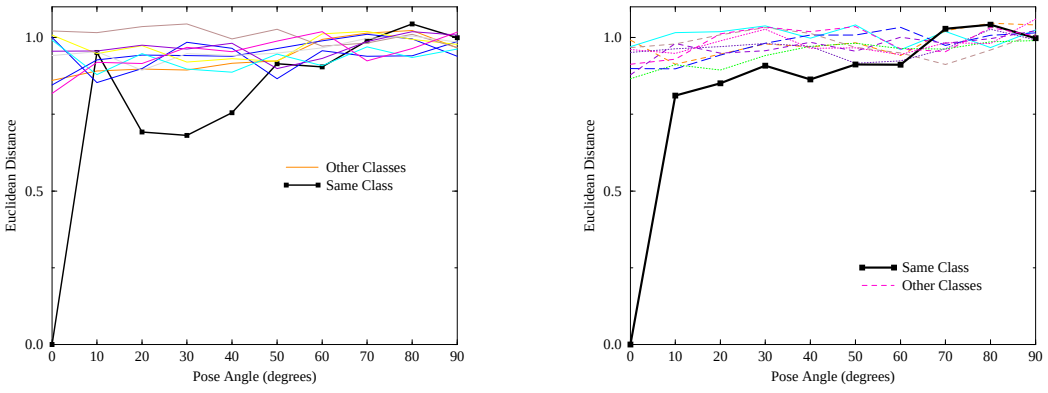
(e) 6×6

Figure A.10: As for Figure A.8, but using face-centering rather than nose-centering during face localization.



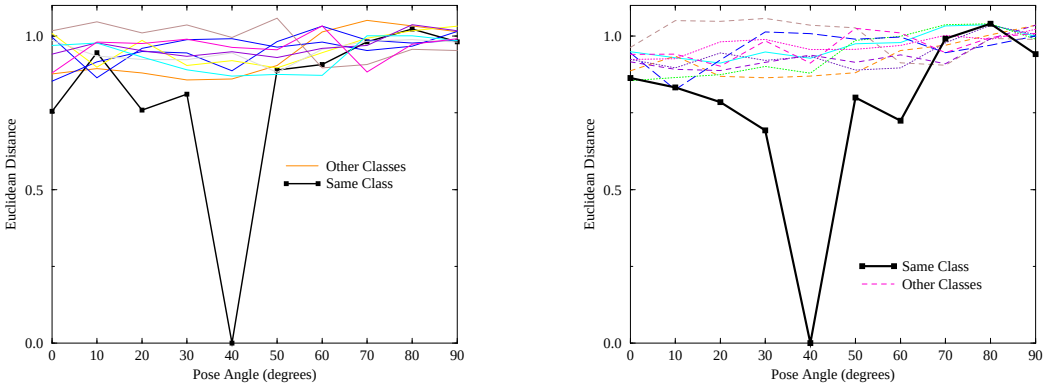
(a) 25×25

Figure A.11: As for Figure A.8, but using classes based on pose, rather than identity. Each line represents a pose class, each point on x-axis an identity class (person).



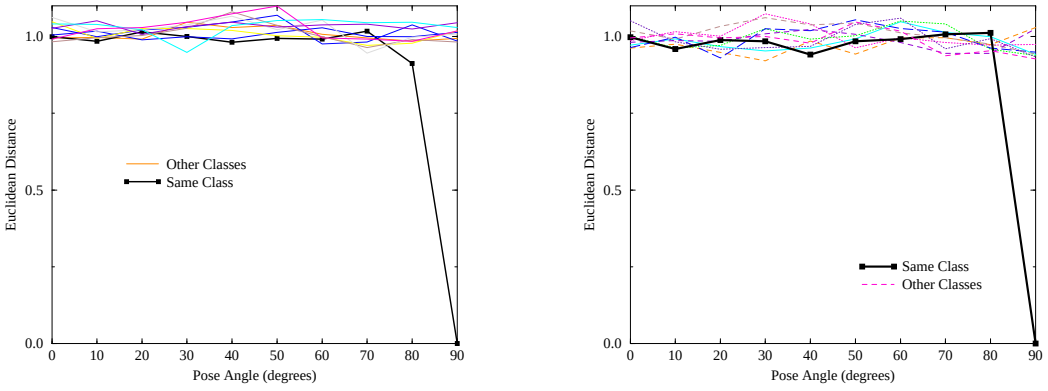
(i) (ii)

(a) Pose angle 0° (frontal)



(i) (ii)

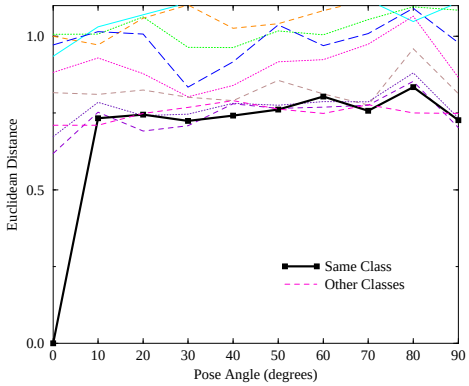
(b) Pose angle 40° (mid-angle view)



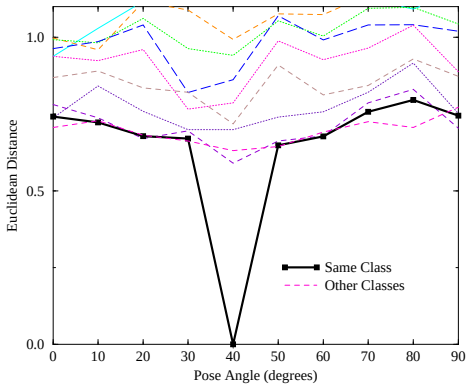
(i) (ii)

(c) Pose angle 90° (profile)

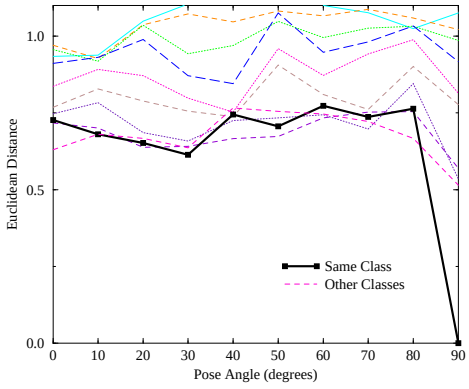
Figure A.12: Euclidean distances from single 25×25 face images to all others from the Sussex database with single scale DoG preprocessing.
Comparing images at specific pose angles from (i) class 1 and (ii) class 8 to all others. Each line denotes one class: the thin lines showing inter-class distances, the thick intra-class distances. For comparison, Figure A.13 shows the distances for the images using in (ii), but using Gabor preprocessing.



(a) Pose angle 0° (frontal)



(b) Pose angle 40° (mid-angle view)



(c) Pose angle 90° (profile)

Figure A.13: Euclidean distances from single 25×25 face images to all others from the Sussex database with 'A3' Gabor preprocessing.
Comparing images at specific pose angles from class 8 to all others. Each line denotes one class: the thin lines showing inter-class distances, the thick intra-class distances. For comparison, Figure A.12(ii) shows the distances for the same images, but using DoG preprocessing.

A.3 QMW Image Sequences

The image sequences used in the tests reported here are the result of collaboration with Stephen McKenna and Shaogang Gong at Queen Mary and Westfield College (QMW), London, and are split into two types of sequence: Primary and Secondary.

A.3.1 Primary Sequences

The Primary image sequences are intended to provide suitable data to train the system for the on-line classification, see Figures A.14 and A.15. They consist of a person moving from one profile view to the other whilst sitting on a chair, to limit body movement. All sequences are taken against a plain, mid-grey background to limit background effects, such as recognition based on distinctive background.

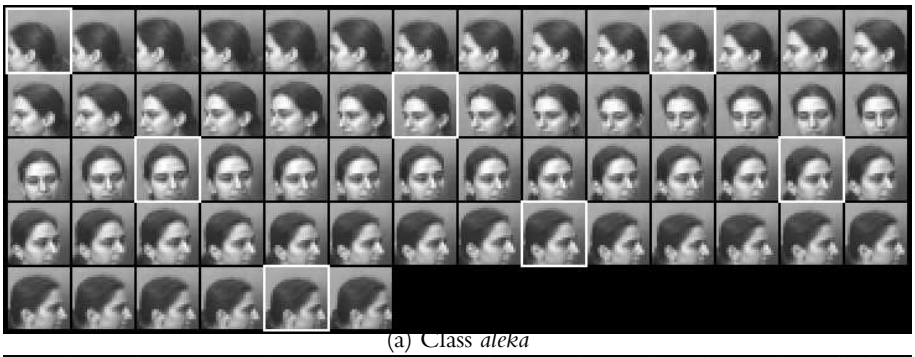
Eight Primary sequences have been collected so far, each featuring a different person. They are variable in length, depending on how fast the person moved from one side to the other, ranging in length from 62 frames to 94 frames, 554 images in total.

A.3.2 Secondary Sequences

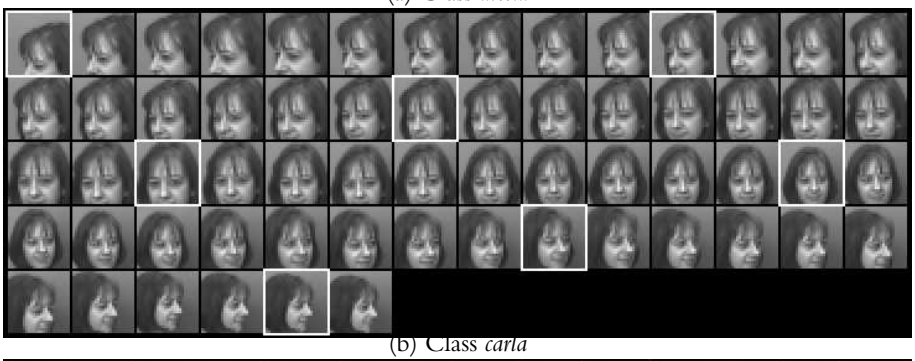
The Secondary image sequences are intended to simulate an on-line source of test images, and are several times longer and much more variable in pose and lighting than the Primary image sequences. Their main purpose is to simulate tracking in an unconstrained environment that could easily be encountered by a real-life application.

They consist of one person moving around a room, allowed to arbitrarily move from side to side and stop and start movement against a cluttered, changing background. This mean that lighting and background detail for a Secondary sequence for a particular person will be radically different to that in the Primary sequence for the same person. Only one preliminary Secondary sequence has been collected so far; this has 169 frames, see Figure A.16. This typical sequence of images from a motion-based head tracker illustrates that perfect registration of the head and face can never be guaranteed, except by manual methods. Future developments of the face detection scheme can be expected to discard any non-face frames, improving recognition, whilst maintaining temporal continuity.

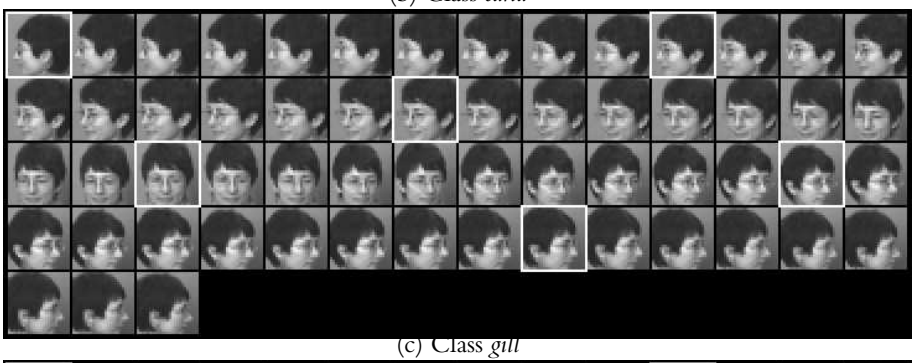
It is also hoped that closer cropping of the faces will be feasible. The image sequences have not been hand-optimised – all data used for this chapter is exactly as output by the QMW processing stages to give an accurately as possible a simulation of an automatic tracking system providing real-life data.



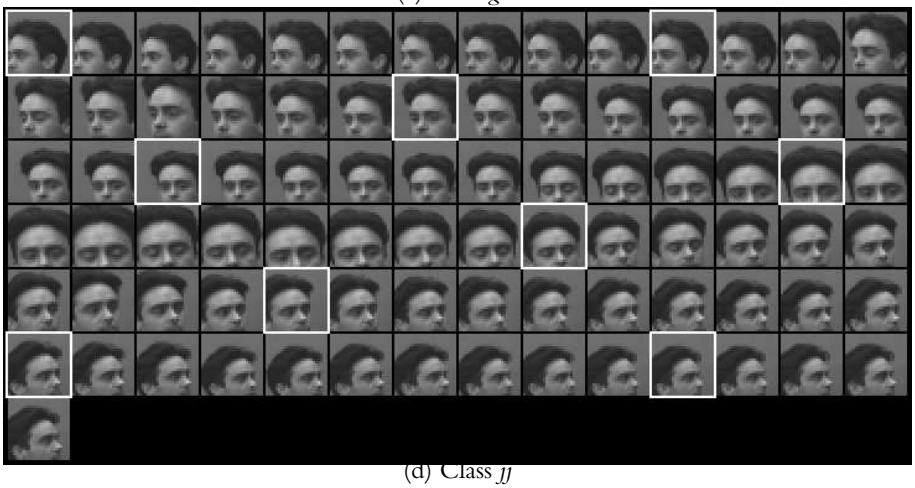
(a) Class *aleka*



(b) Class *carla*

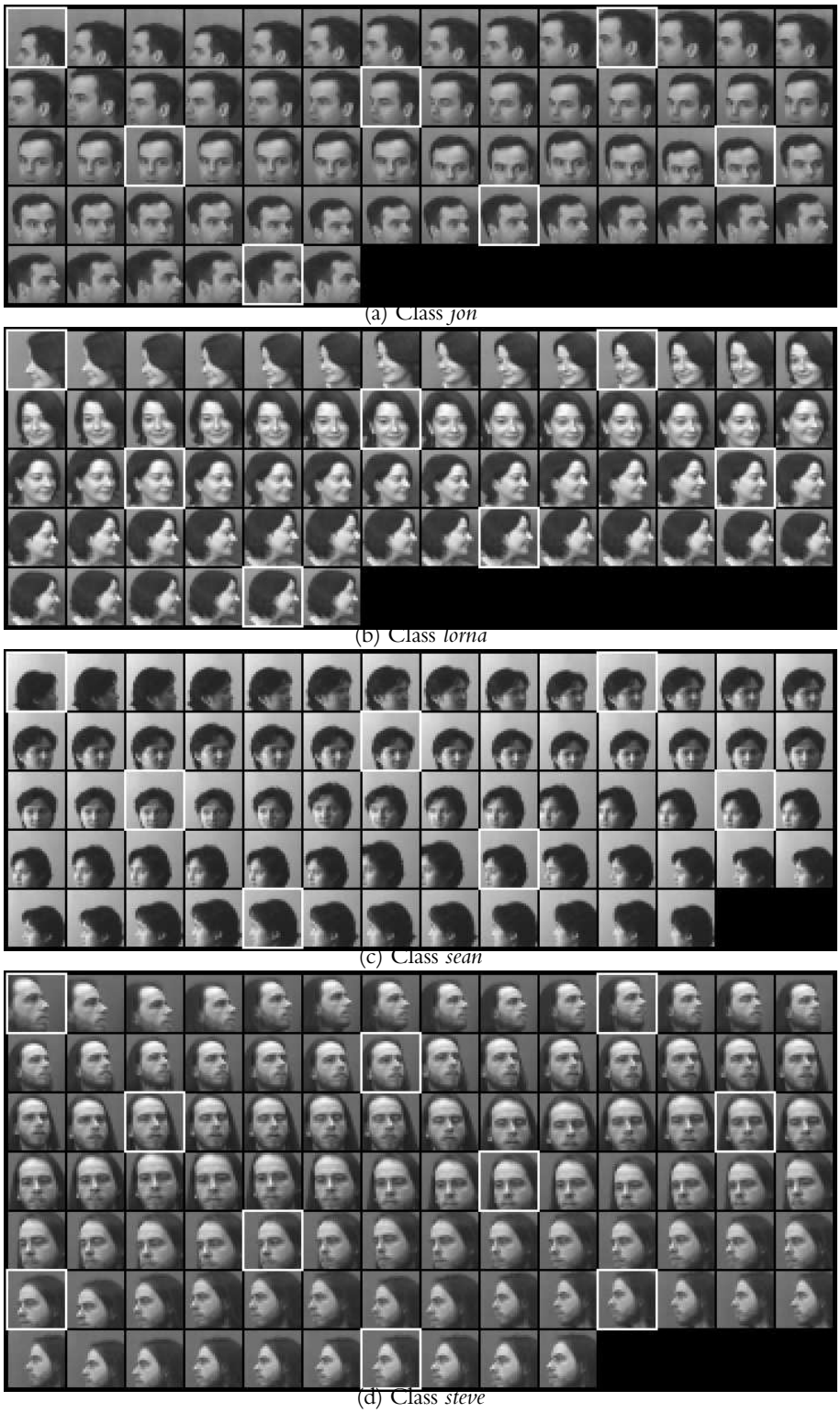


(c) Class *gill*



(d) Class *jj*

Figure A.14: The first four of eight complete QMW 'Primary' image sequences, after segmentation but before preprocessing (boxes indicate frames used for training with a selection interval of 10). Figure A.15 shows the second four sequences.



(a) Class *jon*

(b) Class *lorna*

(c) Class *sean*

(d) Class *steve*

Figure A.15: As Figure A.14, but the second four of the eight QMW sequences.

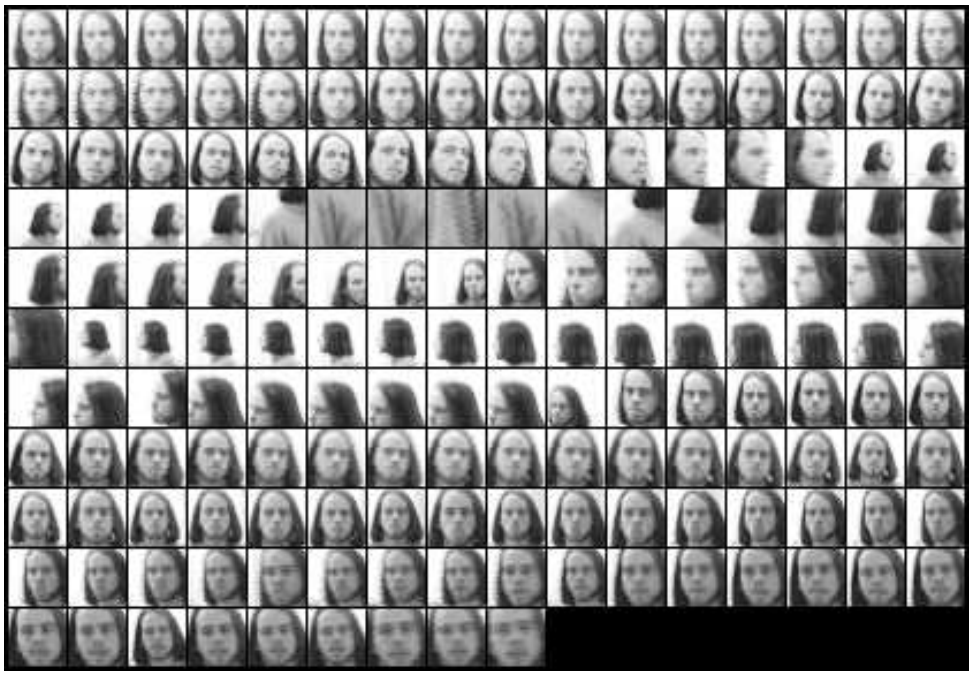


Figure A.16: A complete Secondary sequence for class *steve*, after segmentation but before pre-processing. This shows the high level of lighting and pose variation which was designed to test the RBF network's generalization to conditions different to those used for training. As only front-view face detection has been implemented at this stage, some non-face frames are included and profile views, although segmented, are incorrectly scaled.

Appendix B
Radial Basis Function Network Specification

We employ a Gaussian radial basis function (RBF) neural network model as proposed by Moody and Darken (1988, 1989) (also proposed similarly by Broomhead and Lowe (1988), Poggio and Girosi (1990a, 1990b)). This combines a supervised layer from the hidden to the output units with an unsupervised layer from the input to the hidden units, see Figure B.1. The network model is characterised by individual radial Gaussian functions for each hidden unit, which simulate the effect of overlapping and locally tuned receptive fields.

The size of our RBF network is determined by 1) the size of the face data, as the number of pixels gives the number of input units, 2) the number of training examples, which gives the number of hidden units, and 3) the number of individuals to be distinguished (the number of classes), which gives the number of output units.

The RBF network's success in approximating non-linear multidimensional functions is dependent on sufficient hidden units being used and the suitability of the centres' distribution over the input vector space (Chen et al., 1991). Each training example is assigned a corresponding hidden unit, with the image vector used as its centre, as is common with Regularization Networks (Beymer & Poggio, 1996). This approach should not lead to over-fitting because each image in the dataset contains unique 3-D information. Normalisation of hidden unit outputs means that a 'bias' hidden unit is not required (Musavi et al., 1992).

Table B.1 gives the notation used in equations in this appendix to describe the RBF network. The rest of the appendix is split into two parts, the first describing the training and activation of the hidden unit layer, and the second the output unit layer.

B.1 Unsupervised Learning

The unsupervised part of the training procedure for the RBF network is concerned with determining centre vectors and widths for the hidden units. For this implementation of the network, it is simple to find the centre vectors, as we assign one hidden unit to each training example so that each training vector becomes the corresponding centre vector. Therefore, our main discussion in this section is about methods for determining the width values for the hidden unit Gaussian functions.

B.1.1 Hidden Unit Widths

Each hidden unit (Gaussian basis function) is given an associated width or scale value, σ, which, signifying the standard deviation of the function, defines the nature and scope of the unit's receptive field response (Saha & Keeler, 1990). This gives an activation that is related to the relative proximity of test data to the centre vector associated with the hidden unit, allowing a direct measure of confidence in the output of the network for a particular pattern. In addition, patterns more than slightly different to those trained will produce very low (or no) output.

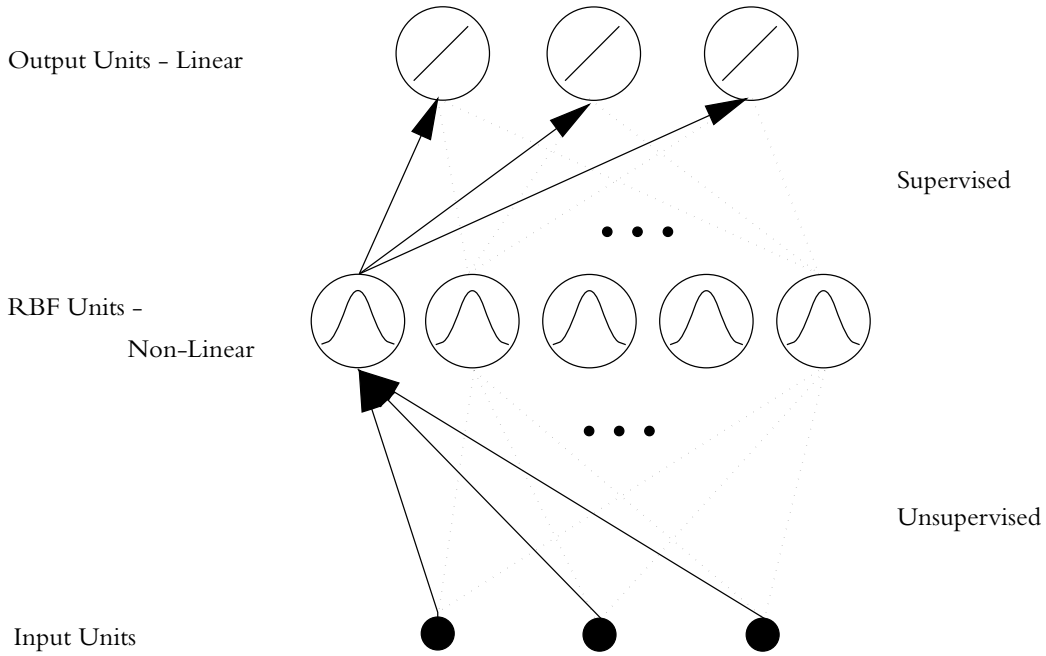


Figure B.1: General layout of a radial basis function (RBF) neural network.

Term	Meaning
L	Number of training patterns
l	Index for pattern
N	Number of input units
j	Index for input unit
$\mathbf{j}$	Input in vector notation
H	Number of hidden units
h	Index for hidden unit
$\mathbf{c}_h$	Centre vector for hidden unit h
ϕ_h^l	Activation of hidden unit h with pattern l
I	Number of classes and output units
i	Index for output unit
t_i^l	Target activation for output unit i with pattern l
o_i^l	Actual activation for output unit i with pattern l

Table B.1: Notation used in equations to describe the RBF network.

Heuristic for σ Values	Initial %	% Discarded	% After Discard
Individual 1-NN	92	44	96
Individual Mean/16	80	2	81
Individual Mean/4	82	22	89
Individual Mean/2	84	48	100
Individual Mean	92	48	100
Individual Mean$\times$2	82	52	100
Individual Mean$\times$4	72	58	95
Individual Mean$\times$16	70	62	94
Global 1-NN	88	50	96
Global Mean	92	54	100

Table B.2: Effect on test generalisation of changing heuristic for calculating σ values for the hidden units for standard 50/50 RBF networks with 'E3' 3-orientation Gabor preprocessing (see Appendix C for details).
'Individual' indicates that each hidden unit has a separate value, whilst 'Global' indicates the mean of all the individual values was used for all hidden units. '1-NN' indicates that the distance to the single closest hidden unit only (the 'nearest neighbour') was used, whereas 'Mean' indicates that the average distance to all other hidden units is used. The Individual Mean results also show where the values were factored by a constant value.

Moody and Darken (1988) used an adaptive P-nearest neighbour heuristic to determine the centre σ values. We calculate individual σ values for each of the H hidden units from the mean Euclidean distance between the centre $\mathbf{c}$ of each hidden unit α and all others h in the global form of the formula from Stokbro et al. (1990, p.606):

$$\sigma_\alpha = \frac{1}{H\sqrt{2}} \sum_h \sqrt{(\mathbf{c}^\alpha - \mathbf{c}^h)^2} \tag{B.1}$$

As mentioned, an alternative would be to use a fixed number P of closest distance values to determine the mean value, but this would require an extra parameter (which would require optimisation). Although the values in Equation B.1 may become closer to the overall mean value for large numbers of hidden units, it was felt this was a better approach than to have such extra parameters, which can interrupt automatic operation.

Generalisation performance for RBF networks is dependent on the appropriate choice of the σ centre width values for the hidden units (Musavi et al., 1992). However, Table B.2 shows that in practice, the precise form of σ calculation was found not to affect generalization greatly (for our particular task, at least), particularly after discard, but the use of individual mean values did give the best results. RBF networks using very small σ values tended to lose performance because of less effective discard (due to hidden units' outputs becoming more polarised), whereas those with larger σ values also lost performance but this time with a high proportion of discarded classifications. This table also shows that the use of a single nearest neighbour gave slightly worse results than obtained with mean values.

B.1.2 Hidden Unit Activations

The (unnormalised) output u for hidden unit h (for a pattern l) uses a Gaussian function, which can be expressed as

$$u_h^l = \exp[-\frac{(r_h^l)^2}{2\sigma_h^2}] \tag{B.2}$$

where, in this case, r is the Euclidean distance:

$$r_h^l = d_E(\mathbf{j}^l, \mathbf{c}_h),$$

$$d_E(\mathbf{j}, \mathbf{c}) = \sqrt{\sum_{x=1}^{N} (j_x - c_x)^2}. \tag{B.3}$$

This is the distance between the N-dimensional input vector $\mathbf{j}$ and hidden unit centre $\mathbf{c}$. Note that the calculation of u_h^l in Equation B.2 does not require the square root of the expression in Equation B.3 to be calculated.

The hidden layer output is then normalised (Moody & Darken, 1989):

$$\phi_h^l = \frac{u_h^l}{\sum_h u_h^l} \tag{B.4}$$

The input-to-hidden weight connections w_{hj} shown in Figure B.1 are fixed and equivalent to the elements of the centre vector $\mathbf{c}_x$ for Equation B.3.

B.2 Supervised Learning

The supervised part of the training procedure for the RBF network is concerned with determining suitable values for the weight connections w_{ih} between the hidden and the output unit layers. The output o for the output unit i for a pattern l is

$$o_i^l = \sum_h w_{ih} \phi_h^l. \tag{B.5}$$

There are two main techniques for determining these weights, which can both be seen as minimising the error measure (cost function) $\mathcal{E}$ of the network

$$\mathcal{E} = \sum_l \mathcal{E}^l = \sum_l \sum_i [t_i^l - o_i^l]^2 \tag{B.6}$$

where t_i^l is the target output value for output unit i when the network is presented with training pattern l. The first is an iterative method using gradient descent, whilst the second is a 'one-shot' method using Singular Value Decomposition.

B.2.1 Gradient Descent Calculation

Weight adjustment for the hidden-to-output layer can be made with the Widrow-Hoff delta learning rule (Widrow & Hoff, 1960), also known as LMS (least mean square) rule. Convergence of the RBF network during training is defined as the point when the error measure for the network (Equation B.6) goes below a pre-determined 'error limit' value, which needs to be established by trial and error. The error δ for output unit i (for a pattern l) is

$$\delta_i^l = t_i^l - o_i^l. \tag{B.7}$$

This is combined with two more fixed parameters which control the speed of change, η, the learning rate, and γ, a momentum term, to give the change in weight value Δw_{ih}

$$\Delta w_{ih}^{l} = \eta \delta_i^l \phi_h^l + \gamma \Delta w_{ih}^{l-1} \tag{B.8}$$

Hertz et al. (1991) suggested initialising the weights w_{ih} to the target output values to speed training, so that $w_{ih} = t_i^l$.

The gradient descent method can be slow, requiring many iterations, and uses several arbitrary parameters – the error limit, learning rate and momentum term. Direct methods of calculation, as discussed below, do not require such parameters and can be calculated quickly.

B.2.2 Pseudo-Inverse Calculation

An alternative method to calculate the weights between the hidden and output layers is to use the matrix pseudo-inverse method (Poggio & Girosi, 1990a), using Singular Value Decomposition (SVD) (Press et al., 1986), which allows an exact solution to finding w_{ih} in a single processing stage.

This process is best explained by rerepresenting the training equations for the network in vector notation. It has already been shown that the error $\mathcal{E}$ of the network in Equation B.6 will be zero when $o_i^l = t_i^l$, that is, the actual output of the network matches the target values. Equation B.5 in this case becomes

$$\sum_h w_{ih} \phi_h^l = t_i^l, \tag{B.9}$$

As Bishop (1995) shows, the values w_{ih} can be estimated from this equation using the singular value decomposition to find the pseudoinverse of the matrix whose elements are ϕ_h^l. This approach is more useful than gradient descent, as it allows almost instantaneous 'training' of the network, regardless of size.

Appendix C
Preprocessing Techniques

This appendix describes the specific implementation of the two preprocessing techniques used: Difference of Gaussians and Gabor wavelet filtering.

Preprocessing of the images is an important intermediate step, as the input representation contributes a great deal to the learn-ability of the task. Highlighting relevant parts of the information (leading to reduction in the dimensionality of input) and providing moderate invariance to normal environmental illumination (Marr & Hildreth, 1980) are important to us here. This is in contrast to tackling strong, incidental lighting, which is very much more difficult (Moses et al., 1994).

C.1 Difference of Gaussians (DoG) Preprocessing

Where there is a change of intensity in an image, peaks or troughs are found in the first derivative of the intensity, and zero-crossings in the second derivative. To isolate the latter, Marr and Hildreth (1980) suggested the $\nabla^2 G$, or *Laplacian of the Gaussian*, operator.

The Laplacian of the Gaussian can be closely approximated by a computationally efficient *Difference of Gaussians* (DoG) operator, which is constructed from two different Gaussians G of the form

$$G(x,y) = \frac{1}{\sigma^2} \exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{C.1}$$

where the space constants σ usually have the ratio of 1:1.6 to give the best approximation to the Laplacian.

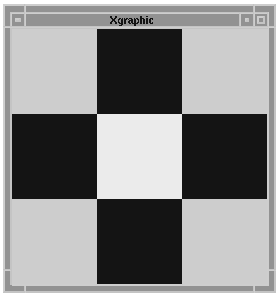
We have constructed DoG filter masks using the POPVISION `convolve_gauss_2d` library routines (further details in Popvision (1994)), which were also used to control the convolution of the image.

Figure C.1 shows the masks produced in this way at varying DoG scales. Figure C.2 shows the result of convolution of these filters with a 25×25 face image from the Sussex database. Because of the low resolution of the image, the number of convolved samples was quite low when the larger DoG filters were used.

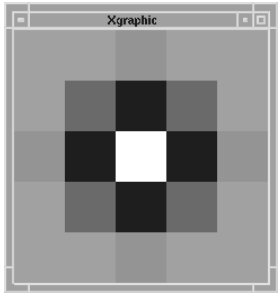
C.1.1 Varying Face Resolution

Table C.1 shows the different DoG scale values used for different image resolutions.

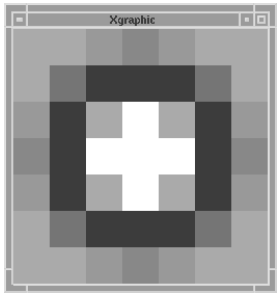
The smallest size of image (6×6) does not have a correspondingly small scale factor, due to the low resolution of the data in general. The scale value cannot be less than 0.15, as this gives a 3×3 filter mask. 3×3 is the smallest mask size there can be, as the mask has to be an odd number of pixels across, and a 1×1 mask would obviously not be much good.



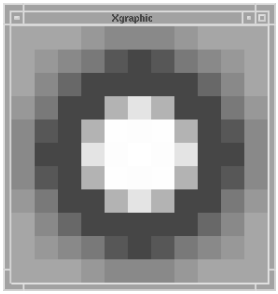
(a) Scale 0.15 (Mask 3×3)



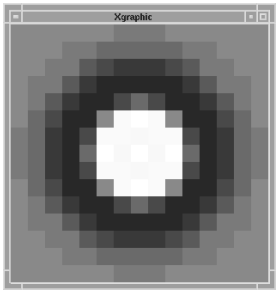
(b) 0.4 (5×5)



(c) 0.8 (7×7)

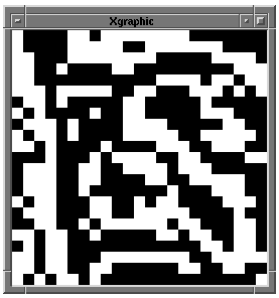


(d) 1.2 (11×11)

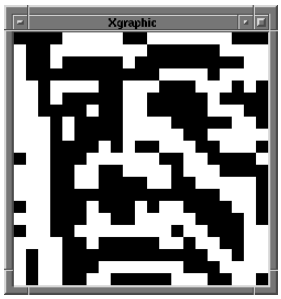


(e) 1.6 (15×15)

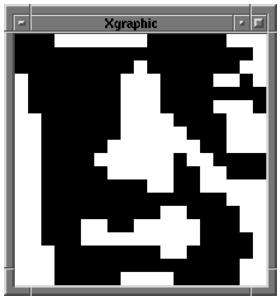
Figure C.1: Filter masks created from a range of DoG scales used for preprocessing.



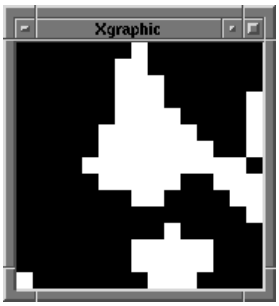
(a) Scale 0.15 (23×23 sampled values)



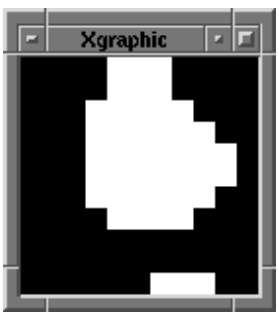
(b) 0.4 (21×21)



(c) 0.8 (19×19)



(e) 1.2 (15×15)

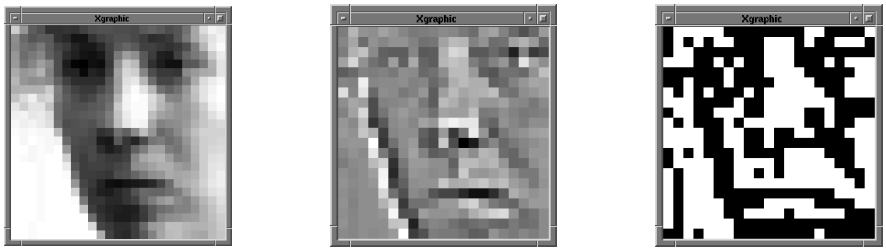


(g) 1.6 (11×11)

Figure C.2: Convolved vales from a 25×25 image using DoG preprocessing at different scales.

Original Resolution	DoG Scale	Convolved Resolution	Samples per Image
100×100	1.3	90×90	8100
50×50	0.8	44×44	1936
25×25	0.4	21×21	441
12×12	0.15	10×10	100
6×6	0.15	4×4	16

Table C.1: Resolutions of face data used from the Sussex database, and the DoG preprocessing values for each image size.



(i) (ii) (iii)

(a) full range of grey-levels

(i) (ii) (iii)

(b) reduced range of grey-levels

Figure C.3: Effect of different ranges of grey-levels for DoG preprocessing using a 25×25 image. (i) before preprocessing (ii) after non-thresholded DoG preprocessing (iii) after thresholded DoG preprocessing.

C.1.2 Image Grey-Level Range

The range of grey-levels present in the images can be reduced, if it is considered that the areas of low and high pixel values are not useful areas of face information (that is, the skin tones will be represented by mid-values). To see how useful this would be, the 8-bit grey-level range of 0–255 present in the images was reduced to 20–65. All values outside this range were set to the minimum or maximum values, effectively removing all detail from these pixels. Such an approach may also reduce the effect of specular reflectance and strong shadows. Figure C.3 shows the effect of such reduction in grey-level range both before and after DoG preprocessing.

C.1.3 DoG Gradients *vs.* Zero-Crossings

With a typical, grey level image, such as Figure C.3(a)(i), DoG convolution will give continuously-valued (with both positive and negative gradient values) gradient information, as shown in Figure C.3(a)(ii). Where these values change from one sign to the other is the 'zero-crossing' point; if the values are thresholded at 0 into either 0 (for negative) and 1 (for positive), the boundaries between black and white are the zero-crossings for the image, as shown in Figure C.3(a)(iii). To test how useful it was to explicitly concentrate only on this boundary point, preprocessing was carried

out with and without this thresholding stage, producing two forms of processed information:

Gradient DoG This is non-binarized.

Zero-crossings DoG This is the binarized form of the information.

It may be possible to remove noise during the binarisation process by ignoring zero-crossings which only have small values on each side, but we have not been able to try this yet.

A third type of information, the zero-crossings contours, as illustrated in Marr (1982, p.69), are the lines following the zero contour around the gradients. Such gradient line formation requires high resolution data, so that the areas on either side of the lines are accurately separated. For the low resolution face data used here, these zero-crossing contour lines were found to be almost the same as the binarized data, and so tests were not done on this type of data. It should also be noted that such bare zero-crossings lines are not as informative as the binarized data, as they have lost the sign of the gradient value. The retention of the areas of positive and negative values in the binarized data makes explicit whether the change-over at the zero crossing is from positive to negative or the reverse.

Moreover, there is evidence that plain line drawings (which look quite similar to zero-crossings) are not a good representation for face recognition (Bruce et al., 1992). The provision of shading information, for example by filling in darker areas with solid black, allows greatly improved recognition. The binarized form of the DoG preprocessed data we use gives a similar effect. Euclidean distance comparisons using such areas of light and dark ought to be more robust than those between pure zero-crossings contour information (which consist of narrow black lines on white) and therefore provide greater test generalization.

To see why this is, one has to imagine the Euclidean distance between two images which are identical except one has been shifted across by one pixel. For a representation of general areas of light and dark, many pixels in both images will still be the same, whereas for the zero-crossings representation (each zero gradient line is one pixel wide), even a shift of one pixel will be sufficient to make most pixels be out of registration and so no longer have the same value when compared. The latter representation will have a very much larger Euclidean distance between the two images than the former, which will make them harder to distinguish from truly different images in a more general classification task, and therefore will not allow good generalization.

C.2 Gabor Filter Preprocessing

We have selected 2-D Gabor filters (Daugman, 1988) as an alternative preprocessing method, as it provides oriented information, which, we hope, will provide input information for the network in a more useful form than the previous methods. One disadvantage of isolated orientation-specific value is that if a full convolution of the image is carried out, more values are output than input (as there is a data value for each pixel for each orientation required). In addition, there are sine and cosine components of the Gabor filter, which doubles the number of coefficients produced.

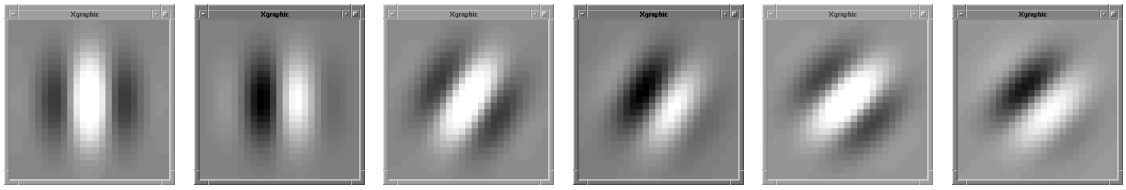
C.2.1 Gabor Scales and Orientations

The Gabor masks were constructed using the POPVISION `gabormask` library routines, using the parameters σ for width and ϕ the orientation of the mask. σ is based on p the period of the harmonic component:

$$\sigma = \frac{p}{2\sqrt{2}}.$$

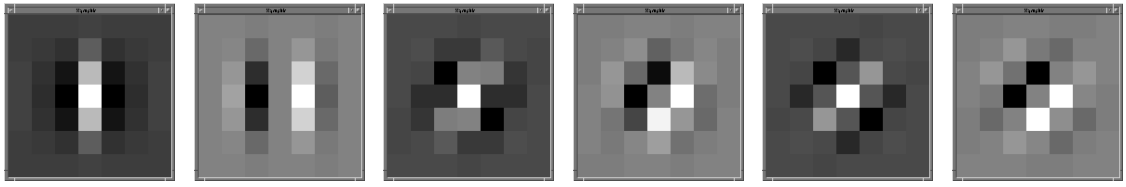
The real (cosine) component, C, of the Gabor mask is calculated as

$$C(x,y) = N\exp(-\frac{x^2+y^2}{2\sigma^2})\ \cos(x'\omega), \tag{C.2}$$



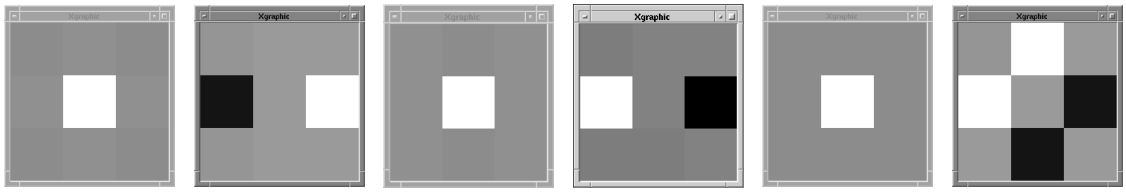
(i) (ii) (iii) (iv) (v) (vi)

(a) 25×25 masks created from a Gabor function of period 13.



(i) (ii) (iii) (iv) (v) (vi)

(b) 7×7 masks created from a Gabor function of period 3.



(i) (ii) (iii) (iv) (v) (vi)

(c) 3×3 masks created from a Gabor function of period 1.

Figure C.4: Gabor filter masks of different orientations, created by Gabor functions of three different oscillation periods.
(i) 0° real (cosine) element (ii) 0° imaginary (sine) element (iii) 30° real (iv) 30° imaginary (v) 45° real (vi) 45° imaginary.

Scheme	Orientations (degrees)	Scales	Over-lapping	Matrix	Coefficients per Image
A1	90	4	No	Square	170
A3	0, 60, 120	4	No	Square	510
B1	90	4	Most	Square	170
B2	0, 90	4	Most	Square	340
B3	0, 60, 120	4	Most	Square	510
B4	0, 45, 90, 135	4	Most	Square	680
B6	0, 30, 60, 90, 120, 150	4	Most	Square	1020
C3	0, 60, 120	4	Less	Square	510
D3	0, 60, 120	4	No	Circular	420
E1	90	3	No	Square	42
E2	0, 90	3	No	Square	84
E3	0, 60, 120	3	No	Square	126

Table C.2: Types of Gabor sampling schemes tested, with filter orientations and number of coefficients sampled per image.

where

$$x' = x\ \cos(\phi) + y\ \sin(\phi),$$

and

$$\omega = \frac{2\pi}{p},$$

and N is a real normalisation constant. The imaginary (sine) component, S, is

$$S(x,y) = N\exp(-\frac{x^2+y^2}{2\sigma^2})\ \sin(x'\omega). \tag{C.3}$$

C.2.2 Gabor Sampling Schemes

In order to reduce the number of coefficients calculated for each image, sparse sampling schemes were constructed, with a range of scales. Each sampling point will have a number of coefficients, one each for the sine and cosine component, multiplied by the number of orientations used. Each scheme is referred to by a letter and an optional number denoting the number of orientations used, for example, 'B3' is the B sampling scheme with 3 orientations.

The 'A' and 'E' square matrix sampling schemes had the least amount of overlap on sampling points. Others were tested which used large amounts of overlap on the sampling receptive fields, or circular sets of sampling points; Table C.2 summarises the different sampling schemes used. Tables C.3(a) and (b) show the sampling arrangements for the 'A' and 'B' square matrix sampling schemes, with Figures C.5(a) and (b) showing how these masks were positioned to cover the image area. Note that the 'A' scheme only covers 24×24 at the 8×8 scale and the some overlap was needed to fit the 2×2 and 4×4 sampling levels.

The 'C' square matrix sampling scheme (Table C.3(c) and Figure C.5(c)) has a midway level of overlap between 'A' and 'B', the scales used intended to retain fine detail from the original image. The 'D' circular matrix sampling scheme was devised in order to reduce the number of coefficients still further and is described in Table C.3(d) and Figure C.5(d).

Number of Samples	Period	Mask Size
1×1	13	25×25
2×2	7	13×13
4×4	3	7×7
8×8	1	3×3

(a) 'A' sampling

Number of Samples	Period	Mask Size
1×1	13	25×25
2×2	10	19×19
4×4	5	11×11
8×8	3	7×7

(b) 'B' sampling

Number of Samples	Period	Mask Size
1×1	13	25×25
2×2	9	17×17
4×4	4	9×9
8×8	2	5×5

(c) 'C' sampling

Number of Samples	Period	Mask Size
7	4	9×9
7	3	7×7
19	2	5×5
37	1	3×3

(d) 'D' sampling

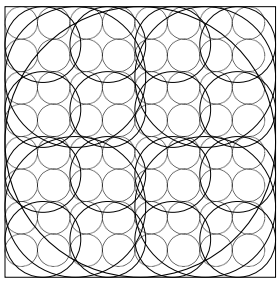
Number of Samples	Period	Mask Size
1×1	13	25×25
2×2	7	13×13
4×4	3	7×7

(e) 'E' sampling

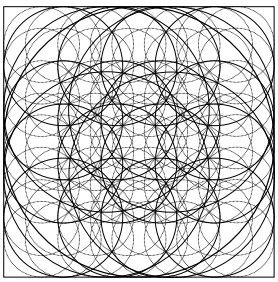
Table C.3: Sampling and filter masks used for different Gabor preprocessing schemes.

Scale Combination	Sampling				Coefficients
	1×1	2×2	4×4	8×8	
1	•				6
2		•			24
21	•	•			30
4			•		96
41	•		•		102
42		•	•		120
421 (E3)	•	•	•		126
8				•	384
81	•			•	390
82		•		•	408
821	•	•		•	414
84			•	•	480
841	•		•	•	486
842		•	•	•	504
8421 (A3)	•	•	•	•	510

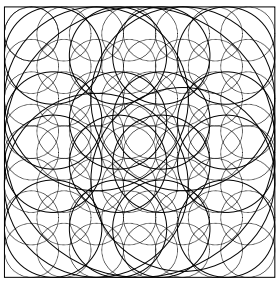
Table C.4: Numbers of coefficients for different A3 Gabor filter scale and sampling combinations: The '8421' arrangement is equivalent to standard A3 sampling, '421' to E3 sampling. See Table C.3(a) for details of filter masks at each sampling level.



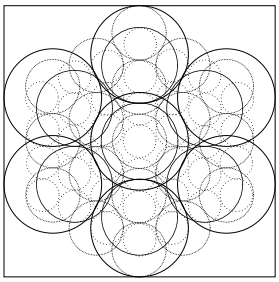
(a) 'A' least overlap, square matrix



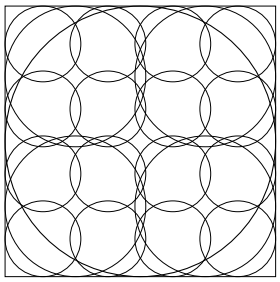
(b) 'B' most overlap, square matrix



(c) 'C' less overlap, square matrix



(d) 'D' least overlap, circular matrix



(a) 'E' least overlap, 3-scale square matrix

Figure C.5: Sampling positions for Gabor sampling schemes.

The fairly coarse alignment to pixel boundaries in the low resolution 25×25 image area means that some actual sampling positions do not coincide with their exact mathematical position.

Table C.4 shows the number of coefficients used in each combination of filter scales. Each number in the combination title refers to one scale, so '1' has only one scale: 1×1, whereas '81' has two: 8×8 and 1×1.