FACE KERNEL EXTRACTION FROM LOCAL FEATURES

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2005

By
Maria Pavlou
School of Electrical Engineering and Electronics

Contents

List of Tables

List of Figures

Abstract

A unified hybrid (local/global) model for face representation and recognition is presented. Non-linear kernel feature representations called *feature kernels* are used at the localised image level to cope with non-linear and complex appearance-based variations of face images. It is shown that feature kernels of internal facial features and facial landmarks are more useful for discrimination than other regions. By taking advantage of kernel combinations, a number of feature kernels is easily combined to forma single representation of face images based on local features. This allows the representation to be compact and hence the learned model is more generalisable over the parameter space of facial variations. For an optimal feature-based representation a progressive feature selection approach is employed to find combinations of feature kernels which optimise a suitable discriminant criterion. The extracted sets of feature kernels combine to form *face kernels* that are representations of face images based on their most salient and discriminant feature. It is shown that the selection of features by the model closely resembles an intuitive selection and can be empirically recognised as being the most salient and relevant for face recognition. A multiresolution representation is also proposed with face kernels extracted from images at varying resolutions combined to form *multiresolution face kernels* (MRFKs). This allows the representation to be flexible and robust by employing detailed and coarse information for learning. Similarly, a hybrid MRFK is proposed which also incorporates a global kernel representation of face images. The MRFK and HMRFK representations are evaluated on a large subset of the FERET database. It is shown that with a suitable selection of feature sizes and image resolutions the face kernel can deliver improved recognition performance over a holistic approach. Error rates of 22% are archived using a MRFK with only 8 features and 20.5% using a HMRFK with one feature. These results are achieved despite the broad scope of variations in the data set and the relatively small number of training examples per identity.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

First and foremost, I would like to thank Our Lord Jesus Christ and The Heavenly Father Who have guided me and given me strength.

To my loving parents and siblings who have not left my side and supported me in my struggle, without them this work would never have been possible.

I would like to thank my supervisor Prof Nigel Allinson for being patient, understanding and supportive throughout the project period and whose kindness and generosity will not be forgotten.

Thanks go to the friends I have come to know along the way. Especially to Melanie Platt who has been like a mother to me, who has always had an open hand and welcoming arms. Her guidance and experiences have been invaluable during my long stay in the UK. To Nicholas Kyele for being a true friend, through good times and bad. To Thalis Anaxagoras who has always been the sweetest person, always fun-loving and caring.

Finally I would also like to thank all the people in the former Vision and Information Processing Group of UMIST, Manchester who I have come to know and who have helped in numerous ways.

Chapter 1

Introduction

Research on biometric methods has gained renewed attention in recent years brought on by an increase in security concerns. The recent world attitude towards terrorism has influenced people and their governments to take action and be more proactive in security issues. This need for security also extends to the need for individuals to protect among other things their working environments, homes, personal possessions and assets. Many biometric techniques have been developed and are being improved with the most successful being applied in everyday law enforcement and security applications. Some biometric methods include:

- DNA Sequence Matching.
- Face Recognition.
- Finger Print Identification.
- Hand Geometry Identification.
- Iris Identification.
- Voice Recognition.
- Signature Recognition.
- Retina Identification.

A common biometric technique being used with great success in forensics and law enforcement is finger printing along with the recent and arguably ultimate biometric method, DNA profiling. Although these methods are unique specifiers or indicators of identity they are not easily applied. DNA profiling requires a sample from which DNA rich tissue can be harvested and then analysed in a laboratory with specialist equipment. Such samples are usually obtained from materials left behind, eg. at the scene of a crime, or by obtaining samples directly from individuals. This applies to finger printing in a similar way and hence the acquisition of these measurable materials is (in some situations) intrusive on the subject. Another intrusive biometric is retina recognition which records the pattern of fine capillary veins in the back of an individual's eye. Confirmation of identity requires a subject to interact with special imaging equipment. Similarly for iris recognition specialised imaging and tracking equipment is needed to locate and capture the iris. Both iris and retina recognition require the voluntary interaction of the subject with the measuring equipment and hence cannot be easily applied in large scenarios. Security applications requiring surveillance of a large number of individuals such as in shopping malls, schools, public places and private property must be non-intrusive. Non-intrusive methods generally requiring no contact or special interaction with individuals and can even be applied without their knowledge.

Face recognition is one such non-intrusive biometric which attempts to assign or confirm identity to individuals by learning and recognising their faces. Face recognition is a natural and straightforward biometric method used by humans to identify each other. Humans are able to detect and identify faces in a scene with little or no effort [15]. This ability has inspired much thought and research into understanding the process and development of Automated Face Recognition (AFR) systems. With the rapid development of technology in the 1990s, the application of face recognition in the commercial world was made possible. Now many applications use face recognition [117] including controlling access to secure areas and verifying the identity on a passport. Many approaches to face recognition have been developed for application in real world scenarios. A collection of these technologies, from academic institutes and private companies, were evaluated between September 1996 and March 1997 with the FERET database [95, 101]. The evaluation strategy employed images of 1196 subjects under varying conditions such as illumination and pose. Although the AFR systems which were tested did perform well, the failure of some systems (which were later commercially applied) to perform as advertised showed that further research and development were still needed. Face recognition is still a subject of great interest in the research literature. With recent advances in machine learning techniques and increasingly powerful computers the development of AFR systems is improving. Ongoing research in psychological and psychovisual analysis of face recognition in humans has further guided and fueled developments in AFR systems. These factors have also inspired and facilitated research into areas surrounding face recognition such as facial expression recognition [5, 103], gesture recognition and human-computer interaction.

1.1 Face Recognition

As mentioned in Section 1, faces are a unique identifier of a person and face recognition has its own advantages over other biometrics. One of the greatest advantages is the non-intrusive nature of face recognition requiring no voluntary interaction by the subject. This means that face images can be easily obtained with good quality, not always easy with finger prints which may be degraded due to (among other things) smudging, cut skin or dry skin. With the development of cheaper and higher quality CMOS imaging technologies any number of facial images is easily acquired from a range of environments suitable for surveillance and security purposes. However the multitude of sources for face images under varying conditions makes the location and recognition of faces in images a complex problem. These variations differ as broadly as the sources from which the images were obtained [39]. Factors such as visibility and hence image quality and resolution play a large role in the introduction of noise. Variation in illumination conditions is another key factor in facial image acquisition. As all images are a product of a Lambrertian reflection of light on an object, changes in the type and direction of the illuminating source greatly affects the object's appearance. In-depth orientation (pose) changes of the subject are also very important variations along with expression changes, aging, partial occlusions and general changes of the subject's appearance (disguises, beard, accessories, etc.).

The development of a complete AFR system is therefore not an easy task which, considering all the possible variations, must incorporate robust systems for the (1) acquisition, (2) location, (3) normalisation and (4) recognition of faces. In this work it is assumed that points (1) and (2) are handled by a separate process leaving only the face recognition task. This is common practice in the research

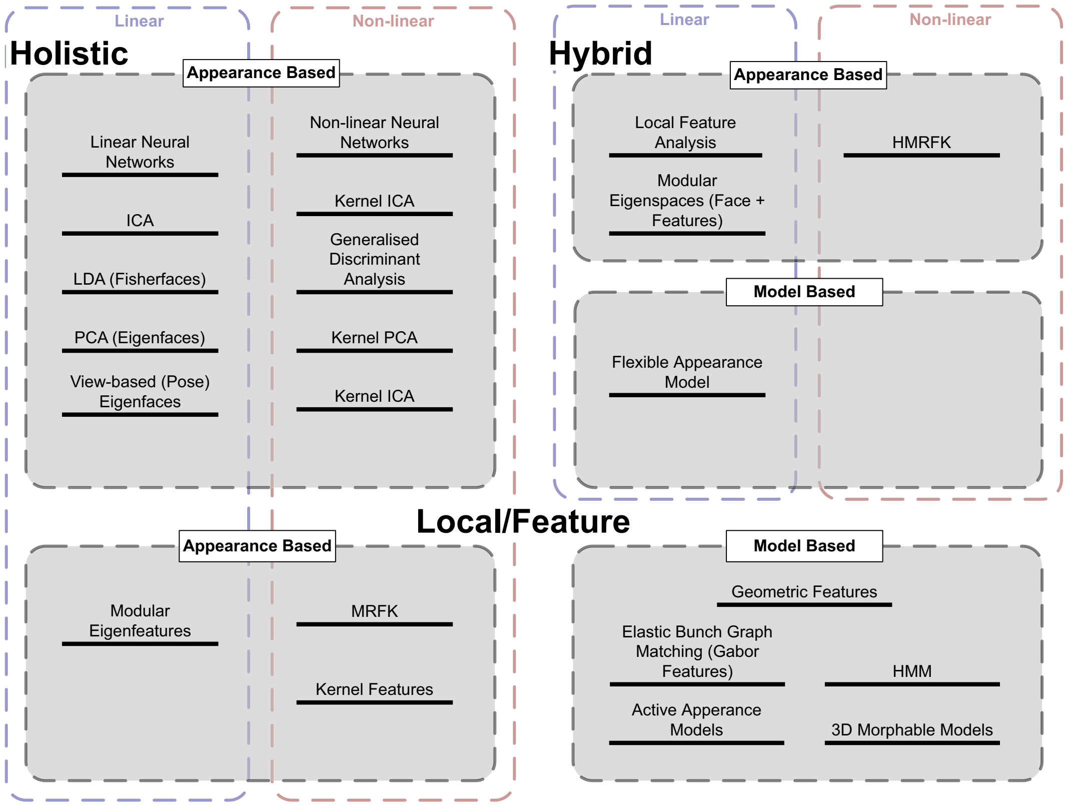


Figure 1.1: Structured outline of methods employed in face recognition.

literature where previously acquired face images are used for development and benchmarking. Normalisation of face orientation and illumination is applied as a preprocessing stage and is discussed in further detail in Chapter 5. After face images have been standardised processing or learning can be applied. The way faces are represented for processing can be divided into two main groups, feature-based and holistic approaches. The local or feature-based approach operates on local structures or image features. Facial landmarks such as the eyes, nose and mouth and non-structural features such as birthmarks are examples. Properties and relations of these features can then be used to construct a representation of faces. The holistic approach treats the raw face data as a whole and attempts to extract a suitable representation which can be used for classification. A further division can be made within each group using a model-based approach and an appearance-based approach. The former uses structural coding of the face, ie. relative positions of facial features. The latter however employs no high level structural coding of the face except that which is inherent in the structure of the image pixel data. A number of other groups can be defined in the methods of face representation, and Figure 1.1 gives a basic outline of these approaches.

Holistic appearance-based approaches are generally related to template matching and have been widely and successfully adopted in face recognition. Methods such as the eigenfaces approach [114] extract features in a face subspace using principal component analysis (PCA) on the image data. The face space is spanned by a number of orthogonal eigenfaces [110] which can be used to represent faces in a compact and low dimensional manner. Appearance-based approaches however are greatly affected by the variations of facial images such as face orientation and size and lighting conditions. Furthermore linear representations such as eigenfaces fail to capture useful intrinsic[1] variations within the face class in the presence of extrinsic variations such as pose changes and illumination. In these circumstances feature-based methods offer a better representation of the face class and its typical variations.

Model-based methods which combine local image statistics and facial structure representations can also sufficiently model the variability of face images. One such approach uses local Gabor wavelet (feature) representations placed in a mesh or at specific landmarks over a set of face images. The feature representations and their geometrical relations form a global structural graph known as a Bunch Graph. *Elastic Bunch Graph Matching* (EBGM) [119] allows the feature graph a degree of flexibility where pose and expression changes can be incorporated into the face representation. Systems such as these produce compact representations and can accommodate variations in viewpoint and, with the use of Gabor features, illumination. However the feature locations used are generally manually placed over facial landmarks or previously supplied by a feature location pre-process. This means that errors (human error) can be passed on when incorrect feature locations are provided. More interestingly feature placements made manually may not be data-driven. This is the case even for some feature location methods which are tuned to locating a set of intuitively chosen prototypes. These prototypes are typically templates of the eyes, nose, mouth and other landmark facial features. The intuitive selection or placement of such features does not necessarily constitute an optimal selection in terms of the image data. An alternative approach learns a feature selection directly from the data and forms a solution for the required task. However, the performance of such an approach can be hindered by other factors. The most important factors

[1]Intrinsic variations - variations exhibited between objects in the same class.

Table 1.1: Terms typically used in face recognition.

Holistic approach	The whole face region is used as input for learning and classification.
Feature-based approach	Features based on local geometry, local statistics and location are used for learning and classification.
Hybrid Approach	Combination of whole face and local feature information for learning and classification.
Appearance-based approach	Learning and classification with face representations based solely on image pixel data.
Model-based approach	Structural framework constructed (trained) from face information (local and/or global) and used for classification.
Single-view approach	Learning and classification from only one view of appearance or from multiple views transformed to the canonical view. This approach is typically related to in-depth orientation changes of faces.
Multiple-view	Learning of and classification with templates (local and/or global) based on multiple views of appearance. When an unseen example is to be recognised it must be matched against all the stored (learned) view templates.
Linear model	Processing of face data and structures (especially in appearance-based methods) assumed of containing linear relations. Typically used for dimensionality reduction.
Nonlinear model	Processing of face data and structures which have been projected into some nonlinear feature space. It is assumed that nonlinear relations in the data become linear relations in the feature space.

being the quality and quantity of the available learning data. In this section the face recognition problem has been introduced. Some approaches commonly used in processing and describing face images have also been briefly discussed. The following section briefly expands on some of these approaches and outlines some desirable properties of an AFR system.

1.2 Motivation

Current non-linear application methods of face recognition based on recent developments in kernel machines have shown great success in separating the intrinsic and extrinsic variations in face images. As these variations are complex [6, 89, 107, 109], basic linear descriptions such as PCA fail to capture the non-linear nature of the variations [1, 107]. This is evident from the inferior performance of such methods as compared to non-linear appearance based approaches

[107, 120, 121] and complex feature based approaches [53, 78]. Non-linear approaches such as Support Vector Machines (SVM)[115] and Kernel PCA (KPCA) [105], therefore, may be better suited to the task of face recognition. The use of these non-linear techniques has predominantly been applied in appearance-based systems which use whole face images in a holistic approach. Therefor only holistic non-linear variations of appearance are learned in these cases and there is a lack in representing any facial structure.

Non-linear localised feature model: Hence a model is needed which unifies non-linear local feature representations with facial structure.

Such a feature based approach will also require a selection of localised features for the representation model. A distinction is now made between image features and facial features. Many existing techniques use localised image features, while some others use localised facial features or landmarks. While both might correspond to the same image region, they differ in certain respects. A hand-crafted model relies on *a priori* knowledge of the task at hand. However, a model based on a predefined set of localised facial features (e.g. nose or mouth) will neglect information about features that are specific to a particular person, such as birthmarks. The model is constructed and not learned, therefore it may satisfactorily process the analytical descriptions of variations but it may not be applicable to other variations. Care must be taken, however, if a system uses localised image features and does not incorporate information about corresponding facial features, since its performance may quickly degrade.

Data-driven salient features: A data-driven model is therefore needed which is based solely on the data statistics. A set of salient local features and the feature structure must be learned directly from the data.

This allows all possible image variations to be addressed within a single framework. Other typical image variations such as pose changes have traditionally been approached in two ways. From a model structure standpoint, the view-based approach attempts to handle different instances of variation with separate representations or learning units. For example, learning a representation model of faces at a certain angle of pose. A number of these models (corresponding to other pose angles) can then be pooled together to produce a complete representation. One such system uses pose eigenspaces to create a set of pose subspaces learned using only discrete variations of pose in a set of face images. Each pose subspace however requires a separate classifier. This is typical of such view-based models and complicated processing schemes (eg. majority voting and weighting schemes) are required to combine the multiple classifiers. A further consequence of such an approach is the lack of generalisation. A rigid model only covers a parameter space of variations discreetly and cannot be easily applied to variations not described in the model. The parametric model approach however learns all variations using a single representation model. This means the model smoothly and continuously covers the parameter space and furthermore requires only one classifier.

Generalised object-based model: An appearance-based multiple-view model which is compact, generalisable and easily applied to the data.

Typically appearance-based approaches fall into this category where a single subspace is constructed describing all the variations present in the data.

1.3 Thesis Statement and Contributions

Considering the three points made in the previous section a flexible face representation model is proposed. In order to capture the complex nature of variations in face images, non-linear kernel discriminant analysis is applied to localised regions of face images. Corresponding local regions from an ensemble of face images are used to construct a *kernel matrix*. A kernel matrix or Gram matrix can informally be regarded as a pairwise similarity matrix between all pairs of samples in an ensemble. In this work the kernel matrix of a localised region (feature) is henceforth called a *feature kernel*. Feature kernels are constructed from a number of localised regions across the face area, each being used to apply kernel discriminant analysis (KDA or GDA). From KDA a feature subspace is obtained which maximises the between-class and minimises the within-class distance of representations from the ensemble based on the feature region. To measure how well each feature subspace achieves this task a discriminant criterion is employed. The criterion becomes an indicator of which localised (feature) regions are most discriminant and hence which feature regions may be most suitable to represent a face. By exploiting a property of the kernel, feature kernels can be combined to form new feature kernels. Any number of feature kernels can be combined to produce a new feature kernel which ideally will improve the discriminant and hence descriptive capacity of its feature subspace. In this manner combinations of feature kernels are found that maximise the discriminant criterion. Hence combinations of varying numbers of feature kernels form representations based on features from varying localised regions over the face area. These new combined feature kernels are henceforth called *face kernels*. Construction of face kernels therefore automatically selects optimal feature placements based on the data. Furthermore the new face kernel inherently incorporates feature structure, only a specific combination of feature kernels produces the optimal face kernel.

To enhance the representation a multiresolution approach is proposed whereby face kernels are extracted from the face data at varying resolutions. As the dimensions of the kernel matrix are based on the number of samples in the ensemble data and not the image resolution, face kernels from varying resolutions can easily be combined. These combinations form *multiresolution face kernels* or MRFKs. This approach can be extended further with kernels being applied to the whole image area at varying resolutions. By combining these holistic kernels with the feature-based face kernels a new *hybrid multiresolution face kernel* or HMRFK is proposed. Contributions of the proposed face models are now discussed in light of the requirements presented in Section 1.2 above:

Non-linear discriminant feature representations. Non-linear analysis is applied locally to face image regions. A kernel matrix is constructed from the image region of all samples in the ensemble. As face images in the ensemble are normalised with respect to orientation, the kernel matrices therefore represent feature regions common to all samples. *Feature kernels* are therefore data-driven representations of localised image features.

Learn optimal feature placements. Commonly used feature node placements are relaxed by learning optimal placements. Feature placements are found by learning optimal combinations of salient feature kernels to form *face kernels*. This process is again data-driven with no *a priori* knowledge of face structure being supplied.

Structural framework based on features. A representation of an ensemble of faces is achieved using combinations of *feature kernels* to construct *face kernels*. The *face kernel* inherently incorporates local feature representations and facial structure based on the specific set of *feature kernels* with which it is constructed.

Unified multi-observer representation. A *feature kernel* can be considered as an observer of the face ensemble. Each observer is based on a different image region or feature. The *face kernel* unifies the multi-observer representations into a single representation. A *face kernel* is used to construct one non-linear discriminant subspace requiring only one classifier.

Unified multi-resolution representation. *Face kernels* from varying resolutions can also be considered as multi-observers which can also be easily combined. *Multiresolution face kernels* therefore unify representations of multiple features at varying resolutions.

The kernel matrix therefore affords flexibility in representation and structure. Kernel matrices can be based on a number of variational characteristic of interest. For example it would be possible to construct appearance-based *pose kernels* of faces at varying pose. A combined kernel could then incorporate a number of discrete poses into a generalisable representation. This approach is however not explored here and is left for future work. This work takes advantage of the kernel matrix and its non-linear representation to form a flexible face model. The aim of this work is twofold. Firstly to apply the face kernel model to the task of face recognition on a data set of faces with multiple variations and to measure its performance. Secondly a comparison is made between typical salient features selected by the model and typical features deemed as being salient for face recognition in humans. The aim of this is to assign a degree of biological plausibility to the face kernel model which learns a successful face representation without affording it any *a priori* knowledge of the task.

1.4 Thesis Outline

In this chapter the motivation and basic outline of the work presented in this thesis are given. In Chapter 2, the most relevant literature including the state-of-the-art systems of interest is reviewed. Open problems are discussed and solutions are proposed with a description of the face kernel model. Chapter 3 presents a brief overview of the non-linear kernel methods to be applied. The structure of the *feature kernel* is developed and briefly evaluated as a 'single-view' parametric face recognition approach. In Chapter 4, the 'single-view' approach is extended to develop the *face kernel*. Issues concerning the resolution, size and number of *feature kernels* employed in a representation are explored. Additionally a method for the selection of *feature kernels* is presented and evaluated. Finally the hybrid face kernel structure is introduced. In Chapter 5, a new criterion function is presented for the extraction of face kernels with the aim of improving processing performance while preserving the quality of feature selections. The *face kernel* models are applied in a face recognition task. Finally recognition performance is explored and compared to similar AFR systems using a number of different model structures and parameters. Chapter 6 concludes by summarising the contributions of this work and presenting ideas for future work.

Chapter 2

Face Recognition

From the time soon after we are born and our eyes can begin focusing, we start looking for the faces around us. In fact Pascalis, et al. [92] show that our parents' faces are the first objects we begin to learn and recognise. As we grow this skill is further perfected until it becomes second nature and we can recognise faces seen in the worst of conditions and over long distances. Human recognition processes are remarkably efficient at recognising faces utilising visual, audio and contextual information to name a few. However good the brain is at remembering and recognising faces it can only do so for a relatively small number of identities. Understanding how the brain performs this skilled task and thereby performing it using modern computers, which potentially can learn a large numbers of faces, is a goal which has inspired 30 years of research.

Since the early 1950s researchers in psychology [18], and since the 1960s in the engineering literature [12], have been trying to understand and develop face recognition processes. Both groups of researchers have benefited from their shared interest, one by guiding the development of machine face recognition systems based on biological and psychological observations, and the other by developing tools and methods for making those observations. In the next section some of the research in psychophysics and neuroscience surrounding face recognition is briefly reviewed. The following sections give a review of the face recognition literature, and conclude with a brief discussion of a number of recent studies that have some resemblances with the work of this research.

2.1 Psychological Research

One of the main issues in the literature is whether or not face recognition in the human visual system (HVS) is a dedicated process. Does the HVS have separate processes for recognising faces and non-face objects? According to Ellis [30] there are indications that face recognition might be a dedicated process. One such indication is the inability of Prosopagnosia patients to recognise familiar faces. These patients compensate by recognising people they are familiar with by using other cues such as voice, hair colour, dress and accessories. Interestingly however these patients can recognise if some object is a face. This suggests that there may be a process for recognising various objects inclusive of faces but that there may also be a separate process which "appreciates" the face object and can therefore perform identity recognition. Recent work by Gauthier and Logothetis [35] and Gauthier, et al. [36] suggest that face recognition may not be a dedicated process. The boundary between visually processing faces as objects with its constituent parts and visually "appreciating" faces for identity is blurred and much of the literature is contradictory. Work by Pascalis, et al. [92] indicates that neonatal babies prefer to look at and follow faces. This might suggest that infants are born with the basic ability to recognise faces. They also note that infants have limited vision, seeing only with low spatial resolution, and that they can only initially recognise the upper parts of the face and head. In tests the infants failed to follow or look at images of faces whose foreheads and hairline had been covered. However as the infants grew they began to recognise faces even when the occlusions were present. This might give strength to the suggestion that face processing is a holistic process at first, becoming more feature specific as familiarisation increases.

The question of whether face recognition is a holistic analysis of the face or whether it uses more specific feature analysis has also been researched in the literature. Bruce [14] and Bruce, et al. [16] suggest that the HVS may use global or holistic representations as a precursor or pre-process for finer feature recognition. However when conspicuous and obvious features deviating from the norm are present, the holistic representation seems to be of less significance. Recollection tests showed that faces with these distinctive features were more memorable and that the HVS might be focusing on such features. One interesting experiment by Thompson [113] known as the 'Thatcher Illusion' uses an upright face and turns its features upside-down producing an unpleasant image not particularly

Figure 2.1: The Thatcher Illusion: Inversion of features in an inverted face (top right) seems normal, whereas inversion of features of an upright face (bottom right) seems unpleasant.

identifiable as a face (Figure 2.1 bottom row). Inversion of the Thatcher-ised face however appeared fairly normal (Bartlett and Searcy [106], Figure 2.1 top row). Inversion of faces has been used in many studies of the importance of configural information in upright face processing. Farah, et al. [32] showed that faces which were inverted were harder to recognise (also Hochberg and Galper [44], Rhodes, et al. [100]) and that this was more evident than for other objects [124]. Work by Young, et al. [125], Rhodes, et al. [100], Kemp, et al. [49] and Searcy and Bartlett [106] among others demonstrate that encoding of upright faces relies on the relational information (or configuration) of facial features and that inversion impairs this encoding more than the encoding of other types of information such as isolated features.

2.2 Holistic Face Recognition

Holistic face recognition methods have been successfully applied in the research literature. The majority of face recognition methods which have been developed are based on a holistic approach. The holistic method makes no distinction of facial landmarks or features but considers the whole face as one object to be processed and recognised. Additionally most holistic methods fall into the category of appearance-based approaches where images of varying facial appearance are processed. The aim is to extract salient and compact descriptions of faces of individuals while ignoring redundant information. In general it has been shown that significant redundancies exist in natural images [102] and when such images are limited to a class of objects such as faces the redundancy is even greater[1] [93, 131]. Holistic processing can be disadvantageous, however, as faces can vary markedly in expression change and rotation in depth causing occlusions and varied appearances. In fact Moses, et al. [1, 86] have shown that variations due to expression, rotation and lighting are greater than variations between faces of different subjects. Holistic face recognition methods must therefore be able to accommodate such variation in order to be successful.

One of the first steps taken towards holistic face recognition by Sirovich and Kirby [52, 110] involves the representation of faces using low dimensional PCA projections. Principal component analysis is used to extract salient descriptions or features from an ensemble of faces. These features or low dimensional bases describe the modes of greatest variance in the face data. This is achieved by finding the eigenvectors of the covariance matrix constructed from the face data. The number N of images used is typically much less than the dimensionality of the image data and so the covariance matrix is degenerate. Only the first N eigenvectors are found and ordered in descending value of the respective eigenvalues. Any image in the face data can now be approximately reconstructed by using a weighted combination of the eigenvectors also known in this case as eigenfaces. Eigenfaces with high eigenvalues account for most of the variance in the data and so contribute more to the approximation of any face, and it is possible to achieve a good approximation with relatively few high order eigenfaces. This supports the view that images of the same class contain large redundancies. At this point eigenfaces were primarily used for the representation of faces until the technology was pushed towards application by Turk and Pentland [114] who used eigenfaces for detection and identification. Once the eigenfaces for a data set are found every face can be represented by a vector of weights obtained by projecting the face images into the eigenface bases. Any test image can then be projected into the eigenspace to obtain its own weight vector. A possible match between the test image and any trained image is achieved by locating the trained image whose weight vector is the closest or most similar using some metric (usually Euclidian distance). Impressive recognition results are reported using a large data set with

[1] This is true for images of a class which are normalised with respect to scale, translation and rotation.

variations in pose and lighting. However with variations in scale performance was low and this was accounted for by Akamatsu, et al. [3]. Their work shows that proper alignment and scaling of faces is important for good performance and robustness. It has also been shown that irrelevant facial portions such as the hair, neck, shoulder and background often provide misleading information to a face recognition system [24]. With the application of feature alignment, normalisation and cropping, these preprocessing steps greatly improve recognition performance and give an early indication into the importance of local features. It is not difficult to conclude that redundancy (within the same class of images [102]) is further reduced when significant facial features are aligned. In the case of eigenface analysis when facial features are aligned and scaled their comparison (analysis) is more intuitive. Eigenface analysis is considered as being an appearance-based approach using no structural framework for the location of features. However the preprocessing applied to face images for normalisation can be thought of as a structural framework imposed on the data and in effect pixel-wise local feature analysis is carried out. The variation of features and appearance of faces can still be great even when normalised and especially between images of the same subject [1, 86].

To enhance classification Moghaddam and Pentland [84] use a probabilistic subspace method modeled on intra-personal and extra-personal variations in face images. The extrapersonal subspace is believed to account for representative variations, such as those captured by the standard Eigenfaces method, while the intrapersonal subspace accounts for variations due mostly to changes in expression. An equivalent approach which learns the variations of faces and other objects within their individual classes was used by Swets and Weng [112]. Linear Discriminant Analysis (LDA) is applied to a set of face and non-face images where intra-class and inter-class scatter matrices are constructed. A set of Most Discriminating Features (MDF) are then found to construct a feature space which minimises the distance between representations of objects of the same class and maximises that between different classes. This approach was applied exclusively to faces in Belhumeur, et al. [10] and called the Fisherface method. Good results are reported and show that the Fisherfaces method outperformed that of Eigenfaces. Improvements to the basic LDA approach were proposed by Zhao [129] and Zhao, et al. [130] with a regularised combination of PCA and LDA subspaces. The use of weighted distance metrics to improve performance was also reported by Zhao [129], Etemad and Chellappa [31] and Liu and Wechsler [66]. To date a number of other improvements to LDA have also been proposed [23, 73, 126] which deal with the *small sample size* SSS problem [75]. Such techniques referred to as Direct LDA (D-LDA) take advantage of the null space of the within-class scatter matrix commonly discarded by previous methods. This null space is believed to contain significant discriminatory information and is used to enhance classification performance.

Other fundamentally different approaches have also been applied in the holistic recognition category and will be briefly mentioned. Liu and Wechsler [67] develop an adaptive representation based on Evolutionary Pursuit (EP). EP seeks to learn an optimal basis for data compression and pattern classification. This is achieved using strategies similar to *Genetic Algorithms* (GAs) where evolution of the learning system is driven by a combination of performance accuracy (classification ability, empirical risk) and class separation (discriminating ability, confidence interval). Good results are reported compared to the eigenfaces [114] and Fisherfaces [10] methods. Lin, et al. [65] propose a fully automatic face detection and recognition system based on a probabilistic decision-based neural network (PDBNN). Two holistic feature representations are used from the upper portion of the face: normalised intensity and edge features. Each feature is fed to a PDBNN which devotes one of its subnets to the representation of a particular class or person. Classification results are obtained by fusion of the outputs of each PDBNN. The PDBBN's one-class-one-network architecture is beneficial as is converges to a solution faster than other neural network systems and has greater generalisation ability.

Other types of feature representations are also possible. For example Chien and Wu [25] use a multiresolution wavelet transform to extract *waveletfaces*. LDA is then employed to learn a discriminant *waveletface* representation. Gabor wavelets and discreet cosine transforms have also been used in face recognition as holistic feature representations. Finally, *independent component analysis* (ICA) has been proposed by Lee [59] which can extract high-order statistics from input images. It is well known that analysis techniques like PCA only extract representational features based on first and second order statistics such as pair-wise relationships between pixels. However, for the complex variations in face images [107] Independent Component Analysis (ICA) has been employed to extract important discrimination information where PCA may be inadequate. Use of ICA for face recognition have been proposed by Bartlett, et al. [7, 8] and Liu and Wechsler [68]; with both set of researchers reporting superior results over PCA-based methods. However, Moghaddam [83] and Shakhnarovich [107] showed that an ICA-based approach did not provide significant advantages over PCA-based methods. This disagreement may be explained by the difference in the experimental setup and the choice of similarity measure. ICA performed better in the former approaches using more difficult time-separated images from the FERET database. Aditionally the latter approaches used face images, which were tightly cropped, leaving no information regarding hair and face shape and were lower in resolution. These factors when combined made the recognition task much harder. In [107], PCA and ICA had an equal performance of $\sim 77\%$ correct recognition rate; however ICA exhibited the largest performance variation. Their experimental results suggest that seeking non-Gaussian and independent components may not necessarily provide an enhanced representation for face recognition.

Large variations in face appearance due to varying pose, illumination and expression typically lead to highly complex and non-linear distributions of face images [107]. This is demonstrated in Chapter 3. As a result linear appearance-based systems such as PCA and LDA do not perform well in such circumstances [74, 120, 121]. One way of dealing with this problem is by employing a "divide and conquer" approach. In this way a complex non-linear problem can be broken down into a number of smaller tasks tackled by linear systems [107]. A mixture model is then constructed by the combination of these linear systems. The choice of decomposition methods can be based on the type of variations within the data. In face recognition the view-based approach can be used where images of similar appearance and variation are clustered together. For example, Pentland and Moghaddam [94] divide the pose variational face space into smaller sets of fixed pose angles. Each set of faces at a specific pose are then represented using a separate pose eigenspace. To classify a test face its pose is first determined after which its projection in the relevant pose-eigenspace is compared to stored examples. However, this approach is rigid and only covers the variation space discreetly and cannot be easily applied to variations not described in the model. An alternative way of dealing with large variations employs generalised non-linear learning techniques.

Recently advances in kernel methods have propelled the development of many non-linear algorithms for machine learning and pattern recognition [115]. These kernel-based learning methods project a complex problem into a higher-dimensional feature space via a non-linear mapping function. The aim is to linearise the complex problem in the new feature space where traditional linear methods can then be employed. Examples of these systems include Support Vector Machines (SVM) [115] and the nonlinear formulations of kernel PCA (KPCA) [105], and kernel LDA also knows as kernel discriminant analysis (KDA) [80] or generalised discriminant analysis (GDA) [9]. Application proceeds similarly to their linear counterparts, an appearance-based method where the whole face image is used. Li, et al. [62] and Kim, et al. [50] use KPCA for face recognition while Li, et al. [64] and Liu, et al. [70] applied KDA to learn a face subspace for pose variation. These techniques have been applied to face recognition and have shown some success, however not always better than the linear methods [43, 61, 122]. A number of enhancements to KDA have also been proposed. Dealing with the SSS problem, Lu, et al. [74] propose Kernel Direct Discriminant Analysis (KDDA) the non-linear formulation of D-LDA and Liu, et al. [71] propose NLDA and NKFDA. To enhance their previous work Liu, et al. [69] propose a new kernel function, called the Cosine kernel, and employ a new feature selection method and feature classifier. Recently Yang, et al. [120, 121] proposed a KPCA plus LDA approach which applies discriminant analysis on the KPCA features. They report good results on a medium to large data set containing pose and illumination variations.

The basic concept of kernel methods and how they are applied to face data is described in Chapter 3. Kernel methods have proven to be useful for describing the non-linear distribution of appearance-based variations in face images. However they are still based on a holistic approach and suffer in circumstances where gross variations occur such as near profile poses and occlusions. In such cases feature-based methods have proven to be more robust. Often *a priori* information of the face is used in a structural framework in conjunction with local features to build a face model. As mentioned in Section 2.1 there is also psychophysical evidence that suggests structure is important in face recognition. The following section gives a brief review of such local feature-based and structural methods.

2.3 Local Feature-based Recognition

A local representation describes a face by a set of local features and is preferred to a global representation because of its robustness to partial occlusions and its flexibility to non-rigid shape transformation. Some of the earliest work in feature-based face recognition was presented by Kanade [47] and Kelly [48]. Their work falls under the category of structural matching: from the geometry of local features of the face a set of sixteen facial parameters are measured. The normalised ratio of distances, areas, and angles between features are used to compensate for the varying size of faces in the image. Faces of varying people could then be checked for similarity against a reference set by using a simple distance metric. The short comings of this method lie in the nature of the features used. Due to variations in facial features and face rotations in depth, distance measurements between features are not always straightforward. Also, human error is present as these measurements are obtained manually. Finally, the features obtained only describe shape information and do not make use of the appearance of the face and its local features, ie. texture information.

It may be beneficial to use shape and texture information as was shown by Craw, et al. [28]. In their work a number of facial points are used to extract a facial texture region. The faces are then warped to fit a reference face shape which produces a shape-free face. They showed that the performance of a PCA-based face recognition system similar to the eigenface system [114] was the best when a face was represented by both shape and shape-free texture PCs. Similar work by Lanitis, et al. [58] uses an automatic system for locating facial landmarks (Active Shape Model [26]) based on local gray-level profiles. Again a shape-free texture area is obtained by warping. A face is then represented by the concatenation of PCs of shape, shape-free texture and local gray-level profiles. This representation was applied to face tracking, pose estimation and face recognition. In both of the above cases, shape-free information considers the whole face area and as such builds a holistic shape-free PC model. Local information is only incorporated separately in the shape information and local gray-level profiles. Hence no correlation exists between the variations in shape and texture. The model assumes there are only basic linear variations like in-plane translation and rotation. However, with large pose variations shape and texture information may not be independent.

On way of modeling these variations is to use localised texture information in conjunction with the shape information. Although shape information is useful in face representation it is better suited to describing geometric variations such as in-depth and in-plane rotations. In contrast, texture information seems to capture more identity specific information as shown by Craw, et al. [28]. To capture local information a representation of the local texture is often used. These representations are usually obtained by applying some transform to the localised image area. Examples are Gabor wavelet, discrete cosine, and multi-resolution wavelet transforms. Although it is possible to use the localised raw image texture, these transform representations offer some desirable properties such as insensitivity to illumination and rotation variations.

Recent examples of the use of Gabor features include work by Yang, et al. [123] and Zhang, et al. [127]. They convolve a Gabor wavelet of varying scale and orientation with fontal face images of a data set. Similarly to Moghaddam and Pentland [84] a two class problem is constructed by finding the difference between intra-personal and extra-personal pairs of corresponding scale and orientation feature representations. A selection of local Gabor features is then learned to form a set of classifiers for recognition. Instead of applying a feature transform to every pixel in a face image (eg. convolution with Gabor wavelets) it is also possible to divide the image into a number of smaller blocks or regions. This is usually done to reduce the correlation (redundancy) between adjacent subimages. A feature transform can then be applied to these regions and then used to learn a representation for recognition. In [90, 91] Pan and Bolouri divide face images into 16x16 blocks and apply a discreet cosine transform (DCT) to each subimage. A selection of DCT coefficients from the subimages is then used to train a back-propagation neural network. Zhao and Grigat [128] employ a similar approach and compare three representation and learning methods. A number of subimage division schemes including the use of the whole image are employed to learn and construct representations for face recognition. Three feature representations are extracted for each subimage namely, maximum likelihood, LDA and DCT features. The subimage features from each division scheme are then concatenated to form a single feature. Classification proceeds via a sum rule fusion of classifiers from each division scheme. Gottumukkal and Asari [41] also use a similar image subdivision. Although the above methods incorporate local texture information, a clumsy division of the image into local regions may not be optimal. Certain sub-blocks may contain no or little salient information and may only describe portions of important facial features.

A more intuitive approach would be to extract local regions placed over facial landmarks such as the eyes. In [60], Li, et al. extract a set of 8 overlapping regions of the eyes, nose, mouth, forehead and chin as well as a tightly cropped whole inner face region. A classifier is then learned from different weighted combinations of feature representations for each region. This approach can also be called a modular approach where views of similar class-specific features are each described by their own model and classifier. This approach was first applied by Pentland, et al. [94] who represented a face by a set of local PCA models (eigenfeatures) located on four distinctive facial regions (i.e. eyes, nose and mouth). These eigenfeatures were also used for the detection of features and performed better than the eigenfaces in the recognition task.

Another well known approach combines the use of shape (structure) and local feature information into a multiple-view approach. Okada, et al. [88] and Wiskott, et al. [119] employ Gabor wavelet features at facial landmarks. A flexible face structure or elastic bunch graph (based on the dynamic link architecture (DLA), Buhmann, et al. [19], Lades, et al. [56]) is combined with the local Gabor features description onto one system. Elastic Bunch Graph Matching (EBGM) of new unseen face images proceeds by allowing flexible localised matching of known features to the test features. In this way the face graph can deform to suit variations in shape and local texture profiles [54, 78]. An identity match is then found by comparing the similarity of the localised Gabor features and graph structure of the the test face with a collection of known examples. The EBGM methods has proven to be very effective for face recognition and was one of the top performers in the FERET evaluation tests [88].

The studies mentioned above have used some form of local information in their systems. These local regions are either rigidly chosen or placed over facial landmarks. To some extent the selection of these regions is based on an intuitive knowledge of facial structure and the importance of facial features. However these features may not always be optimal in building a face representation for the task of recognition. It may be better to learn an optimal selection of features directly from the data rather than using a manual selection.

In [37], Gökberk, et al. extract a rectangular lattice of Gabor features from a collection of images and learn feature weights based on a pose estimation and recognition task. A two stage process is then employed, first pose estimation via the selected pose sensitive Gabor features into one of three pose views. Secondly face recognition proceeds by using the identity discriminant Gabor features for the selected pose view. They conclude that in frontal images, horizontally located grid points in the upper half of the image are more important for recognition whereas in half and full profile images, a vertical axis dominates. This grid arrangement of features is a sufficient yet crude approach for the task of pose estimation and face recognition, and is not very accurate in terms of revealing which facial features may be most important.

To investigate the discriminative power of Gabor features (nodes) in a face graph, Krüger [53] learned the weights of nodes based on face discrimination. The results indicate that the eyes are more important for discrimination of half profiles and frontal views compared to the mouth and chin. Similarly in [54], Krüger, et al. use a learning algorithm to find a suitable subgraph which only contains those nodes important for locating the head and identifying its pose. Kalocsaia, et al. [46] also apply a statistical analysis to learn weights for the contribution of Gabor feature locations, frequencies and orientations to discrimination ability. Their results show that the hairline area with the forehead and eye regions provide useful information while the mouth, nose, cheek and lower part of the outline region are the least useful part of a face for face recognition. A similar work by Grudin [42] found that the eyes and mouth are more stable for recognition, whereas the hair and nose regions have large variations and may be unreliable.

A different approach to extracting and selecting local features is the biologically inspired approach of Penev and Atick [93]. In their work *local feature analysis* (LFA) is used to extract local features from global PCA bases. These features are located at every position in the image and resemble localised filters whose outputs are decorrelated as much as possible. The best set of topographic filters for a face is then selected based on the reconstruction error. It was shown that features from the eyes nose and mouth were of importance. However, as their facial images were not tightly cropped the majority of important features were located around the facial edges.

This problem was also evident in the other studies mentioned above and it has been shown that only internal facial areas should be used in such statistical approaches [24]. Another point of interest in the feature based approach regards the relative importance or ranking of the selected features. This issue has been dealt with in the majority of studies mentioned above by learning a set of weights corresponding to each feature. These weights typically assign some level of importance or relevance to the individual features. The value of the weights assigned to each feature have commonly been based on two strategies. The first method measures the performance of a feature for the task at hand (in this case, classification error) and assigns higher valued weights to those features performing best. The combination of the best performing features is then assumed to give the best performance of the combined system. A second approach assigns higher weights to weakly performing features (called boosting [34]), and finds a summed combination of these features to maximise performance.

These strategies are commonly known as classifier fusion schemes and typically only employ knowledge about the discrimination ability of individualised features. The alternative approach is to consider the discrimination ability of a combination of features into one feature template. A number of methods can be applied for the search of optimal feature combinations and are explored further in Section 4.4.1. Furthermore a recent example employing such a feature selection technique is explored in Section 2.4 along with other feature based approaches which are relevant to this work.

2.4 Related Work

A number of recent studies are now highlighted which have some resemblances with the work presented here. Kim and Kim [51] take into account prior knowledge about facial structures and apply local correlation kernels in an SVM architecture. The knowledge employed here is that faces are composed of several characteristic structures which can be effectively used to form representations of these structures and their relationships between faces. Typically an approach employing the whole image area in the SVM kernel builds a representation based on the pixel-wise correlations between images and hence the faces contained within. In their approach representations are formed based on correlations between facial structures, which are defined as the correlations between the pixels constituting these structures. This is achieved by using local regions of the face in the kernel representations and the final representation is obtained by summing up the local correlation kernels. In this way correlations of between and within class pairs of images are used to train a set of one-per-class SVMs which are combined to a two-layer neural network classifier. Good results are reported, although a small data set is used and regions of external facial features were included in the representation which can provide misleading results [24].

As briefly described in Section 1.3 the work presented in this thesis also takes advantage of local correlation kernels. The above example considers only correlations between any two image pairs, ie. the two-class problem, as a consequence of the SVM approach being used. Alternatively the proposed model considers the multi-class problem allowing for a single representation model to be learned. This in turn is beneficial, as all variations and samples are described by one model and only a single classifier is needed to identify a test sample. An open problem not approached in the above example regards the selection of local correlation kernels. Their approach employs a sliding window at a certain sampling interval. Similar to using a rectangular lattice of Gabor features their approach results in the inclusion of areas which may not necessarily be of any benefit. A better approach may be to select a set of local kernels based around the facial landmarks as in a face graph. However this assumes some prior knowledge and the selection of local kernels may not be optimal in terms of the data.

A step towards extracting features from the data is adopted by Wallraven, et al. [116] who use local correlation kernels and SVMs for object recognition. A standard Harris-type corner detector is used to detect a number of interest points which are represented by a set Gabor features. The interest points extracted from each image are, however, not structured. One image may contain interest points not present in another image and vice versa. This requires a matching step for each image pair comparison in turn, each feature in one image is matched with a feature in the other based on a maximum kernel response and summed to produce a final kernel response. Good results are reported on three data sets. Their feature based approach performed best on the face data and this was attributed to the large amount of discriminative features in the images compared to the other data sets. Although the feature detector used in this case was shown to have high repeatability and robust performance [104] not all the features chosen may be useful. A corner detector may select a number of features, some similar to features chosen by hand and others which may be less intuitive. Additionally the feature detector is a separate unit and not influenced by the SVM representation being learned. An alternative approach is to learn a model representation and feature selection simultaneously and directly from the data.

One recent approach which explores this idea in a similar way is proposed by

Gokberk, et al. [38]. A feature selection algorithm is employed in this case to find a selection of Gabor features for a recognition task. Three selection strategies are employed. The first strategy measures the importance of a Gabor feature based on its individual classification performance. A combination of the top performing features can then be used to construct a single feature for classification. This assumes that each feature contributes independently to the discrimination performance and does not hinder the performance of other features in the combined representation. A second strategy considerers the relative information gained by sequentially combining a new feature with existing ones in a set. The third strategy builds on the later and employs a backtracking step to remove the least useful features from an existing feature subset to overcome the nesting effect. These strategies are discussed in more detail in Section 4.4.1. In term of their classification performance it is found that strategies employing sequential selections and backtracking steps perform much better than a hand picked set or lattice of features.

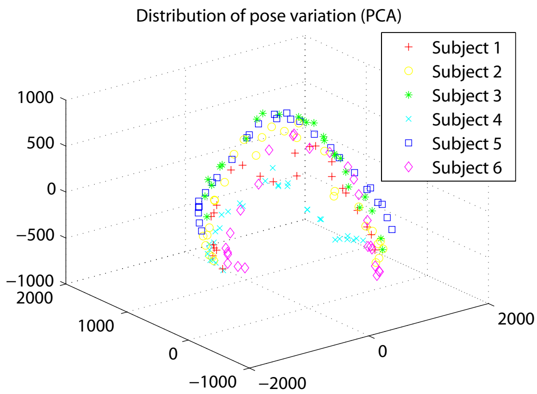
A similar data-driven selection approach using local kernel is applied in the proposed face kernel model in Section 4.4.

Chapter 3

Feature Kernels

Kernel analysis has become increasingly popular in the recent literature and has been applied to face recognition. Subspace methods like Principal Component Analysis (PCA) by Turk and Pentland [114] and Moghaddam and Pentland [84] and Fisher Linear Discriminant (FLD) by Belhumeur, et al. [10] have successfully been applied to face recognition with good results. However, due to their linear nature they can at best only describe pattern information based on second order statistics, i.e. pixel-wise covariance among pixels. Changes in face images due to viewpoint, illumination and expression variations are highly nonlinear and complex [63, 107]. It has been observed by Moses, et al. [86] that variations between face images of the same person due to illumination and viewing direction are greater than image variations between different persons. In order to capture these variations and successfully classify them, statistical methods using higher order statistics such as the relationships between neighbouring pixels are needed. Consider Figure 3.1 which shows the distribution of a set of six pose varying faces in PCA space. This distribution is representative of the high-dimensional vectors space obtained when using images of faces under different viewing conditions. The geometry of this pose varying face-space is curved with distinct regions defined by the scatter of the individuals. In [11] it was shown that multiple views of a face form non-convex (yet connected) regions in PCA space. This can be seen in Figure 3.1(b), which shows only the three largest principal components and is distinctly non-linear. With the inclusion of more principal directions and other modes of variation such as illumination and expression the resulting distribution describes a highly complex and nonseparable manifold. As a result linear models such as PCA and LDA can easily fail to perform well when large variations are

(a)



(b)

Figure 3.1: (a) A selection of images of two subjects with pose variation and (b) the distribution on the first three principal components in a PCA subspace for six subjects.

present in the data [63].

The 'kernel trick' was first used by Aizerman, et al. [2] and Boser, et al. [13] to represent the input data to a system as nonlinear features. The input space is mapped by means of a nonlinear function, expressible as dot products, to a convenient feature space in which the input data is nonlinearly related. This kernel trick remained largely unused in the machine learning community, except within Support Vector Machines (Vapnik [115]), until Schölkopf, et al. [105] developed Kernel Principal Component Analysis (KPCA). It was shown that KPCA is able to extract nonlinear features and therefore provide better recognition performance over PCA. However KPCA, just like PCA, extracts features along the direction of greatest variance across all the training samples and so can retain unwanted variations due to lighting and expression change. KPCA, although being non-linear, is like PCA in that it is optimal for reconstruction from low-dimensional bases and not necessarily optimal for discrimination purposes (Belhumeur, et al. [10]). The application of PCA and LDA to face recognition is initially reviewed in this chapter followed by the development of non-linear methods. The non-linear approach of KPCA and the kernel discriminant variants KDA, GDA and KFD are described. In the concluding sections principles are introduced for the construction of a new face recognition model.

3.1 Kernel Analysis

The most obvious way someone may think of to compare images of different faces would be by using template matching with a similarity metric like correlation. Template matching, however, can become very unrealistic as image size increases causing computation time to increase. Consider a set of data $X = \{x_k\} \quad \forall k = 1, \ldots, M$ where $x_k \in \mathbb{R}^N$. When dealing with large data sets and large dimensionality N, feature extraction methods are used to determine an appropriate subspace of dimension m from the original feature space where $m \leq N$. One such feature extraction and dimensionality reduction method is PCA. PCA seeks to minimise the mean square error criterion by constructing a low-dimensional orthogonal space from the largest m eigenvectors of the $N \times N$ dimensional covariance matrix of M N-dimensional data. Let X represent a data matrix of size $N \times M$, the covariance matrix S_T is,

$$S_T = (X - \mu)(X - \mu)^T \tag{3.1}$$

where μ is the mean of the data matrix. Eigenvectors V and eigenvalues Λ are then found satisfying,

$$S_T V = \Lambda V \tag{3.2}$$

The linear transformation from the original N-dimensional input space to the m-dimensional reduced feature space is then achieved using $\tilde{V}$ of dimension $N \times m$,

$$Y = \tilde{V}^T X \tag{3.3}$$

The matrix Y $(m \times M)$ now contains the M m-dimensional transformed patterns. In face recognition the column vectors of V represent feature vectors that best account for the distribution of the faces in X. V is called the face subspace where faces in X are projected to obtain the transformed data vectors Y. Essentially the transformed data vectors are weights that describe the contribution of each feature vector in the face subspace. The weight vector y of a test face x,

$$y = \tilde{V}^T x \tag{3.4}$$

can then be compared using a suitable metric, commonly a distance metric, to check for similarity with any of the transformed data vectors in Y. PCA is an unsupervised analysis method and considers the input data X as a whole even though samples within the training data X may belong in different classes. In contrast, LDA exploits the within and between-class information of the data to construct a more descriptive face subspace. A set of labels is usually required for each sample in X defining its class or identity ownership and hence LDA is a supervised method of analysis. The Fisher criterion [33] is then used to construct the discriminant subspace and is defined as the ratio of the between-class scatter over the within-class scatter,

$$J_{Fisher} = \frac{S_B}{S_W} \tag{3.5}$$

where,

$$S_B = \sum_{i=1}^{C} c_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.6}$$

and

$$S_W = \sum_{i=1}^{C} \sum_{j=1}^{c_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T. \tag{3.7}$$

Here μ_i is the mean of each of C classes where each class has c_i samples. Similarly x_{ij} is the jth sample in X belonging to class i. Fisher linear discriminant (FLD) analysis now seeks to find a transform V such that,

$$V_{FLD} = \arg\max_V \frac{|V^T S_B V|}{|V^T S_W V|} \tag{3.8}$$

of the eigensystem

$$S_B V = \Lambda S_W V. \tag{3.9}$$

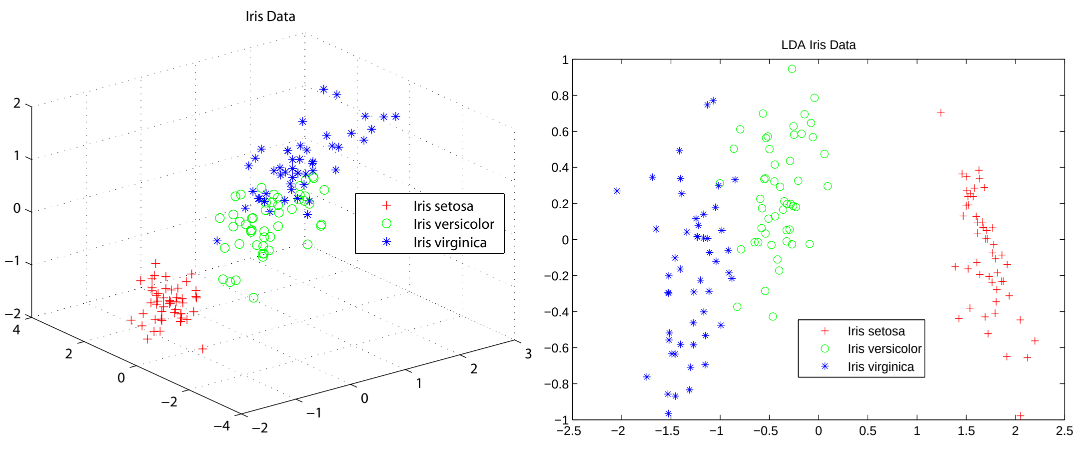


Figure 3.2: Scatter plot of (a) the normalised Iris data (showing only three of four dimensions) and (b) its projection on the first two discriminant directions using the LDA method.

As with PCA (3.3), data vectors can then be projected into the FLD subspace to obtain the feature weight vectors to be used for classification.

Consider Figure 3.2(a) showing the first three dimensions of the Iris data set and (b) its projection onto the discriminant feature space. Fisher's Iris data [33] gives sepal and petal dimensions for three different species of iris, Setosa, Versicolor, and Virginica. It can be seen that FLD (LDA) performs well in separating the classes in two dimensions of the original four dimensional data set. With the use of the within-class information, LDA methods perform better at recognition tasks than PCA [77]. This is, however, not true if there are insufficient samples per class (Small Sample Size problem or SSS) and as shown by Martinez and Kak [77], switching from PCA to LDA is not always useful. They fail however to tackle the problem of defining a minimum number of samples per class to obtain good results with LDA. Intuitively having as many as possible samples per class above a minimum of 3 is reasonable based on their unstable result using 2 samples per class. Despite having sufficient samples, LDA (and PCA) are still linear analysis methods and can at best only describe relationships between face images based on second order statistics. Although LDA attempts to account for variations within the sample classes (in face recognition) these variations are mostly larger than any variation between individual classes ([86]). If the relationship of pixel data from the facial features is considered across changes between facial expressions and pose it can be seen that a linear system may not be suitable. The face manifold across all real world variations is complex and hence a nonlinear representation would be more suitable [63, 74, 107].

3.1.1 Kernel PCA And Kernel Discriminant Analysis

Here a brief formulation of the KPCA method is described (for a detailed description, Schölkopf, et al. [105]), followed by the formulation for kernel discriminant analysis (KDA). Kernel analysis is generally achieved by projecting some data x_k, $k = 1, \ldots, M \quad x_k \in \mathbb{R}^N$ from the input space to a higher dimensional feature space F using a non-linear mapping function Φ

$$\begin{aligned} \Phi : \mathbb{R}^N &\rightarrow F \\ x &\mapsto z. \end{aligned} \tag{3.10}$$

The covariance matrix in the feature space is then expressed as,

$$S_T^{\Phi} = \frac{1}{M} \sum_{j=1}^{M} \Phi(x_j)\Phi(x_j)^T, \tag{3.11}$$

whose eigenvalues and eigenvectors must satisfy,

$$\lambda V = S_T^{\Phi} V. \tag{3.12}$$

Since the solutions V lie in the span of $\Phi(x_i), \ldots, \Phi(x_m)$, equation (3.12) can be formulated as,

$$\lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot S_T^{\Phi} V) \quad \forall k = 1, \ldots, M, \tag{3.13}$$

and by defining an $M \times M$ kernel matrix K as,

$$K = (k(x_i, x_j))_{i,j} = (\Phi(x_i) \cdot \Phi(x_j))_{i,j} \quad \forall i, j = 1, \ldots, M \tag{3.14}$$

equation (3.14) can be expressed as

$$M\lambda a = Ka. \tag{3.15}$$

Therefore K is diagonalised to obtain the eigenvectors V and the projection of a test sample x with image $\Phi(x)$ is

$$q = (V \cdot \Phi(x)) = \sum_{i=1}^{M} a_i k(x_i, x). \tag{3.16}$$

As with LDA, KDA attempts to maximise the between-class scatter and minimise the within-class scatter of samples in feature space. Defining the between-class scatter of the input data in feature space as S_B^Φ

$$S_B^\Phi = \frac{1}{M} \sum_{i=1}^{C} c_i(\bar{\phi}_i - \bar{\phi})(\bar{\phi}_i - \bar{\phi})^T \tag{3.17}$$

where $\bar{\phi}_i$ is the mean in feature space for class i of a total of C classes and c_i the number of samples in that class. $\bar{\phi}$ is the mean of all M samples.

$$\bar{\phi}_i = \frac{1}{c_i} \sum_{j=1}^{c_i} \phi(x_{ij}). \tag{3.18}$$

The within class inertia is defined as

$$S_W^\Phi = \frac{1}{M} \sum_{i=1}^{C} \sum_{j=1}^{c_i} (\phi(x_{ij}) - \bar{\phi}_i)(\phi(x_{ij}) - \bar{\phi}_i)^T \tag{3.19}$$

To maximise the discriminant ratio of S_B^Φ / S_W^Φ the eigenvalues and eigenvectors must satisfy the follwing expression,

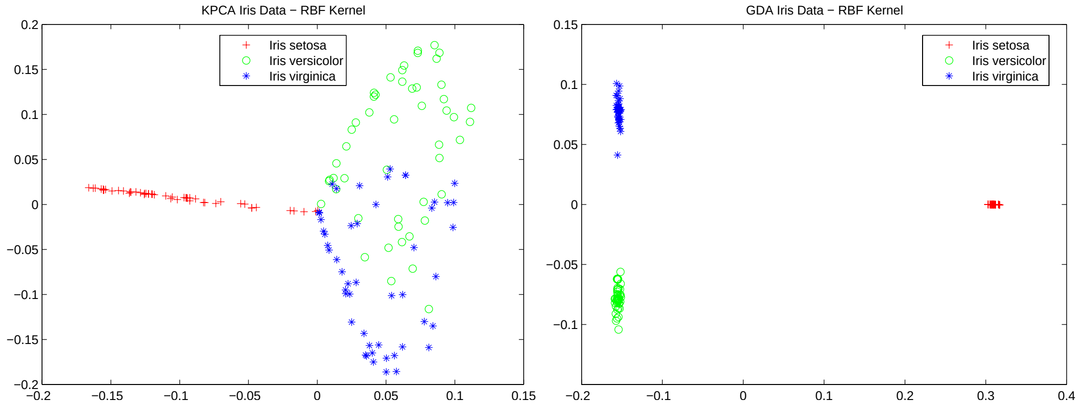
$$\lambda S_W^\Phi V = S_B^\Phi V. \tag{3.20}$$

As with (3.13), equation (3.20) can be formulated as

$$\lambda \Phi(x_k) S_W^\Phi V = \Phi(x_k) S_B^\Phi V \tag{3.21}$$

leading to,

$$\lambda K K a = K Z K a. \tag{3.22}$$



(a) (b)

Figure 3.3: Projection of the Iris data on the first two significant features of (a) KPCA and (b) GDA using an RBF kernel with $\sigma = 0.5916$.

Then eigen-analysis is used to find the eigenvectors a and eigenvalues λ maximising

$$\lambda = \frac{a^T KZKa}{a^T KKa} \tag{3.23}$$

The projection of a test sample x with image $\Phi(x)$ is,

$$q = (V \cdot \Phi(x)) = \sum_{i=1}^{C} \sum_{j=1}^{c_i} a_{ij} k(x_{ij}, x). \tag{3.24}$$

Figure 3.3 shows the IRIS data projected onto the first two directions of the non-linear (a) KPCA and (b) GDA subspaces. It can be seen that although KPCA does separate the data there is a small degree of overlap. However projecting the data onto more of the non-linear principal directions will achieve a better separation. This is not required with GDA which can successfully separate the three classes into tight clusters (maximising the between and within-class scatter) with only two discriminant directions of the GDA subspace. Similarly with the face images, it can be seen in Figure 3.4 that GDA performs better at describing the pose variation in the data as opposed to LDA and KPCA.

3.1.2 KDA and The SSS Problem

As noted in Section 3.1 discriminant analysis methods encounter problems when dealing with the intra-class scatter matrix S_W. This matrix is almost always singular in face recognition tasks where the number of samples per class is small

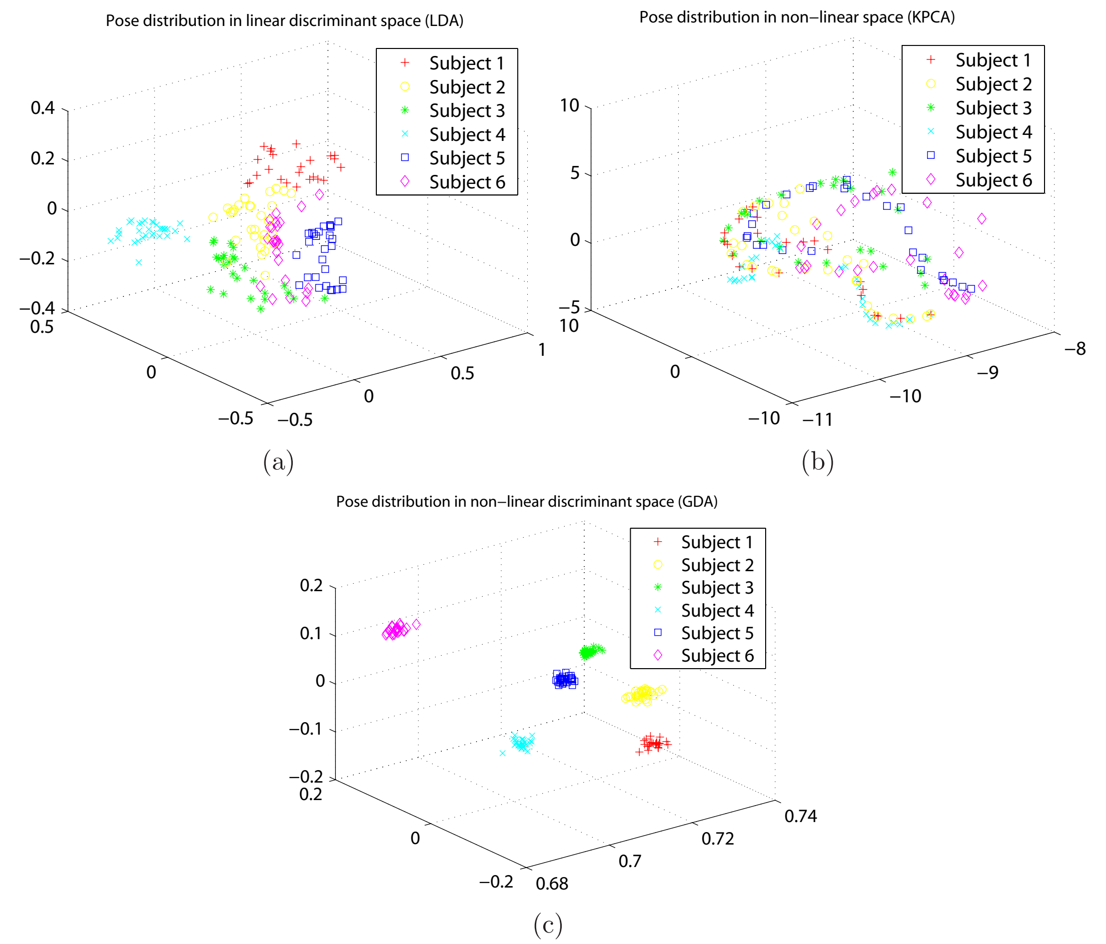


Figure 3.4: The distribustion of a set of pose varying faces in (a) LDA (b) KPCA and (c) GDA subspaces.

and the total number of samples in the learning set n is much smaller than the number of pixels d in each image. It is possible to find a transform matrix V that reduces the within-class scatter to zero making the discriminant ratio (3.5) undefined. Various methods have been used to overcome this problem in the linear and nonlinear formulations of discriminant analysis. The most common method uses PCA to project the samples onto a reduced space so that the resulting within-class scatter matrix S_W is nonsingular [10]. Formally a transform V is found such that,

$$V = \arg\max_V |V^T S_T V|, \tag{3.25}$$

with S_T being the total data covariance matrix. Samples are then transformed with V into a lower dimension space. The discriminant transform is now found by optimising W for the transformed samples,

$$W = \arg\max_{W} \frac{|W^T V^T S_B V W|}{|W^T V^T S_W V W|}. \tag{3.26}$$

The combined transform is now formulated as

$$U^T = W^T V^T \tag{3.27}$$

and a test sample x is projected as

$$y = U^T x. \tag{3.28}$$

A similar method has also been used in kernel discriminant analysis by Baudat and Anouar [9] with GDA where K in equation (3.23) is first diagonalized as,

$$K = V \Lambda V^T \tag{3.29}$$

and equation (3.23) now becomes,

$$\lambda = \frac{\beta^T V^T Z V \beta}{\beta^T V^T V \beta}. \tag{3.30}$$

Both these methods remove the null space of the within-class scatter matrix which has been shown to contain significant discriminant information [23, 126]. Many alternative methods have been proposed for dealing with the SSS problem in linear discriminant analysis [23, 72, 73, 126] and nonlinear kernel discriminant analysis [74, 120].

3.2 Kernels for Local Features

Local feature based face recognition has enjoyed mixed success in the face recognition literature. Moghaddam, et al. [84] consider using eigenfeatures where, instead of using the whole face area, smaller areas placed over facial features are employed for analysis. Eigen-decomposition of these smaller areas produces eigenfeatures of the eyes, nose, and mouth that outperformed eigenfaces when small numbers of principal components are used. Work by Lanitis, et al. [57] also uses local information in combination with global shape information to improve recognition performance. Other methods like *Local Feature Analysis* (LFA) [93] and *Elastic Bunch Graph Matching* [119] use local features for face recognition with good performance. In this Section, the application of kernel methods to local features or areas of the face is introduced. It is shown how discriminant subspaces constructed from local feature are useful and can be used in a recognition task similar to eigenfeatures.

3.2.1 Local Kernels

Instead of employing the entire face image as the input vector, restricted regions of the image are used as in [84]. Localised regions of the face are selected by defining a sliding block window over the whole image. These regions represent a local sampling of the face image and contain various features of the face as well as important internal facial features or facial landmarks. The identified local regions are then used for kernel analysis. In other words the local kernel used is defined as for a normal kernel except that a smaller vector input from the block window is used. Let $r_{u,v}$ signify a block window or region with dimension $u \times v$ of a whole face image with dimension $U \times V$. By taking a sliding window over the whole face image there are a possible $R = (U - u + 1) \times (V - v + 1)$ regions which can be used. The term r is now used to signify one of the possible R regions and hence a single local kernel is define as,

$$k^r(x, y) = k(x^r, y^r) = \exp\left(-\frac{\|x^r - y^r\|^2}{2\sigma^2}\right) \tag{3.31}$$

where x^r and y^r signify the input vectors from two samples using the region of block window r. With the same region extracted across a set of training data a GDA subspace can be constructed for that region. Projections of training data and test data for each region can then be extracted and used for classification. As the total number of regions R can be quite high it is not desirable to construct GDA subspaces for all local kernels. This is due to the increased processing required for all possible regions which may not necessarily be useful. Instead local kernels must be found which are most discriminant and therefore more useful for region selection and classification.

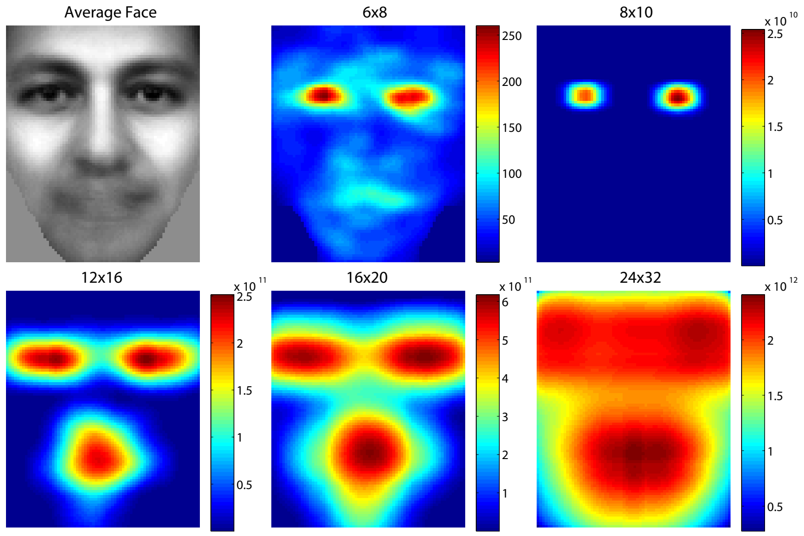


Figure 3.5: A typical discriminant criterion response over a face image (of size 88×76) for different window sizes. Window size is indicated above each image.

3.2.2 Feature Discrimination

A discriminant criterion is used to investigate how well a particular local kernel and hence its discriminant subspace may perform. Again using the same principle as used in GDA each subspace, corresponding to the local kernel $k^r(x, y)$, is evaluated as to what degree it can maximise the criterion C_{Proj},

$$C_{Proj} = trace\left(\frac{S_B}{S_W}\right). \tag{3.32}$$

S_B and S_W are the between and within-class scatter matrices of the projected data q_k, $k = 1, \ldots, M$.

$$S_B = \frac{1}{M}\sum_{i=1}^{C} c_i \bar{q}_i \bar{q}_i^T \tag{3.33}$$

$$S_W = \frac{1}{M}\sum_{i=1}^{C}\sum_{j=1}^{c_i} q_{ij} q_{ij}^T \tag{3.34}$$

Here $\bar{q}_i$ is the centroid or mean of the projected samples of class i. The criterion C_{Proj} is a useful measure since the discriminant function in GDA uses this same ratio of terms. However, some problems can arise; for example, if the number of samples for a certain class is small, their projections can be set to the same point in the feature space. This causes the denominator in (3.32) to become zero and the criterion becomes ill-posed. To counter this problem the minimum value in the scatter matrix S_W is fixed by adding the identity matrix of a small number μ. That is

$$S'_W = S_W + \mu I. \tag{3.35}$$

Taking into account the available computation precision and the nature of the processing platform μ is chosen to be an order of magnitude less than the maximum precision possible. In this case $\mu = 1e^{-14}$, and this ensures that the minimum within class scatter of a class is not zero but also not substantially influenced by this modification. A typical response based on the discriminant criterion of local kernel subspaces created from areas over the face can be seen in Figure 3.5. Low criterion values are coloured dark blue and increasing through Cyan, Green, Yellow and dark Red. A Gaussian kernel is used in this case, as in equation (3.31) with $\sigma = 90$. Although other kernel functions can be used, for example polynomial and sigmoid kernel functions (among others [105]), the Gaussian kernel has been used extensively in kernel analysis on face images with great success.

It can be seen that regions with a high criterion response are positioned over significant internal facial landmarks such as the eyes, nose and mouth. GDA subspaces therefore are most discriminant over these facial areas for a range of window sizes. Suitable window sizes however are important for good recognition performance. In Section 4.2 it is shown how smaller window sizes may not be well suited for recognition purposes as they may be more sensitive to slight changes or shifts of local facial components that they cover. Additionally smaller window sizes capture less data and hence can affect the amount of discriminating information available. Once a suitable window size is chosen appropriate local kernels are selected. These local kernels are then used to represent faces through their most discriminant features.

3.3 Evaluating Local Feature Kernels

It has been shown that kernels can be applied to local features of the face to obtain a number of feature-based subspaces similar to the modular eigenfeatures of Pentland, et al. [94]. Regions can now be selected over facial landmarks or distinguishing facial features to construct a number of feature kernel subspaces. This selection can be based on the typical criterion response shown in Figure 3.5. Kernels can then be constructed for the selected local feature regions and called

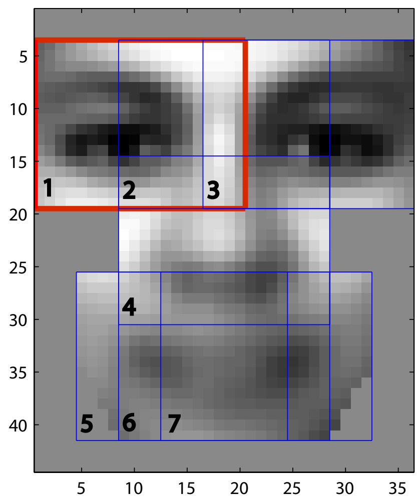


Figure 3.6: A typical selection of feature regions for an images size of 44x36. Each region is numbered in the bottom left corner of the 16×20 region block.

'feature kernels'. A discriminating kernel feature subspaces is then constructed for each feature kernel. These feature subspaces make up a modular representation for a data set of faces. For the task of classification the analogous features of any test sample are then projected into the respective subspaces. The test sample is now classified based on its features and hence a classifier is required for each feature subspace. These classifiers can then be combined (in various methods [4, 55]) to obtain a single classification result. This approach is now adopted to evaluate the performance of such a modular kernel feature subspace approach.

3.3.1 Experimental Methodology

A manual selection of seven 16×20 pixel feature regions is made for this experiment and are shown in Figure 3.6. The choice of region size is made so as to reasonably incorporate the landmark facial features. Each feature region is emphasised by a bounding box and labeled in the bottom left-hand corner by a number. These regions are placed over the eyes, nose and mouth as suggested by the discriminant response maps in Figure 3.5. Additionally the selection of features also closely resembles those made in [94]. For each feature region a feature kernel is then constructed from the training data to construct a set of feature kernel subspaces. A small subset of 100 individuals is chosen from the FERET database [96]. Each individual has seven images with variations in pose, illumination and expression of which three are randomly picked to form the training

Table 3.1: Percentage classification error per feature and with majority voting using all features.

Feature	Mean % Error
1. Left Eye	27.1
2. Mid Eyes	27.3
3. Right Eye	26.5
4. Nose	54.5
5. Left Mouth	63.6
6. Mid Mouth	62.8
7. Right Mouth	63.3
Maj. Vote	30.0

set. The remaining four images are used to form the test set. Classification performance can be measured in a number of ways, either based on the individual feature classifiers or on a number of their combinations. A basic Euclidean distance classifier is used here to ensure classification results are based on the performance of the features and not their classifiers [85]. For the combination of feature classifiers a majority-vote method is applied. This is a straightforward fusion scheme and is only used in this case as a demonstration of how local feature-based representations can be combined. Other classifer fusion schemes exist (see [4, 55]), and may perform better for this experiment, however, this is not within the scope of this thesis. For the majority-vote scheme each feature classifier will vote a test sample as belonging to an individual in the training set. The majority vote and the final classification is assigned as that individual which receives the most number of votes. Each training and testing process is run ten times to obtain an average classification result.

3.3.2 Results

Table 3.1 shows the average percentage classification error results when each feature is used separately for classification and also in a majority-vote classifier using all seven features. The results show that although central regions of the face have been selected as feature kernels most of them perform poorly in classification when considered on their own. With the exception of features 1,2 and 3 which correspond to the regions over the eyes the remaining feature kernels are poor classifiers. Combining the classification results for all the feature kernels into

Table 3.2: Percentage classification error per feature and with majority voting using features 1, 2 and 3.

Feature	Mean % Error
1. Left Eye	26.1
2. Mid Eyes	26.5
3. Right Eye	26.3
Maj. Vote	22.2

a majority voting scheme does however decrease classification errors. A further improvement might be expected by making a better choice of feature kernels based on these results. A further test is applied using only those features performing best in classification. Only features 1, 2 and 3 are used in the majority-vote classifier and the results are shown in Table 3.2. Only a slight improvement is gained with the final combined error rate dropping by $\sim 8\%$. However this shows that a good selection of features will improve performance. As only three features are used in this case there is only a slight improvement in performance due to the small majority of 2 against 1 in the voting scheme. As more salient feature classifiers are employed the performance can be expected to improve.

The method of choosing the best features is still an open problem. As mentioned above the selection can be based on the criterion response of the feature kernel subspaces. The criterion can be evaluated for all possible regions of the face and the selection based on the higher n performing subspaces. This is similar to the approach applied in this experiment, however, the combination of the selected features may not be optimal. The features are selected based on their performance ability as individual discriminators or classifiers and not on their combined performance. Furthermore this approach is process and time intensive. The task becomes difficult as the number of regions R (in Section 3.2) can be large. This is worsened even further when the total number of training image samples increases. Finding the eigenvectors a of equation (3.23) can become very difficult as the dimensions of the kernel matrix K increase proportionally with the number of samples. It is obvious that this process is quite time consuming and may also require large amounts of memory if the data set is large.

3.4 Conclusion

In this Chapter the feature kernel is defined as a kernel function applied to a limited region or structure of an image, in this case focusing on the face and its features. It is shown how nonlinear discriminant analysis can produce strong (discriminant) feature spaces using regions situated around landmark facial features. The areas predominantly around the eyes are shown as being the most discriminant followed by the areas around the mouth (Figure 3.5). This result is promising and shows a degree of similarity to the human affinity for landmark features such as the eyes. In one study on humans, [87], it was found that sensitivity to changes made to the eyes is enhanced with increased familiarisation of faces. From the results obtained in this Chapter, it could be said that when discriminant analysis is used to familiarise (learn) regions of a face the most sensitive (discriminant) areas are those of the eyes, nose and mouth. More interestingly no supervision regarding the relevance of features has been afforded to the nonlinear discriminant analysis method applied in this Chapter. Hence we conclude that nonlinear discriminant analysis used in this manner shows some degree of biological plausibility. This is not meant to imply that a biologically inspired model is explicilty being developed or that any biological similarities are necessarily beneficial. However, our developing system does bear some unplanned resemblances to the HVS. By using these discriminant feature areas, representations can be constructed for the purpose of automatic face recognition. As shown in Table 3.1 it is possible to achieve reasonable results, but it is not clear what weight of importance must be assigned to each feature when considered for recognition. This becomes important when using a fusion scheme to consolidate multiple classifier results. In the next Chapter a property of kernels is exploited which allows the construction of a single subspace for any number of feature kernels. This reduces processing demands and also requires only one classifier for any number of feature kernels.

Chapter 4

Multiresolution Face Kernels

In Chapter 3 local feature kernels were introduced and it was shown how they can be used to construct representations for discrimination and recognition similar to the eigenfeatures method [84]. While constructing feature spaces for facial landmarks or other facial features can be beneficial the representations remain distinct and separate making it difficult to construct a simple representation for the whole face. In this chapter the closure property [108] of kernels is exploited to construct a global representation based on featural parts. The kernel structure is flexible and able to incorporate any number of feature kernels at multiple resolutions. This kernel structure is called a face kernel. Commonly used feature locations are avoided and a feature selection algorithm is employed to learn an optimal selection of features for constructing the face kernel. Key parameters like feature size and image resolution are investigated to find a suitable parameter set for application purposes. The face kernel structure is compact producing only one subspace for classification. The compact nature is beneficial in the recognition task and typical face kernel extraction and classification results are discussed. Finally a hybrid kernel is proposed which combines a holistic face kernel with local feature kernels. Again this multiresolution face kernel is compact and its application and results are explored in Chapter 5.

4.1 Local-Global Combination

As described by Schölkopf, et al. [105] kernels can be constructed from other kernels. As long as the component kernels satisfy the conditions of Mercer's theorem of functional analysis [79] then their combinations are also valid kernels.

Mercers theorem gives a necessary and sufficient condition for a function $k(x, y)$ to be a kernel: its corresponding kernel matrix has to be symmetric and positive semidefinite. (See [27, 115]). Hence local kernels and their kernel matrices can be combined under summation and/or multiplication as long as the resulting and component kernels hold the aforementioned conditions. Such local kernel combinations are used in Wallraven, et al. [116] to represent faces. Points of interest over a face are selected using a corner-like detector which are then compared using kernels and trained with a Support Vector Machine. Sample face images are compared by summing the kernel values of their best matching features combinations. This means multiple feature comparisons take place and the best match is kept. The reason this takes place is because features found by the corner detector are not structured or ordered and the number of features found are not always equal between samples. In this work a similar formulation is used to construct a representation by combining local kernels. Because the most useful features of the face for classification are of interest any data set being used must be properly aligned. This means that when a kernel is calculated for a region r this region is the same region across all training samples. Kernel comparisons can only be made between the same features, hence the 'feature kernel'. In Chapter 3 the discriminant power of each feature kernel is considered to be independent. However feature kernels can be considered as one where a combined discriminative power is achieved by combining a number of feature kernels. By taking advantage of this property of kernels a number of local feature kernels can be combined simply by a summation. Hence a summed kernel of the local kernels $k^r(x, y)$ is,

$$k(x, y) = \sum_{r=1}^{R} k(x^r, y^r) \tag{4.1}$$

This formulation now defines the basic structure of a global kernel or 'face kernel' constructed from local kernels. The formulation in equation (4.1) sums all possible kernel features from the set of regions R in an image. This direct use of all the regions in a face image is not sensible as not all feature regions are good discriminators. Results from the discriminant criterion plots of Figure 3.5 show that regions around the facial landmarks are of greater importance. Additionally, calculating all possible feature kernels would be computationally expensive and wasteful when considering the number of redundant or non-salient feature regions. In order to represent a face only a small subset $S \in R$ of local kernels need to be used. These feature kernels can be selected based on their projection criterion C^r_{Proj}. By selecting an optimal combination of feature kernels a good representation can be constructed for the representation of the whole face based on salient component features.

$$k^S(x, y) = \sum_{r \in S} k(x^r, y^r) \tag{4.2}$$

The choice of what subset S to use can be manually assigned by a user with local kernels intuitively placed over facial landmarks like the eyes, nose and mouth. Automatic assignment of the subset S is also possible by allowing an unsupervised selection of local kernels based on their ability to represent a face. In other words the feature kernels are not considered by their independent discriminating power but by the discriminating power of their combination. Best representations can be found by iteratively testing different feature kernel combinations of varying size. A result is sought which minimises the number of feature regions used and by so doing minimises redundant and non-salient local kernels of the face. As can be seen in Section 3.2.2, regions that have the best separability between identities are those incorporating the eyes, nose and mouth. Hence feature kernels from these regions are expected to be selected to represent a face. In order to tackle the problem of finding the best combination of local kernels a selection algorithm like Sequential Forward Selection (SFS) can be employed. In Section 4.4 a feature selection algorithm is proposed which is used to learn an optimal set of feature kernels to construct the face kernel. However before feature selection is applied consideration must be given to the choice of window size which will be used for the definition of feature regions.

4.2 Feature Size

The size of feature windows is important when considering feature kernels. As an example a feature window can be defined by the approximate size of a facial feature or some subsection. Feature kernels of different sizes produce varying discriminant responses. This variability is exhibited in terms of discriminant power and discriminant facial regions, and ultimately affects classification performance. This is clearly shown in Figure 4.1 where the average separability criterion for a sliding window of varying size has been calculated. A colour bar indicating the

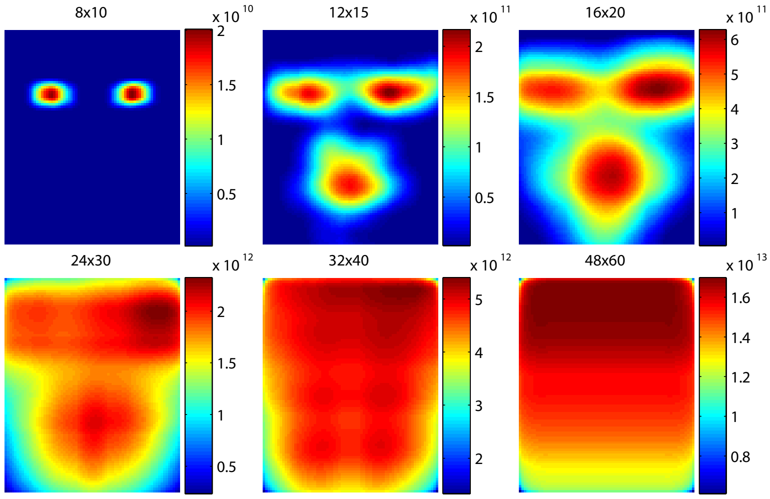


Figure 4.1: Separability criterion maps of various feature window sizes for an image resolution of 88x72. Window size is indicated on top of each image.

criterion value of each pixel colour is situated next to each criterion image map.

As mentioned in Section 3.2.2 areas over the landmark features of the face produce GDA subspaces with a high discriminant criterion response. The size and shape of these areas is dependent on window size and shape. A general trend can be seen where the maximum criterion for each criterion map increases with increasing window size. This is to be expected as larger windows produce larger dimension vectors used to construct the GDA subspace. This increase in dimension results in an increase in detail (variance and useful information) for discrimination and hence a better separability criterion. Figure 4.2 shows the maximum criteria from Figure 4.1 plotted against the ratio of window to image area. When using a high resolution image (88×72) the rate of increase in criterion slows with larger feature windows. Again this is expected as increasing resolution will give rise to an increase in high frequency noise and hence a decrease in the signal-to-noise ratio. One such source of noise is the movement of sharp gray level variations or edges of facial features due to their misalignment or movement between samples. Variations in illumination can also contribute more noise in high resolution images. Therefore, feature windows sizes are sought which best describe salient features of the face, which may include facial landmarks, while reducing noise effects and computation time. Returning to Figure 4.2 it can be seen that at a lower resolution (44×36) feature-image window ratios from 0.11 and above have comparable performance to those used at the higher resolution. This gives a suggestion as to what size feature blocks can be used for the face

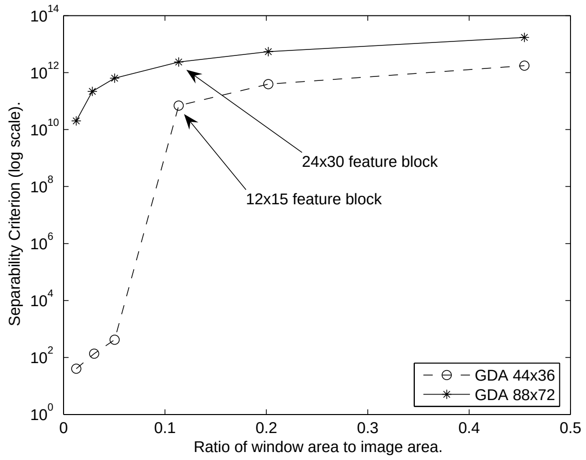


Figure 4.2: Maximum separability criterion of various feature window sizes for image resolutions of 88×72 and 44×36 pixels.

kernel. The use of a separability criterion alone is not a sufficient indicator of what feature size to use. The effect of feature size on classification performance must also be determined. This is explored in Section 4.5 after feature selection and multiresolution is incorporated into the face kernel. Although it is easy to understand how feature kernels can be used at varying resolution to obtain corresponding face kernels, the following Section proposes a method for combining these face kernels into a multiresolution face kernel structure.

4.3 Multiresolution

It is shown [17, 20] that people can recognise familiar faces even from very coarse and low quality images and especially video. People can recognise familiar faces from distances and it is believed [17, 20, 21] that such low resolution representations may be processed in a more general or holistic manner. Low resolution representations contain less high frequency noise and more useful low frequency information [40]. In terms of face recognition low resolution images are more useful in global representation architectures. In contrast, discrimination performance of unfamiliar faces is performed less readily at low resolutions and requires more detailed facial information. Face recognition architectures based on local facial features are therefore better suited to using higher resolution images where enough information is available for feature description (construction). For example if a face image of size 22×18 is used then the selection of a small area containing an eyes or any other feature becomes quite difficult. Selecting the corresponding area of the eye at an even lower resolution results in the selection of an area typically containing a few pixels. This would be insufficient to construct any subspace for discrimination. Higher resolutions allow the extraction of more detailed feature areas and hence are more discriminative. A problem arises however, as shown in Section 4.2, where high resolution features are prone to high frequency noise in the form of illumination variations and slight changes in feature position and structure. The use of multiresolution representations can therefore have mixed results and employing a discriminant analysis technique like GDA becomes important.

By incorporating a combination of feature kernels from a number of resolutions the face kernel representation may be improved. Local features from various resolutions are combined to construct the 'multiresolution face kernel' or MRFK. This combination is achieved quite easily in the same manner as kernels of different features are combined in the face kernel. The MRFK is now described as follows,

$$k^D(x, y) = \sum_{d \in D} k^{S^d}(x, y) = \sum_{d \in D} \sum_{r \in S_d} k(x^r, y^r), \tag{4.3}$$

where D is a set of image pixel width-height resolution pairs in increasing order. Each index d into D increases the resolution pair by 2^{d-1}. For example let the first image resolution pair be 22×18 then $D = \{(22; 18) \times 2^0, (22; 18) \times 2^1, (22; 18) \times 2^2, \ldots, (22; 18) \times 2^{d-1}\}$. S^d indicates a different set of kernel regions r used at each resolution step d. Effectively a set of feature kernels is selected for each resolution level to construct a corresponding face kernel. The face kernels at each level are then combined to construct the multiresolution face kernel. Feature kernel responses can vary greatly with varying resolution even if similar feature areas are being compared. Therefore before multiresolution feature kernels are combined they must be normalisation with respect to resolution bias.

4.3.1 Resolution Bias

With increasing resolution comes an increase in dimension and variance in the data. The same area of a face can be extracted at different image resolutions and the higher resolution feature will have more pixel information and hence a greater variance. When using kernel functions similar to the Gaussian kernel their resulting output can be different at varying resolutions even if describing the same image area. This is because Gaussian kernel comparisons are basically the exponentials of the sum of squared pixel differences between samples. As the number of pixels increase so an increase in the sum of squared differences is expected. When combining feature kernels, especially from different resolutions, it is expected that the kernel values be comparable. Adding a low resolution kernel (whose value might be very small) to a kernel at high resolution (with higher value) has little effect or contribution on the combined representation.

Feature kernels should therefore be normalised prior to their use in a face kernel. The difference in kernel response (in this case a Gaussian) can be seen by looking at the sum of squared difference matrices of an image feature at varying resolutions. A set of 78 image samples from 27 subjects at resolutions of 22×18, 44×36 and 88×72 are used. An image feature of size 8×10, 16×20 and 32×40 is extracted from the top left-hand corner of each image respectively. This feature is used to construct the difference matrices shown in Figure 4.3. In general a ratio of $\sqrt{\frac{22\times18}{44\times36}} = \frac{1}{2}$ can be expected between the range of values in the squared difference matrix of images of size 22×18 and 44×36. Further more a ratio of $\sqrt{\frac{8\times10}{16\times20}} = \frac{1}{2}$ is expected between image features of size 8×10 and 16×20. By multiplying these two values a final ratio of $\frac{1}{4}$ is expected between feature kernels. This can be seen from the range/color index for each matrix on the first row in Figure 4.3. Similarly a total ratio of $\frac{1}{16}$ can be expected between kernel value ranges of images of size 22×18 and 88×72 with their respective feature sizes. Because the number of pixels increases by powers of 4 a proportional increase in the range of the sum of squared differences can be expected. This in turn affects the values of the kernel matrices, and hence the low resolution kernel becomes insignificant and contributes very little in the summed multiresolution kernel.

To counter this effect a normalisation term ξ_d is added to the Gaussian kernel,

$$k^{d,r}(x,y) = k(x^{d,r}, y^{d,r}) = exp\left(-\frac{\|x^{d,r} - y^{d,r}\|^2}{2\sigma^2}\xi_d\right) \tag{4.4}$$

where

$$\xi_d = \sqrt{\frac{D_1}{D_d}\frac{B_1}{B_d}}. \tag{4.5}$$

Let B represent the set of feature window height-width pairs to be used at each corresponding image resolution in D where D_1 is the first and lowest resolution

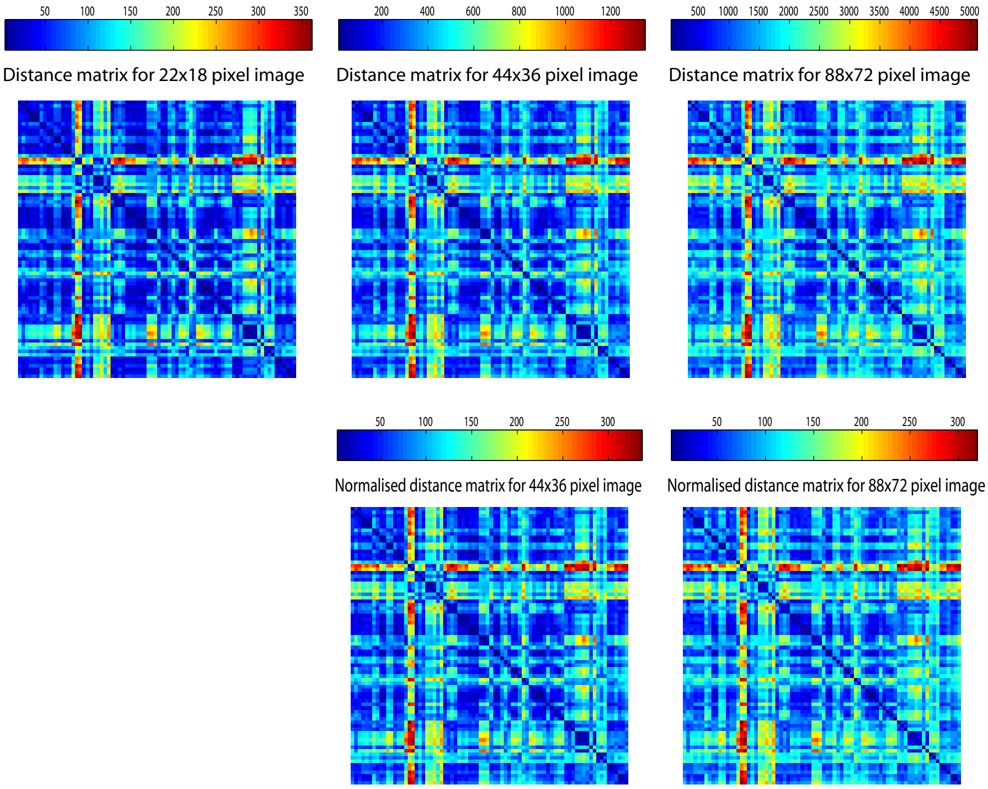


Figure 4.3: Distance matrices for a set of 78 images showing the variation in range of values of Euclidean distance (or sum of squared differences) with varying image resolutions.

and B_1 its corresponding window resolution. All kernels are now normalised with respect to kernels using image and features sizes D_1 and B_1 respectively. The results of applying normalisation to the above difference matrices can be seen in the bottom row of Figure 4.3. All the normalised difference matrices have values within the same range as the lowest resolution matrix. The normalisation of kernel features now ensures that the kernel is not monotonic with respect to resolution and adding face kernels is better conditioned.

4.4 Extraction of Face Kernels

In Section 4.1 the construction of a face kernel is described as the summation of a selection of feature kernels. A good selection of feature kernels is important as there is a wish to construct face kernels that are salient and useful for representation and recognition purposes. Using an intuitive approach it would be possible to select a set of features based on facial landmarks. This would be a selection similar to that used in the Section 3.3 and in the eigenfeatures approach. The use of these intuitively selected feature kernels does not necessarily construct the optimal face kernel. Separate feature kernels might perform well in classification tasks but not when combined as a face kernel. So an optimal set is needed of feature kernel combinations in terms of discrimination or classification. Testing all possible local kernel combinations would be very difficult and time consuming. For example considering that a face image of size 44×36 can have 1,131 sliding windows of size 6×8. To avoid such an exhaustive search feature selection methods are essential.

In general feature selection involves the selection of a subset S of features from a given original set R without significantly degrading the performance or accuracy of the system they describe. The selection is based on the effectiveness of the feature subsets which is usually gauged with a criterion or performance measure such as a separability measure. Usually features are first extracted from a set of data by some analysis or dimensionality reduction, like PCA, after which feature selection takes place. At this point a distinction can be made between feature selection and feature extraction. Feature selection algorithms identify and select the best subset of an extracted feature set with respect to the target task (e.g. classification accuracy). Feature extraction methods, however, create (extract) new features based on transformation or combinations of the original features in a set [98]. For the selection of local feature kernels, the use of all the possible regions of a face image is defined to be the complete representation or set. Now the task is to find a subset of kernel features to construct (extract) salient face kernel representations. A feature selection algorithm is employed to find salient feature kernel combinations based on a separability criterion. A multiresolution face kernel is then obtained by applying kernel feature selection at increasing image resolutions. In this way face kernels are extracted for each resolution level and combined to form the multiresolution kernel.

4.4.1 Feature Selection

Feature selection was first employed by Marill and Green [76] using sequential backward selection (SBS) and later Whitney [118] developed sequential forward selection (SFS). SFS and SBS are two very fast but sub-optimal algorithms. SFS begins with no features in the subset and proceeds to sequentially construct subsets by successively adding an additional feature. The subset S_n of size n is constructed by adding to the subset S_{n-1} of size $n-1$ that single feature that gives the best performance for the new subset. In this way only $N(N+1)/2$ subsets need to be examined where N is the total number of features. SBS works in an analogous manner but starts with the entire set of features and successively removes features in searching for the optimal subset. These algorithms are sub-optimal because they construct nested subsets, i.e. they will choose subsets such that $S_{n-1} \subset S_n$, $\forall n$. This property does not necessarily hold for the optimal subsets and the departure from optimality can be large. Pudil, et al. [99] proposed modifications to the SFS and SBS algorithms with floating sequential algorithms where they developed sequential forward floating selection, SFFS and its counterpart backward algorithm SBFS. The SFFS and SBFS methods have been shown to produce comparable results to the optimal branch-and-bound method, and are generally faster [132].

A modified SFFS algorithm is used in the face kernel structure to find the best combination of feature kernels. In general this is trying to solve the equation,

$$\arg\max_{S_n \subset R}(C_{Proj}(S_n)), \tag{4.6}$$

where $C_{Proj}(S_n)$ is the feature space separability criterion of the projected samples using the selected subset S_n of feature kernels. The criterion C_{Proj} is that described in Section 3.2.2 and the selected set is used to construct the face kernel representation and feature space. Local kernels are calculated for each feature block of the current image size D_d and then passed on to the SFFS algorithm. Sets of feature blocks are selected based on how their respective subspaces project the training data (4.6): for each set of features S_n^d the respective local kernels $k^r(x,y)$, where $r \in S_n^d$, are calculated and added;

$$k^{S_n^d}(x,y) = \sum_{r \in S_n^d} k(x^r, y^r). \tag{4.7}$$

This summed kernel $k^{S_n^d}(x,y)$ is now used to calculate the respective subspace and separability criterion for the projected training data. A pseudo code description of the feature extraction algorithm is shown in Algorithm 1. In this forward searching algorithm, kernel features are added starting with the empty set S_0^d. After each new subset is created S_n^d the algorithm backtracks to test if removing any feature (other than the recently added feature) from the current subset gives

Algorithm 1 Feature extraction algorithm

procedure ModSFFS
  Set $d \in D$ ▷ Selected image resolution.
  Set $r \in R_d$ ▷ Regions based on window size and image resolution
  $n \leftarrow 0$
  $S_n^d \leftarrow \{\}$
  while $n \leq N$ **do**
    Find $r \notin S_n^d$ such that $J(k^{S_n^d}(x,y)) < J(k^{S_n^d}(x,y) + k^r(x,y))$
    if $J(k^{S_n^d}(x,y) + k^r(x,y)) > J(k^{S_{n+1}^d}(x,y))$ **then**
      $S_{n+1}^d \leftarrow S_n^d \cup r$
    end if
    $n \leftarrow n + 1$
    $Backtrack \leftarrow 1$
    while Backtrack **do**
      Find $\arg\max_{r \in S_n^d} \left\{ J(k^{S_n^d}(x,y) - k^r(x,y)) \right\}$
      if $J(k^{S_n^d}(x,y) - k^r) > J(k^{S_{n-1}^d}(x,y))$ **then**
        $S_{n-1}^d \leftarrow S_n^d \backslash r$ ▷ $\backslash$ is set difference
        $n \leftarrow n - 1$
      else
        $Backtrack \leftarrow 0$
      end if
    end while
  end while
end procedure

a better criterion that the previous subset S_{n-1}^d. If not the algorithm continues finding subsets of increasing sizes. Otherwise if a superior subset is found after removing a feature then S_{n-1}^d is updated and backtracking continues. When no more features can be removed to construct superior subsets the algorithm continues in its forward search mode. In Algorithm 1, $S_n^d \backslash r$ indicates the set difference of S_n^d and r while the operator $J()$ is analogous to C_{Proj}. More formally $J()$ is the separability criterion of the projected sample training data using the feature kernel structure specified within the parenthesis.

The feature selection algorithm continues to create subsets of increasing size up until the set containing all possible N features. This is not desirable however due to processing time constraints. Additionally we are only interested in using the most salient feature combinations without any redundant features. By changing N in Algorithm 1 to M, where typically $M \ll N$, construction of face kernels can be based on a varying number of features. Section 4.5 explores how the number of features M effects classification performance of the face kernels, and also discusses the hierarchical ranking and type of facial features typically selected as strong feature kernels. Once feature selection has been completed at

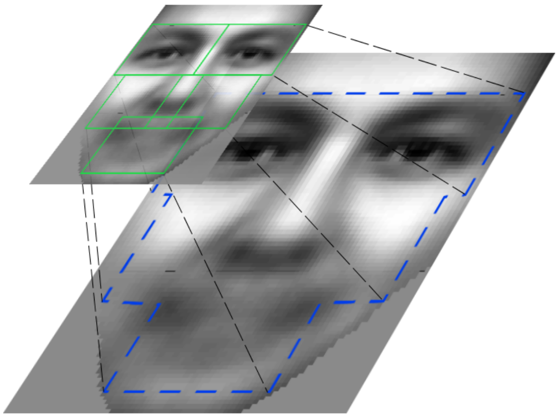


Figure 4.4: Using the combined area of selected feature regions as a guide for consecutive feature searches at the next resolution level.

a certain image resolution further processing can continue at a higher image resolution. The next section looks at how feature kernel position information can be fed forward to consecutive stages of feature selection. This is useful for refining the feature selection search at each resolution level and also reduces processing power requirements.

4.4.2 Multiresolution Pyramid

In order to enhance the face kernel representation a multiresolution face kernel (4.3) has been proposed. To incorporate face kernels from a number of image resolutions feature selection at each resolution level must be applied. Feature position information from an initially extracted face kernel is used in subsequent extractions. The objective of this is twofold. Firstly, reducing processing time by restricting the number of feature kernel combinations being evaluated. Secondly, feature selection is refined by using a coarse-to-fine approach to home in on salient areas of the face.

Before feature selection starts at any resolution level d the set of all possible window regions R_d for a given window size is constructed (see Section 3.2.1). This is the set from which a smaller subset S_n^d is extracted (via Algorithm 1) which represents the set of selected feature kernels. If R_d is large the feature selection process can be slow and using a reduced initial set of features $\tilde{R}_d$ can improve processing time. The size of R_d can be reduced for consecutive face kernel extractions by passing on feature information. This is achieved by restricting

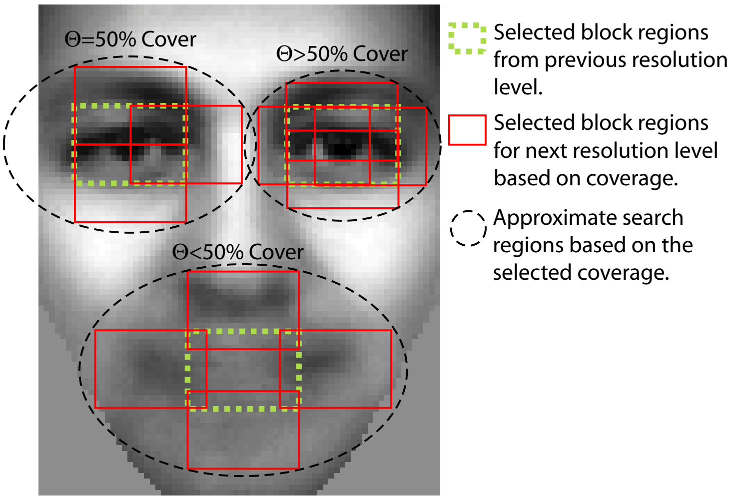


Figure 4.5: The selection of facial regions based on Θ for the construction of the initial pool of regions $\tilde{R}_d$ to be used in feature selection.

feature selection to facial areas which were previously selected. Facial areas which are not selected are assumed to be non-salient. These areas are therefore left out in subsequent feature selections. To achieve the feed forward approach, the face area covered by an initial face kernel is projected onto the next resolution level, Figure 4.4. This is simply done by finding the common area covered by all feature kernels and expressing this as a binary image. This 'coverage' image is then resized to the higher image resolution. The reduced set of valid feature windows for processing at the next resolution level, $\tilde{R}_{d+1}$, are those that cover the valid area (being binary one) of the 'coverage' image by some percentage amount Θ. Varying the value of Θ changes how strictly the initial selection of features is made around the features areas of previous feature kernels. This is shown in Figure 4.5. With small Θ the search regions become larger. The number of selected feature regions therefore increases and $\tilde{R}_{d+1}$ can quickly approach the full initial selection R_{d+1}. Therefore allowing too much growth will increase the initial pool of features and increase processing time. By keeping Θ high (above $\sim 70\%$) a tighter search region is achieved. This still allows growth of the search area around previously selected features but is now more controlled. Smaller search areas relate to smaller $\tilde{R}_{d+1}$ and hence less processing time. At the same time the chance of extracting features at consecutive levels, which may not have been present on previous levels, is possible. The value of Θ is set at 75% for the remained of the thesis in all experimental procedures. From empirical analysis this value allows for sufficient growth of the search area while also reducing the

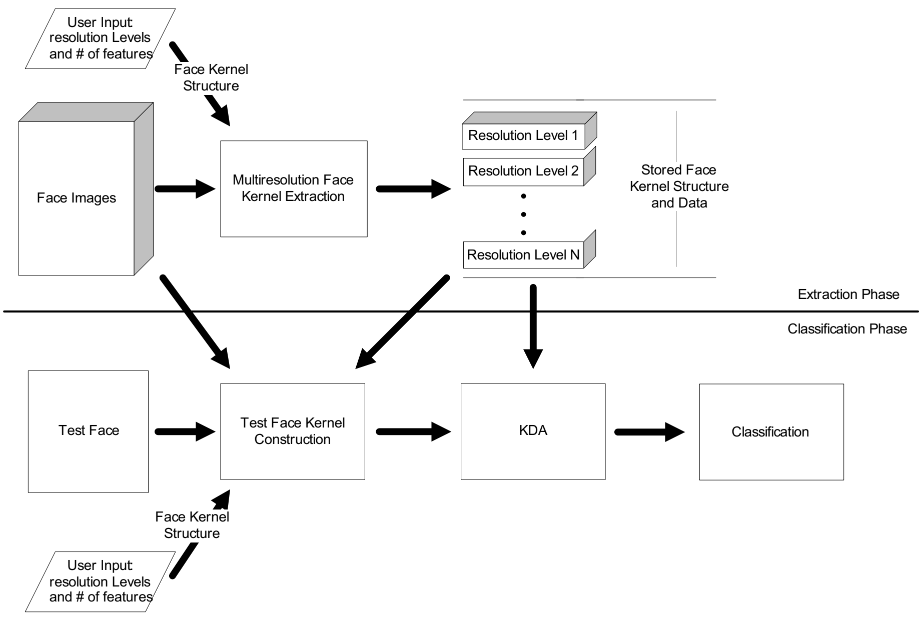


Figure 4.6: Structure of the face kernel extraction and test sample classification process.

initial pool of chosen regions to be used in feature selection.

4.4.3 Face Kernels for Recognition

We have shown how feature kernels can be used to extract face kernels. Starting at a low image resolution feature kernel selection is used to extract a face kernel. Once this is complete the process then moves up a resolution pyramid. The face area covered by feature kernels in a previous resolution level is used as a starting point for the next level of feature selection as described in Section 4.4.2. The extracted face kernels are now combined to construct a multiresolution face kernel (MRFK). Figure 4.6 shows the structure of the face kernel extraction and classification process. For all training samples the MRFK is used to construct a kernel matrix K. If $\{x_1, x_2, x_3, \ldots, x_N\}$ is a set of training samples then the kernel matrix K is as follows,

$$\begin{aligned} K &= (k_{ij}^D)_{ij} \\ &= \sum_{d\in D}\sum_{r\in S_d} k(x_i^r, x_j^r) \qquad \forall\ i,j = 1,2,3,\ldots,N. \end{aligned} \tag{4.8}$$

The kernel matrix is then used to calculate a set of orthogonal bases which represent a discriminant feature space (3.23). Classification of an unseen test face proceeds by using feature areas of the test image at each resolution level to construct test feature kernels. The feature areas used are those which correspond to the feature areas of the previously extracted MRFK. Test feature kernels can then be combined to construct the test MRFK. The test kernel $k(x_{li}, x)$ in equation (3.24) now becomes a multiresolution test kernel,

$$k^D(x_{li}, x) = \sum_{d \in D} k^{S^d}(x_{li}, x) = \sum_{d \in D} \sum_{r \in S_d} k(x^r_{li}, x^r). \tag{4.9}$$

This test face kernel is then centred and projected into the face kernel subspace as in Equation (3.24). Classification is carried out by means of Euclidean distance comparisons with the training data projections. The test face is taken as belonging to the face or class of faces having the lowest Euclidean distance between feature space projections.

The MRFK is a view-based and feature-based representation which constructs only one feature space representation. This is desirable as it does not require multiple classifiers as in the eigenfeatures method. Classification of a test face is done with one classifier without the need of classifier fusion strategies. In the following section the qualitative performance of the MRFK is evaluated. An intuitive discussion of the face kernel extraction is also made based on what type and number of features are selected.

4.5 Evaluating Face Kernel Extraction

Face Kernel extraction is now applied on a small subset of 40 identities from the FERET database [96]. The subset consists of 7 images per person with variations in illumination, expression and pose. Multiresolution face kernel extraction is carried out at image resolutions of 22×18, 44×36 and 88×72 pixels. In order to evaluate the effect of the number of extracted features on classification performance, feature sets of cardinality up to 14 are extracted. Similar to D (the set of image pixel height-width resolution pairs) used in Section 4.3, a number of window height-width pair sets, W, are defined for use in face kernel extraction. For example let the first window height-width pair of a set be 4×5 pixels then $W = \{(4 \times 5), (8 \times 10), (16 \times 20)\}$ where each pair corresponds to the feature

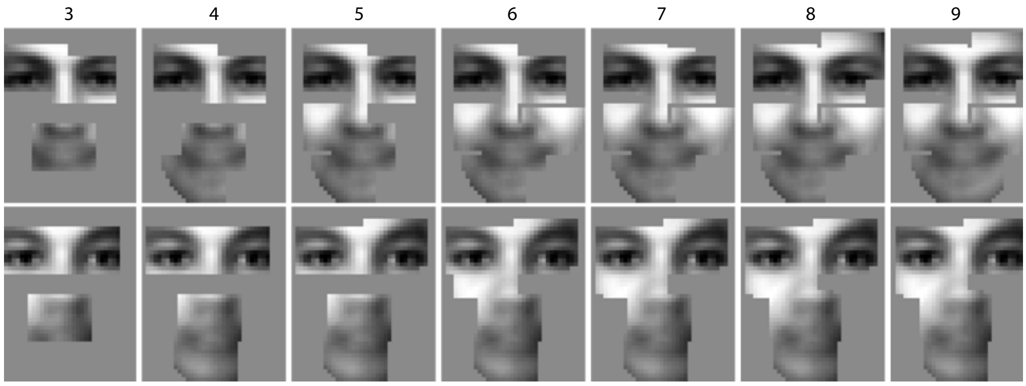


Figure 4.7: Different number of feature sets sizes starting from from 3 on the left to 9 on the right.

window size to be use at each image resolution height-width pair of D. Face kernel extraction is evaluated with the above set and two other sets with starting width-height pairs of 6×8 and 8×10. The training set is made of three images per identity randomly chosen and each extraction and classification is run six times to obtain an average. The 4 remaining images per identity are used as the testing set.

4.5.1 Extracted Features

Figure 4.7 shows the typical face kernels extracted by the modified SFFS algorithm. From left to right each column shows an increasing number of features from the set S_k, $k = 3, \ldots, 9$. The second row shows features selected for a small batch of faces from our training set with variations including pose change. The first row training batch does not have variations in pose. As can be seen the local/global kernel representation successfully extracts internal and salient features of the face. The first three, and perhaps the most important, extracted features are those of the eyes, nose and mouth and as increasing numbers of features are added more of the face is covered. The feature extraction algorithm can therefore construct face kernel representations without any structural or configural information of faces in general. Three important properties of the extracted face kernel have analogies with psychological findings. (1) The face kernel successfully utilises facial landmarks which are known to be the most salient for face recognition and discrimination [87]. (2) Additionally structural information is also present in the face kernel in the form of the positioning of the individual feature

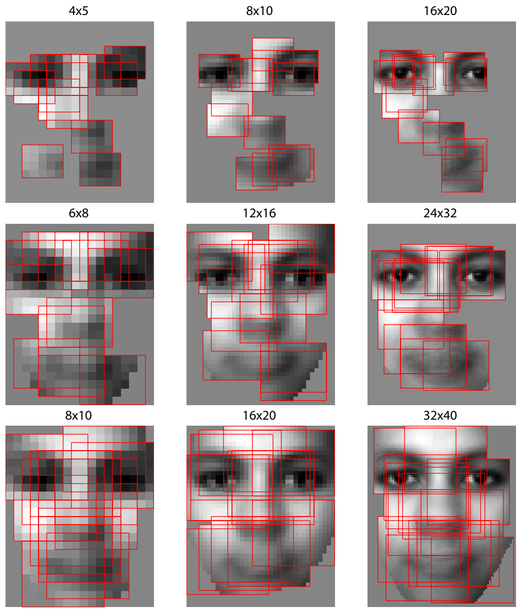


Figure 4.8: Typical extracted feature areas of varying size indicated above each image. The columns from left to right have image resolutions of 22×18, 44×36 and 88×72.

kernels [49, 106, 125]. (3) A degree of feature ranking or hierarchy is evident where the first three most important features selected are those of the eyes, nose and mouth. Other less salient facial areas come afterwards. Figure 4.8 shows some typical face kernels extracted at the resolutions stated above with varying feature sizes.

4.5.2 Number of Features and Feature Size

A return is now made to the question of what kernel feature sizes to use and how many features are needed for recognition. As shown in Section 4.2 it would be possible to use the criterion response to estimate what feature size to use. This would just be a good guess, however, as the criterion maps shown in Figure 4.1 are not very distinct. At an image resolution of 88×72 window sizes of 8×10 are most discriminant around the eyes and inner eyes only and this excludes any other facial features. This window size is also too small to include the whole eye and only captures sections of it. Also the maximum criterion response is small compared to other window sizes. There is accuracy in terms of finding the two most important facial landmarks but not necessarily in terms of classification. The 12×15 and 16×20 feature windows are more general in terms of feature sensitivity being most

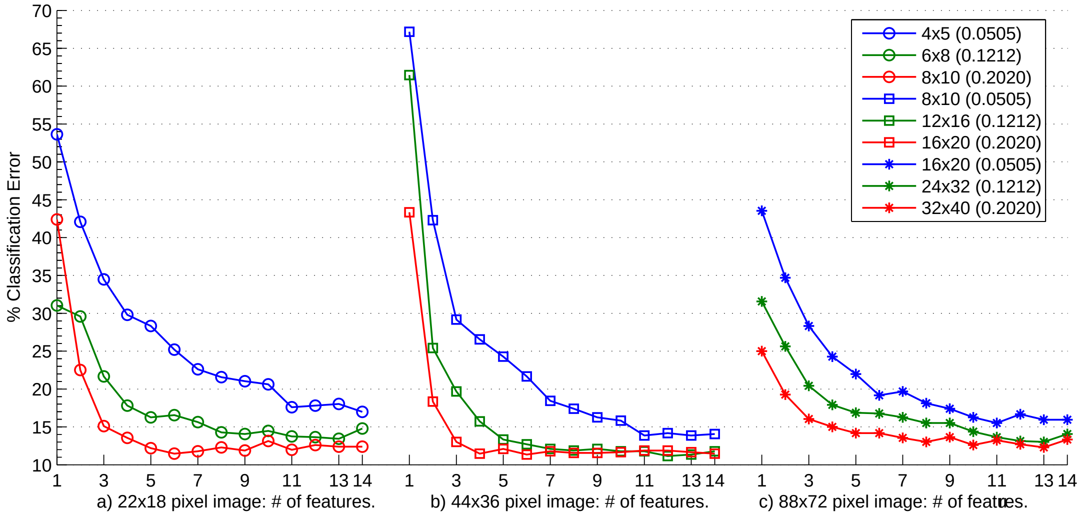


Figure 4.9: Classification error for various feature window sizes for a number of image resolutions.

discriminant at regions around the eyes and midway between the nose and mouth. These feature sizes are again smaller than any whole landmark feature like the eyes or mouth and at best only describe sub-regions of these features. Larger window sizes of 32×40 and 48×60 can be too general by incorporating whole regions of a face including more than one feature. Additionally at high image resolutions larger feature windows also contain more high frequency noise which may affect classification performance. Therefore the medium feature window sizes are left and in this case the 24×30 pixel window which closely matches the size of the facial features.

Figure 4.9 shows the effect of feature size and number on classification error. Firstly it is very obvious that incorporating an increasing number of features is not always justified, and there is a degree of redundancy. For feature windows to image size ratios of about 0.12 and higher the least number of features that can be used is about 8 for a low classification error. It can also be seen that the best results for feature extraction and classification are achieved in the mid range image resolution of 44×36 pixels with feature windows sizes of 12×16 and 24×32. Figure 4.10 is a cross-sectional plot of the results in Figure 4.9 at feature sets of cardinality eight. This clearly shows how feature size affects classification error, large kernel features in the high resolution image perform worse than kernel features in the lowest image resolution. This can be accounted for by high frequency noise in the high resolution images and feature windows.

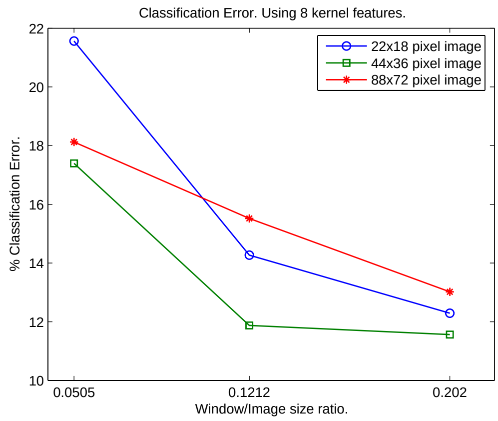


Figure 4.10: Classification error for various feature window sizes and image resolutions using 8 kernel features.

This noise can be in the the form of increased detail (variances), slight image and feature translations and illumination changes. Feature kernels in the mid-range image resolutions perform the best over all.

From these results it is possible now to select a set of image resolutions and corresponding feature sizes for a face kernel representation. Two lower image resolution levels may be used since they perform better in combination with features with an window/image size ratio of 0.20. The elimination of feature kernel extraction at high resolution will also reduce processing time, and can be reduced further when using larger feature sizes. Extraction of the multiresolution face kernel using the above data set takes $\sim$ 20 minutes on an AMD Athlon XP 2800+ running at 2GHz. This is a very large processing time considering the small data set size. In Chapter 5 a method for avoiding the computationally expensive construction of feature subspaces for feature kernel extraction is presented. This is possible while still extracting salient feature kernels. At this point only isolated feature kernel sizes and image resolutions have been evaluated. The multiresolution capability of the face kernel, by combining feature kernels from different image resolutions, is exploited in Chapter 5 where classification performance is evaluated.

4.6 A Flexible Hybrid Construction

It has been shown that the face kernel structure incorporates local and structural information of faces and combines them into one representation. The face kernel representation, at this point, is a feature based representation and can easily be extended using global information into a hybrid representation. The global representation is achieved in the traditional way where the whole face image is used to construct a kernel matrix K_G. The hybrid multiresolution face kernel (HMRFK) can be constructed as,

$$\begin{aligned} k^H(x,y) &= k^L(x,y) + k^G(x,y) \\ &= \sum_{d\in D}\sum_{r\in S_n^d} k(x^r,y^r) + \sum_{d\in D} k(x^d,y^d), \end{aligned} \tag{4.10}$$

where the global kernel is added to the face kernel at each resolution level. This allows the HMRFK representation to contain local and global information at a number of resolution levels. The kernel construction in Equation 4.10 is rigid and suggests that all resolution levels in D should be used. It is possible, however, to be flexible in which features to use and at what resolution levels. Depending on the required classification task or some restrictive conditions, such as process time or memory limitations, a subset of features and resolution levels may be used. This allows the construction of a HMRFK representation based on feature subsets sizes at resolutions that perform best in recognition tasks and that are more robust to any degrading effects like noise. In Section 4.5 it was shown that, in the case of face recognition, very low and high resolution representations perform worse than those at medium resolution in terms of classification error. It was also shown that local features at medium resolution with feature size to face size ratios of ~ 0.125 perform best in classification tasks. Using such information a HMRFK can be constructed with only those resolutions and features which are required.

Chapter 5

Application and Results

The theoretical framework of feature kernels, multiresolution face kernels and hybrid face kernels has been proposed in the previous chapters. Some configural aspects such as the setting of feature sizes, image resolutions and learning strategies have been discussed. Although the kernel representations presented are ones which consolidate multiple image formats and feature structures, the discriminant learning unit which is, in this case, GDA is cumbersome. Multiclass discriminant analysis using GDA alone becomes unusable in real world applications with large data sets containing multiple images per sample. Feature selection for face kernel extraction is similarly affected as it is based on a discriminant criterion derived from feature space projections using GDA. In order to apply these kernel representation effectively to face recognition, modifications can be applied which reduce the computational requirement of processing large data sets. In the following two sections, two such modifications are proposed, one being batch processing and a second which avoids the GDA-projection-criterion process.

The remaining sections are concerned with evaluating the performance of the proposed face kernel models. Performance can typically be evaluated using a number of measures such as processing requirements and complexity, processing time, robustness to noise, and recognition performance. The measures are chosen according to the needs of the final application task; and some tradeoff between accuracy and speed is commonplace. As no particular effort has been put into developing fast implementations of the algorithms used, comparisons based on processing time are not reasonable. In this work evaluation of the face kernel models and their various configurations is based only on recognition performance and error rate. This does not seriously hinder the quality of the evaluation as a large proportion of published face recognition literature is evaluated in a similar manner. Provided comparable data sets and experimental procedures are used, a satisfactory comparison can be made with existing face recognition models.

5.1 Batch Processing

The MRFK feature extraction algorithm given in Algorithm 1 requires the construction of multiple GDA feature spaces. This can become computationally expensive as the size of the kernel matrix for GDA, $K = (k(x_i, x_j))_{1\leq i,j\leq M}$, increases with the total number of samples M. Hence the kernel matrix can become large as the samples per class and number of classes in a data set increases. Large kernel matrices in turn slow down processing time, as diagonalising equation (3.23) becomes more difficult, and requires larger amounts of memory. As an example, a training set containing 200 classes/identities and three images for each identity will produce a 600×600 kernel matrix. Extracting a face kernel with feature set sizes of one to eight means storing a maximum of $\sum_{i=1}^{8} i = 36$ distinct kernel matrices. Using single-precision arithmetic will require ~ 50 megabytes of memory to store the kernel data alone. If a MRFK is extracted then memory requirements will increase according to the number of resolution levels. Furthermore diagonalising a single such kernel in Matlab using its SVD function requires ~ 8.4 seconds running on an AMD Athlon XP +2800 at 2 GHz. Extracting the 8 sets of features from a total pool of $R = 369$ feature regions[1] requires approximately 5544 comparisons of GDA subspaces. Each subspace is constructed by applying SVD twice (see Section 3.1.2) and hence each face kernel extraction requires approximately 25.8 hours[2] of processing time. This is an enormous amount of time and hence realistic application of the face kernel extraction becomes unfeasible.

One way to reduce the computational load is to process the data in smaller batches by constructing a few MRFK representations for the whole data set. A data set of N classes is divided into B batches as evenly as possible with each batch having n_b classes so that $N = \sum_{b=1}^{B} n_b$. MRFK representations are then extracted for each batch of training faces and their respective feature spaces and parameters stored. Extensive experiments indicate that batch sizes should be at least 10% of the whole data set size. Returning to the above example, SVD of a

[1] Using an image size of 44×36 and a feature region size of 12×16.

[2] Assuming all other necessary calculations take up negligible process time.

60×60 dimensional kernel matrix now requires only ~ 0.0048 seconds and reduces the extraction time for one face kernel to approximately 53 seconds. When this is applied to each batch the total process time for the data set is now ~ 8.8 minutes. This is a considerable reduction and means kernel extraction can now be more easily applied to a large data set.

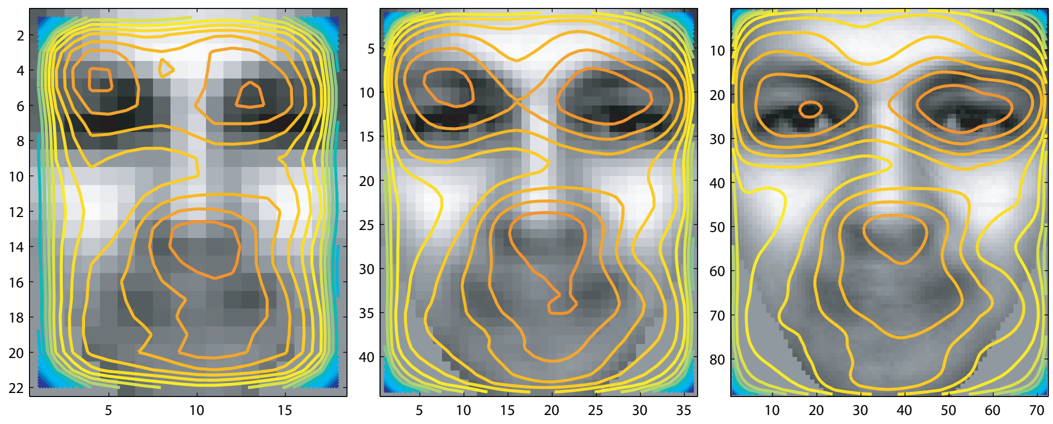
Batch processing can, however, have an adverse effect on the classification of unseen samples. Each MRFK representation is only trained to discriminate its small subset of the complete data set. If a previously unseen sample needs to be classified then that sample must be projected into each batch MRFK feature space. This requires that each batch MRFK subspace is highly specific or tuned to test samples of identities belonging to its batch. Hence specific test samples will be favourably represented in batch MRFK subspaces to which they belong and penalised in others. This can be called the capacity of the feature space. In some cases when constructing feature spaces to represent faces it is generally required that this capacity be large. This means that when projecting faces of persons not used to construct the feature space a good classification of test samples of that person is still achieved. In contrast, for the application of the batched MRFK, feature subspaces with a smaller capacity are more desirable. This is required to reduce misclassifying test samples into incorrect MRFK subspaces and hence incorrect face classes. With the use of discriminant subspaces and the uniqueness of feature set selections between batch MRFKs the capacity of their subspaces can be reduced. This is discussed further in the concluding sections of this chapter.

In this section batch processing of large data sets is proposed and was briefly shown to be effective in reducing the computation load. However a kernel matrix still must be diagonalised and this can still be slow when evaluating a maximum of $n(n+1)/2$ combinations of feature kernels. In the next Section, a method for the extraction of feature kernels is proposed without the need of constructing a feature space, projecting samples and then measuring a criterion response.

5.2 A Kernel Matrix Criterion

As mentioned above, constructing a GDA subspace can be very process intensive especially with large kernel matrices. The construction of a GDA subspace is achieved using Matlab's SVD function, once to diagonalise the kernel matrix (3.29) and a second time to find the discriminant directions β (3.30). This can still be very process intensive and so a basic modification is presented for selecting feature kernels for face kernel extraction. The kernel matrix in equation (3.14) is generally a matrix of pairwise comparison values between all the samples in the training set. The kernel matrix represents all the information needed by a learning machine. This information describes the structure of the non-linear feature space created by the kernel function and the relative positions of the data in the feature space. By directly tuning the kernel matrix it is possible to obtain an improved feature space and hence a better representation via the learning machine. This has been applied in Chapter 4 by finding optimal combinations of local feature kernels based on the subspace discrimination criterion. However as mentioned above this can be computationally expensive especially with a large data set.

An alternative approach is to evaluate the kernel matrix or combinations of kernel matrices directly. The kernel matrix structure reflects the relative positions of the sample data in the feature space. For example, when samples are similar then their kernel values, be they from a Gaussian kernel, polynomial kernel or others, will be very similar. In the case of different classes (eg. face images of various people), samples from the same class will ideally tend to have similar kernel values. Hence with a number of samples x_{ci} of class c and index i the symmetric kernel matrix of size $N = \sum_{j=c}^{C} N_c$ will have block diagonals of very similar value. It is assumed that the kernel function can adequately represent within-class samples as similar (with a high kernel value) and between-class samples are dissimilar (lower kernel value). This strong block diagonal structure is the desirable target structure of the kernel matrix. By taking a ratio of the kernel responses for within-class and between-class samples a measure of the representational strength of the kernel matrix (and hence the discriminant subspace) can be obtained directly. This is called the kernel criterion and is used to measure the fitness of a feature kernel or combinations of feature kernels. More generally a set of feature kernels is sought which undergo a favourable transformation in the non-linear feature space. For an ideal kernel matrix the block diagonal structure will be strong for all classes as equally as possible. Therefore the new metric or criterion must be averaged over all classes. In other words feature kernels are found which are strongly and evenly discriminant for all classes in the training data. The new kernel criterion C_K,



(a) (b) (c)

Figure 5.1: Contour plot of kernel criterion response for image and feature sizes of (a)22x18-3x4 (b) 44x36-6x8 and (c) 88x72-12x16 respectively.

$$C_K = \frac{1}{C}\sum_{i=1}^{C} C_c \tag{5.1}$$

is calculated directly from the kernel matrix where C_c is the criterion for class c,

$$C_c = \frac{\frac{1}{{N_c}^2}\sum_{i_c=1}^{N_c}\sum_{j_c=1}^{N_c} K(i_c, j_c)}{\frac{1}{N_c(N-Nc)}\sum_{p\in\{1,\ldots,C\}\setminus\{c\}}\sum_{i_c=1}^{N_c}\sum_{j_p=1}^{N_p} K(i_c, j_p)}. \tag{5.2}$$

Equation 5.2 is the ratio of the mean of the block diagonal of class c and the mean of all other samples. A typical response of the criterion when applied to a collection of face images is shown in Figure 5.1. Contour plots of the kernel criterion response are superimposed over the average face. More detailed kernel criterion response plots over a range of image resolution and feature sizes can be found in Appendix A, Figure A.1. A high response is suitably concentrated around the facial landmarks similar to the standard criterion response shown in Figure 3.5.

The new kernel criterion is therefore similarly responsive in the central facial regions as is the feature space projection criterion. This criterion is now used instead of the standard criterion which requires the construction of a GDA subspace. The feature kernel selection process is now much faster. From empirical measurements using batch processing and the subspace-criterion approach require on average 20 minutes processing time using a data set of 269 identities with three images each (described in the following section). The feature and images sizes and computer architecture are those specified in the previous section. Using the kernel criterion for feature extraction an average time of 2 minutes is now achieved for the whole data set. The qualitative change in feature selection is small and this is shown from classification results using the new kernel criterion. Good results are achieved as well as reducing the amount of processing power needed to use the face kernel representation effectively.

5.3 Experimental Methodology

Face images are taken from the FERET face database [95, 97]. This is a large database popular for the evaluation of face recognition methods. A subset of the large database is used in this case and incorporates variation in appearance over time, pose, expression and illumination. Pose angles variation is in the range of ± 25 degrees from frontal pose. For the purpose of comparing the face kernel performance with varying illumination two data sets are constructed. The first data set, henceforth *Set A*, consists of 269 subjects from the 'fa', 'fb' and 'ba', 'bj', 'bd', 'be', 'bf', 'bg' subsets of the FERET database. The second data set, henceforth *Set B*, uses *Set A* and is augmented with the 'bk' subset to incorporate illumination variations. Each identity has at least 6 images in *Set A* which vary with pose, age and expression all adding to a total of 1729 images. Each identity in *Set B* has at least 7 images, with variations in illumination and those present in *Set A*, totaling to 1929 images. There are also some variations with respect to facial accessories such as glasses within these data sets but have been kept to a minimum.

The FERET database also consists of a list of eye coordinates for faces in the 'fa','fb','ba','bj' and 'bk' subsets. Eye coordinates for the remaining subsets were manually placed and included in the eye coordinate list. The eye coordinates are used in a preprocessing stage aligning the face images, via a linear conformal transform and bicubic interpolation, so that their eyes are horizontally placed and separated by a fixed distance. The images are then tightly cropped around the face producing images of 88×72 pixels. A mask is then applied to remove areas which are not part of the inner facial region such as the hair and ears. Histogram equalisation and vector normalisation is finally applied to reduce the effect of illumination variations. A Gaussian resolution pyramid is then built for each

Table 5.1: Feature size configurations used in multiresolution processing. All dimensions are in $Row \times Col$ format in pixels.

	Image Resolution		
Configuration	22×18	44×36	88×72
1	6×8	12×16	24×32
2	8×10	16×20	32×40
3	12×16	24×32	48×64

image producing corresponding images of size 88×72, 44×36 and 22×18 pixels. The image pyramid is stored in a data structure which can be accessed during the training and classification process. Training sets are made by randomly selecting 3 images for each subject and the remaining images are then used as the test set.

As mentioned above batch processing is used to reduce the computational load. The data sets of 269 subjects are divided into 9 batches containing 27 subjects and one batch of 29 subjects. Each training and classification test is carried out 10 times to obtain an average performance result. Table 5.1 describes the image and feature size configurations used in each test scenario, while the total number of kernel features to be extracted at each resolution level for the face kernel is set to 14.

5.4 Extraction and Application of Face Kernels

Face learning is initiated by learning each batch of faces in turn. For each batch, processing begins at the lowest image resolution of 22×18 corresponding to resolution level 1. Face kernel extraction is started using all possible feature windows of the chosen configuration (Table 5.1). Once kernel extraction has found sets of up to 14 feature kernels the feature kernel sets $S_1^d, \ldots, S_{14}^d$ are stored, where d indicates the resolution level. The area covered by the set S_{14}^d is projected into the next image resolution level and used as the starting point for kernel extraction as described in Section 4.4.2. At each resolution level, feature kernel subsets are stored as these are later used to construct multiresolution models of the set of faces in the training batch. In this way face kernels of varying feature number and from varying image resolution can be used in recognition tasks.

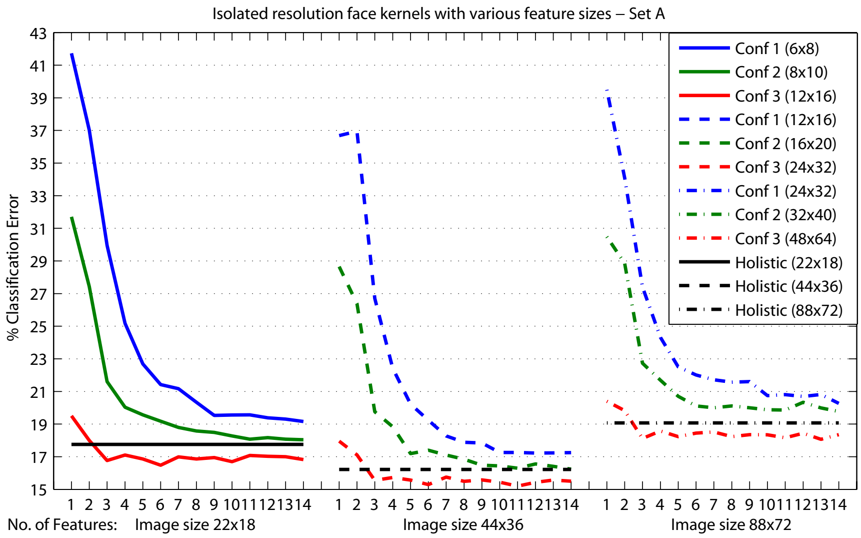
Classification of a new unseen test face is executed as follows. Firstly, a structure is selected by the user defining the size and resolution of face kernels to be used for classification. For example, choosing to combine 8-feature face kernels from level 1 and level 2 and indicated in short as $MRFK_8 - 12$. This structure is used across all the trained batches. For each batch the feature areas of the defined face kernel are used to extract the same regions from the test face. The feature regions are now used to construct the test feature kernels and finally the test face kernels specific to that batch. Once a GDA subspace is constructed for the defined face kernel structure the test kernel can be projected into the subspace and compared to projections of the face kernel for that batch. In this way a test face is compared across all the trained batches but only in specific face kernel subspaces specific to the batch. As a result subspace capacity regulation is achieved because each batch subspace is based on a face kernel with a differing structure of feature kernels and differing set of subjects. Classification is carried out using Euclidean distance comparisons. The test face is taken as belonging to the subject having the smallest Euclidean distance across all the training batches. In the following section, recognition performance of varying face kernel structures are evaluated. The effect of varying the number of feature kernels and resolution in the MRFK representation are investigated using the two data sets described above.

5.5 Results

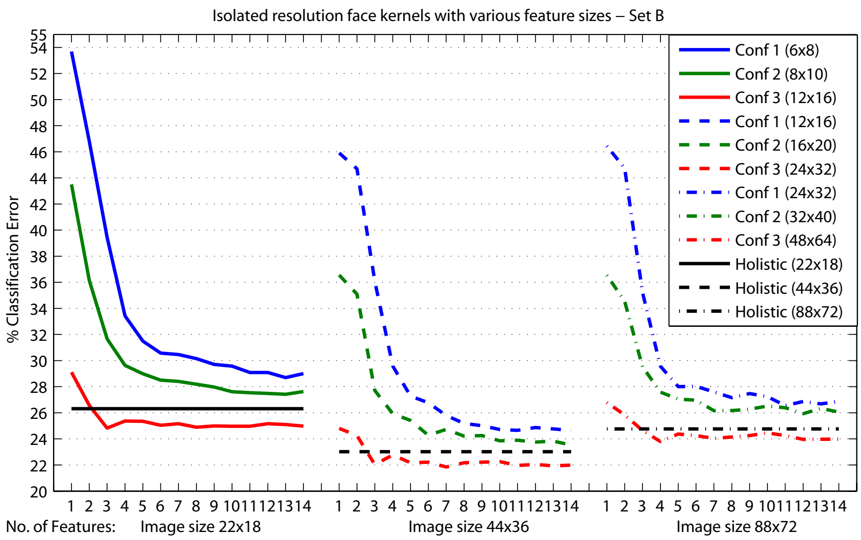
Three face kernel structures are assessed in a classification test. Firstly face kernels of single resolution and feature size are employed to learn data sets *Set A* and *Set B*. This test indicates how performance of face kernels is affected by resolution and feature size. The effect on classification error using multiresolution face kernels is also assessed. Finally the hybrid multiresolution face kernel is assessed by finding good combinations of whole face image kernels and feature based face kernels at varying resolutions. With the methods described above face classification using these various face kernels can be applied and the results are discussed below.

5.5.1 Single Resolution Face Kernel Performance

With the use of single resolution face kernels the effect of resolution and feature size on classification error can clearly be seen. Figure 5.2 show classification error for single resolution face kernels based on feature and image size configurations in Table 5.1. Classification error is plotted against face kernel size (number of



(a) *Set A*



(b) *Set B*

Figure 5.2: Classification error performance using single resolution face kernels with data set (a) *Set A* and (b) *Set B*. Black lines indicate average error rate using a holistic kernel approach at each resolution.

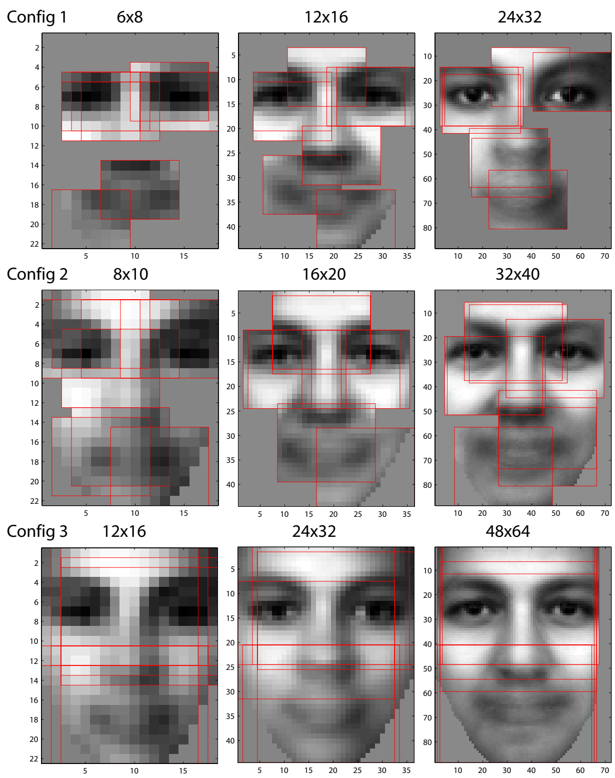


Figure 5.3: Some typical face kernels at varying resolutions. The face kernels are constructed from combination sets of eight feature kernels, S_8^d. Each column is of fixed image resolution with varying feature sizes at each row.

feature kernels), and for comparison the average error rate for the holistic face kernel is plotted in black for each image resolution. The results show that mid range image resolutions perform best, having the lowest error rates. Classification at low resolution may be hindered by low entropy and the lack of discriminating information while high resolution may be hindered by high frequency noise and low signal-to-noise ratios. A trend can be seen where the larger feature sizes of Config 2 and 3 produce lower classification errors. This means that the incorporation of larger areas of salient features improves classification. This may indicate that recognition based on tightly defined facial features with little overlap is not necessarily ideal. Classification with small feature areas is greatly affected by feature misalignment (in the test images) as salient pixel and texture information is lost outside the feature boundaries. As feature window sizes increase incorporating larger feature areas, and perhaps several sections of a couple of facial landmarks, the effect of small feature misalignments is reduced. Now any salient pixel/texture information which may have been shifted (misalignment in test images due to incorrect image registration/normalisation or unseen pose views) may still be within the feature window, provided the window size is sufficiently large.

Some typical features selected by the new kernel criterion feature selection algorithm can be seen in Figure 5.3. Face kernels of size eight (feature kernels, S_8^d) can be seen at varying resolutions. Feature selection is similar to that obtained using the subspace discriminant criterion (Figure 4.8), with feature selections typically incorporating facial landmarks. Small feature kernels (Config 1) are usually tightly placed over facial features and landmarks. This is also true for the intermediately sized feature kernels (Config 2) which also incorporate more of the surrounding areas of features with more overlap. Feature kernels of Config 3 are even larger and nearly half the size of the whole face image. These feature kernels typically divide the face into an upper section (containing the eyes and the nose bridge) and lower section (nose tip and mouth). Feature sizes from Config 3 make the face kernels robust to pose variations in the data set as can be seen from the classification results.

A second trend is also evident from the results (as was noted in Section 4.5.2) where classification error decreases with an increasing number of feature kernels used to construct the face kernel. The additional feature kernels gradually become redundant and improvements to classification error decrease. The rate at which feature kernels become redundant can be used to select a suitable face kernel size (in terms of feature kernels). Each feature size configuration has a different rate with Config 3 requiring only 3 features to achieve a low error rate. This can be seen across all the image resolutions. Similarly face kernels based on Config 2 and Config 1 require 8-10 and 9-10 feature kernels respectively for good error rates.

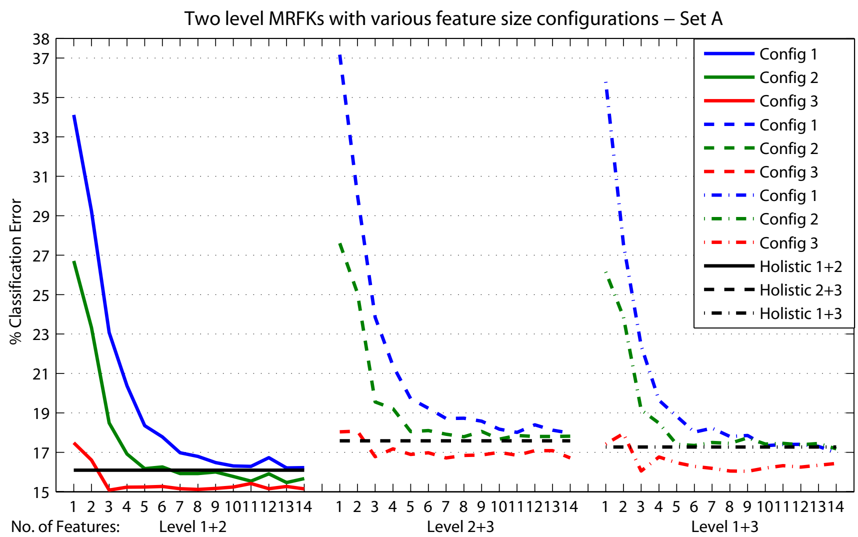
Comparing results obtained using Set A and Set B shows an increase in error rates due to the incorporation of illumination change. The results for Set B also show that mid range resolution face kernels improve performance to a greater extent than when compared with Set A. This further strengthens the reasoning that lower resolution images are less affected by high-frequency noise (in the form of illumination changes in Set B).

In conclusion it is shown that mid range resolution kernels perform best especially with illumination noise. Comparing to a standard holistic approach using GDA (black lines in Figure 5.2), the face kernel has comparable performance with mid range feature sizes (Config 2) and improved performance with larger feature sizes (Config 3). Hence the face kernel constitutes a good representation for use in classification of faces showing a good degree of robustness to pose change. The face kernels applied here have been isolated for one resolution, and in the next section the multiresolution face kernels (MRFK) are applied to the test data and their performance is evaluated.

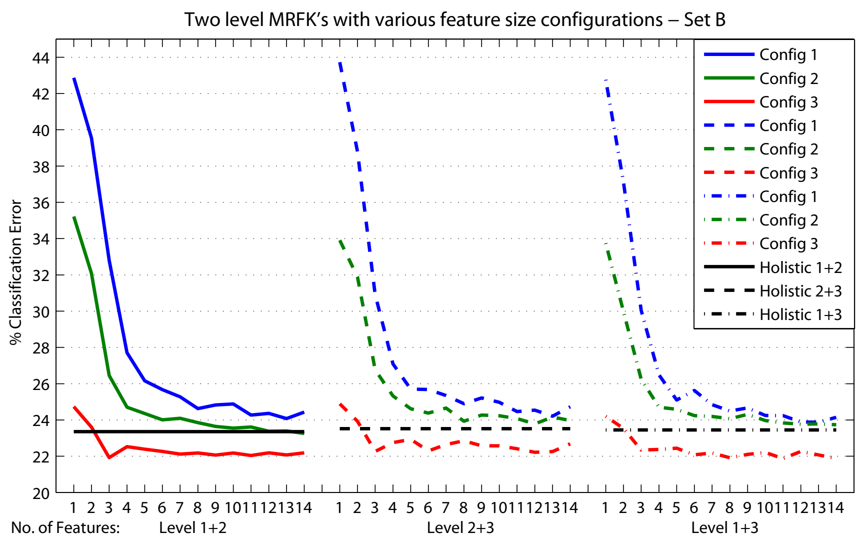
5.5.2 MRFK Performance

By combining face kernels of various resolutions MRFKs can be constructed for use in classification. A set of two and three level MRFKs are used to construct a multiresolution representation to be used for classification. Figure 5.4 shows classification error performance for the two level MRFKs using (a) *Set A* and (b) *Set B*. As with the single resolution face kernel performance the MRFK is similarly affected by feature size and number. The combination of various resolution face kernels produces corresponding error performances. MRFKs constructed using level 1 and 2 resolutions will be henceforth referred to as MRFK-12, and MRFK-23 for the combinations with level 2 and 3, etc. MRFK-12 is the best performing out of the possible three two-level face kernels with the two-level MRFKs having an overall improvement in error rate compared to single level face kernels. MRFK-12 combines a low and mid range image resolution, those where single level face kernels perform best, to achieve a good representation with low error rates. The degrading effect of high resolution on error performance can be seen when comparing MRFK-23 and MRFK-13. The combination of the two higher level resolutions (level 2 and 3) produces an inferior MRFK in terms of classification error. This can be accounted for by the usual high frequency noise which is even more emphasised due to the combination of the higher resolution levels. MRFK-13 performs better than MRFK-23 and, in the classification test using *Set B* and Config 3, equally with MRFK-12. For *Set B* error performance is improved over error rates for the single level face kernels. Error rates between the different MRFK combinations are very similar, with MRFK-23 being only slightly worse. The similar response from the MRFKs suggests that the multiple resolution representations are important especially when illumination variations are present. MRFKs with good performance will generally include low resolution representations (eg. level 1 face kernel).

Comparing feature size configurations of *Set A* and *Set B* shows an improved error rate of Config 3 between Config 1 and 2. Hence the smaller feature sizes



(a) *Set A*



(b) *Set B*

Figure 5.4: Classification error performance using multiple resolution (2-level) face kernels with data set (a) *Set A* and (b) *Set B*. Black lines indicate average error rate using a holistic kernel approach for each resolution level combination.

(Config 1 and 2) are doubly affected by feature movements and sensitivity to illumination noise in higher image resolutions (level 2 and 3). The use of two-level MRFKs can therefore be beneficial as has been shown from the results. This is not true for the all MRFKs and a good selection of feature size and image resolution must be made. Based on results at this point and the available selection of feature sizes and image resolutions, an increase in performance of three-level MRFKs is not expected. The inclusion of the higher resolution representations (level 2 and

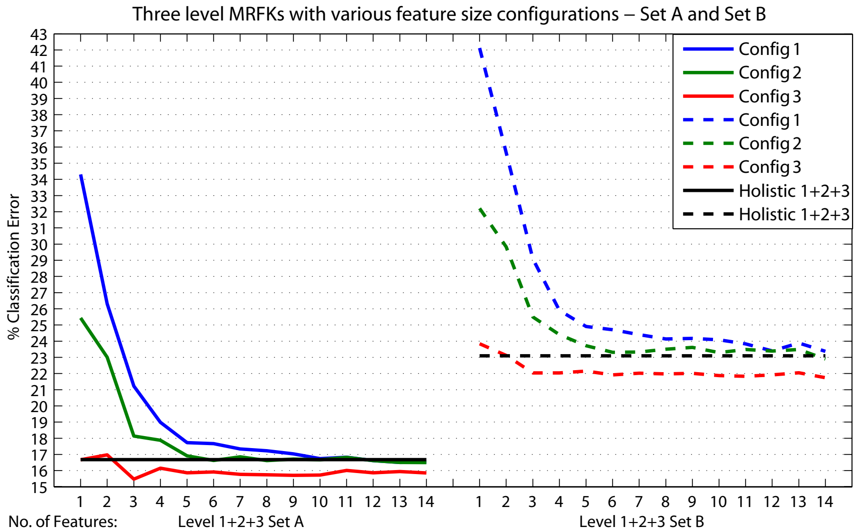


Figure 5.5: Classification error performance using multiple resolution (3-level) face kernels with data sets *Set A* and *Set B*. Black lines indicate average error rate using a holistic kernel approach for the resolution level combination.

3) will have an adverse effect on the lower resolution MRFK. Results for the three-level MRFKs are shown in Figure 5.5. In terms of the lowest attainable error rate using Config 3, no significant improvement is gained. Only a slight improvement is evident for *Set B* where error rates fall just below 22%. Performance with *Set A*, however, is reduced compared to MRFK-12 with error rates now being above 15%. The full multiresolution representation is therefore slightly more robust in the case where illumination noise is present. Smaller feature sizes (Config 1 and 2) show a general improvement between two and three-level MRFKs, approaching error levels close to the holistic kernel baselines. This trend can also be seen between the single and two-level MRFKs. The use of multiresolution therefore improves these small feature kernel representations with the exception of MRFK-12.

Overall face kernels based on Config 3 perform consistently better than the baseline achieved using a holistic kernel. In conclusion MRFKs can be beneficial for representation, although care must be taken when selecting which resolution levels and feature sizes are to be used.

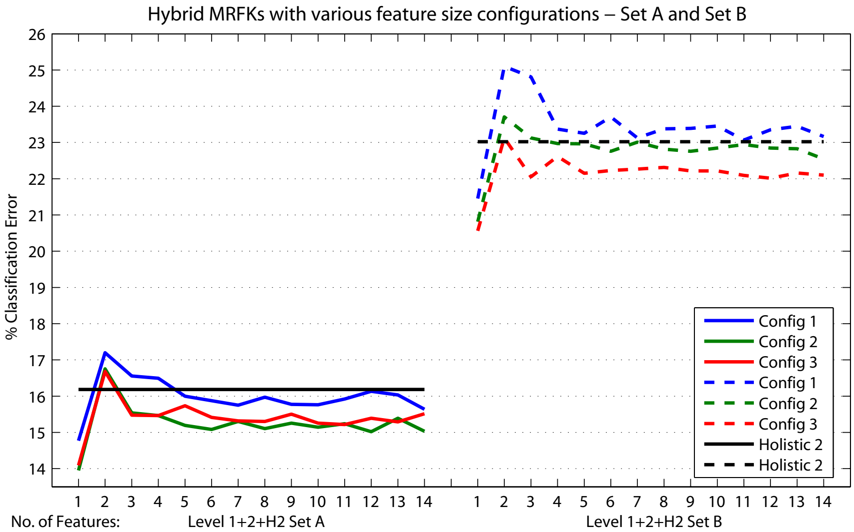


Figure 5.6: Classification error performance using a hybrid MRFK with data sets *Set A* and *Set B*. Black lines indicate average error rate using a holistic kernel approach for resolution level 2.

5.5.3 HMRFK performance

In Section 4.6 the hybrid MRFK was introduced. The HMRFK structure can be quite flexible and many combinations of feature based face kernels and holistic face kernels are possible. Figure 5.6 shows the classification performance of HMRFK-12h2 which combines feature-based face kernels from level 1 and 2 with the holistic face kernel from level 2. Other HMRFKs using level 1 and 3 and any other combinations are also possible but have been omitted as they do not produce good results.

The performance curves shown in Figure 5.6 are characteristic of the general trend in error rate vs. feature set size for the hybrid model. In general it is observed that HMRFKs produce their lowest error rates at the first instance when only the first ranking features from any feature-based face kernels (MRFK-1 and MRFK-2 for HMRFK-12h2) and a holistic face kernel are combined. Further combinations of holistic face kernels with MRFKs of increasing order, however, initially degrades performance quite dramatically and then steadily improves. Although error performance is degraded with the use of higher order MRFKs, the overall error rate is still generally lower than that obtained using a non-hybrid MRFK.

Considering only the face kernel feature sets of more than one feature the following observations are made. In the case of *Set A* (Figure 5.6(a)), Config

1 and 2 have been improved compared to MRFK-12 with Config 1 having a minimum of 15% error from feature set S_6 and onwards. However for the larger feature size, Config 3, the error rate is slightly increased just over 15%. Similar improvements are seen with *Set B* (Figure 5.6(b)) although error rates for Config 1 and 2 have not dropped as significantly as those for *Set A*. Compared to the MRFKs, the HMRFKs using smaller feature sizes have improved with Config 2 now falling bellow the baseline error rate of 23%. The larger feature size HMRFK (Config 3) has not been affected by the hybrid representation and maintains the lowest error rate of 22%.

Alternatively by ignoring higher order feature sets the lowest error rates can be achieved, 14% for *Set A* using Config 2 and 20.5% for *Set B* using Config 3. The HMRFK can therefore produce improved results albeit using only first order face kernels. This effect is however not fully understood. At first it may be expected that the combination of face kernels of increasing order will further reduce error rates. This is unfortunately not realised and is discussed further in Chapter 6.

5.6 Conclusions and Discussions

In this chapter the multiresolution and hybrid face kernels have been evaluated. Because the number of identities and image samples in a good evaluation data set are typically large a number of approaches have been proposed to reduce processing time and memory requirements. As a first step a large data set is divided into a number of smaller batches which are processed separately with each having its own face kernel structure. In this way reducing the final size of the kernel matrix used in processing. Based on the data set size and number of images per identity in the training set, there is a ten fold decrease in the size of the kernel matrix when using batch processing. This is a considerable difference and in turn allows faster processing and less memory requirements.

There is however an adverse effect on classification performance. Classification of an unseen image requires it to be compared with each batch of faces and hence each face kernel. As discussed in Section 5.1 batch processing the data set in this way requires the construction of face kernel representations with a lowered capacity. A certain degree of overfitting of the kernel subspace is required to prevent classifying an unseen image as belonging to an identity in an incorrect batch. However overfitting also means that the face kernel models may fail to perform well on unseen data variations due to the strong bias towards the empirical data used in the learning task. The variations are not only variations of identity but also within-class variations such as pose, expression and illumination. Overfitting therefore can reduce the ability of the face kernel model to accommodate for such within-class variations. The use of a non-linear discriminant learning algorithm in this case can reduce the overfitting risk while optimising a representation based on within-class variations. Constructing an overfitted subspace is not the only way to ensure a 'tailor made' representation for a specific identity. The face kernel structure and its feature kernel sets constitute a batch-specific feature framework. The specific selection of features and their inherent structural framework is optimised for representing the set of faces in the batch. A good classification requires a good match between the learned features of the face kernel and those found in any test image. Another method of enhancing the specificity of a face kernel can also be achieved by tailoring the set of faces used in a batch. These and other enhancements are discussed further in Chapter 6. A second approach was also used in this chapter to further reduce processing time. A kernel criterion was proposed to directly evaluate the effectiveness of feature kernels and their combinations. This avoids the process intensive construction of a discriminant subspace for evaluating each combination of feature kernels in the face kernel extraction process. With the above mentioned modifications an average reduction of 99.8% ($\sim$ 800 times) in processing time is gained for the training set described in Section 5.3.

Apart from the computational gains the kernel criterion is also effective for good feature selection. The kernel matrix represents the projections of the data into the non-linear feature space. Hence evaluating and manipulating the kernel matrix directly can improve the final non-linear projection and hence the learned face kernel subspace. From Figures 5.1 (and A.1 in the Appendix) it can be seen that the kernel criterion responds to feature regions in a similar manner as the subspace discriminant criterion (Figure 3.5). Furthermore, selected feature sets using the kernel criterion (Figure 5.3) are also similar to those chosen by the subspace discriminant criterion. Although no a priori knowledge is afforded to the face kernel model the extracted features are consistently those of facial landmarks. This is especially true for feature sets of low cardinality, and as feature set sizes increase other facial features are included. Taking into account the quality of the extracted face kernels (in terms of the types of features) and the performance results in Section 5.5 the kernel criterion is accepted as being a sufficient replacement criterion.

The concluding sections of this Chapter have applied the face kernel for learning and recognition of face images. A number of MRFK and HMRFK models have been evaluated and have shown favourable results. Firstly in terms of extraction of salient features to form the face kernel and secondly in classification results. By exploiting the property of kernel combination a number of features can be easily unified so as to optimise the kernel representation. This is seen in Section 5.5 where classification error is reduced as larger sets of feature kernels are combined. With a suitable selection of feature size and image resolution the face kernel can deliver improved recognition performance over a holistic approach. Additionally the need for a separate classifier for each feature is avoided as only one discriminant subspace is constructed for the combined face kernel. The face kernel and its discriminant subspace represents a unified and compact data driven model. A multiresolution approach was also proposed in Chapter 4 to extend the face kernel model. The use of information from multiple resolutions can be beneficial. By incorporating low-frequency (identity specific) information and high-frequency (discriminant) information an improved representation can be achieved. This is clearly shown in Section 5.5.2 with MRFK-12 producing improved classification results over a single face kernel. The choice of combinations of feature sizes and image resolutions may be specific to the object being recognised and the quality and variations in images of that object. For example, two and three-level MRFKs on data set *Set B*, containing illumination variations, performed equally well. For data set *Set A* without illumination changes this is not the case, MRFK-13 and MRFK-23 are inferior to MRFK-12. The inclusion of the higher resolution features is believed to introduce high frequency noise which will degrade the representation. However for *Set B* the higher resolution face kernels are needed to form discriminant representations of the high-frequency illumination noise. This is seen where MRFK-13 and MRFK-123 are slightly improved over MRFK-12 and MRFK-23. More generally the MRFKs tend to improve classification error for the smaller feature sets and smaller feature sizes. To further extend the face kernel a hybrid representation is also proposed. The HMRFK produces the best results on both data sets by combining holistic face kernels with the first feature set from a MRFK. HMRFKs of feature sets with higher cardinality also improve classification error over the non-hybrid MRFKs.

Overall the MRFK and HMRFK models perform well and can adequately represent the data set of faces. Experimental evidence indicates that better results are achievable by using larger batch sizes however this gain is opposed by a disproportionate loss in computational speed. Considering the size of the data sets used and the broad number of variations the error results obtained are loosely comparable to recent face recognition approaches. The following section gives a brief discussion of face recognition approaches with similar aspects to the proposed work of this thesis. The performance of these and other face recognition systems is also briefly compared as much as possible.

5.6.1 Comparison

In Chapter 2, a number of feature-based face recognition approaches were briefly described. A comparison is now made based on classification performance between some of these approaches and the work presented in this thesis. Direct comparisons are difficult, however, as the choice of experimental methodologies and data sets reported in the research literature varies largely. Many factors can effect the reported performance of a model such as the quality and preprocessing applied to the images in a data set. Of major concern is the size of the data set and the choice of training and testing protocols used in the reported research literature. Many use small data sets collected under conditions relevant to the aspects of the problems that they examine. Some data sets sometimes exhibit extreme variations in pose or illumination and others are more general with larger numbers of images.

Another point of variability is the amount and type of preprocessing that is applied to face images. An area frequently overlooked in this respect is the proper extraction of the facial region. It has been shown that statistically based methods should only consider internal areas of the face, removing the presence of the background and head hair [24]. This is important especially for holistic and appearance-based approaches as these areas can provided misleading results. The external and non-facial features tend to be highly variable causing the learned face representation to be based on these features and not those specific to faces. Due to these inconsistencies only approaches which are relevant and that have used larger data sets will be discussed.

One approach to the feature selection problem has been proposed in [123] and [127]. Gabor kernels of 5 scales and 8 orientations (at each pixel position of an image) are used to extract a total of 64800 intra-personal and extra-personal features. The goal is to find a combination of these features to form a classifier. This is achieved using AdaBoost to learn a boosted combination of weak classifiers into one strong classifier. Frontal images of 1196 identities from the 'fa' subset of the FERET database are used as the training set while the 'fb' subset is used as the testing set. With combination classifiers of 6 and 8 selected features a correct recognition rate (CRR) of 37.5% and 76% is achieved, respectively. These results are comparable to those obtained with the MRFK on data sets *Set A* and *Set B*. The MRFK performs substantially better (85% CRR) using 8 features with *Set A* on the basis that Gabor features in the above example afford a degree of invariance to intensity due to global illumination. With *Set B* the MRFK results are on par with the above example even though other gross variations such as pose and expression are present in the data set. When a very high number of 700 boosted Gabor features are used an improved result of 95.2% is achieved. The improved results obtained with 700 features however constitutes an overfitted system. AdaBoost selects features which are weak classifiers, i.e. those that do not generalise or classify well and are specific to individual identities. The whole data set can then be represented by combining enough features of this type. A similar approach is reported in [45] using local rectangular filters. Such a system is not easily generalisable and is highly biased towards the training data and may not accommodate unseen variations well. Additionally, the data set used in this case contains no gross variations in appearance and it is questionable if their approach could generalise to unseen views. In contrast the MRFK representation can adequately cope with pose and illumination variations requiring only 8 features to achieve 78% CRR on *Set B*.

Another feature selection approach is proposed in [38]. SFS and SFFS is used to find combinations of local Gabor feature jets for classification. Using 146 identities from the FERET data set with expression and illumination variations a CRR of 91.07% is achieved using 15 features. A genetic algorithm (GA) is also used to find a selection of features achieving a CRR of 96.5% with 15 features. In this example the selected Gabor feature jets are concatenated to form a single feature vector for representation and classification. This is unlike the previous example, employing a weighted combination of individual classifiers. Provided that no large feature variations occur, nearest neighbour matching of the combined feature vectors is very effective. This is seen from their results using a data set containing no pose variations and a good selection of invariant features. Both the SFFS and the GA algorithms select features over the upper face region and it is argued that this results largely due to the expression variations present in the data set, especially in the mouth region. They find that features points around the outer corners of the eyebrows, forehead region, and the outline of the nose provide the most useful information. Hence their approach, which is essentially a template matching approach, consist of matching a set of invariant features. The MRFK doest not perform as well on *Set A* and *Set B*. However, it is not clear whether the former approach will perform as well on a more broadly varying data set. In terms of feature selection, similar choices are made for the MRFK but mostly include features which exhibit small within-class variations. This is seen where the MRFK includes features from the lower facial regions (Figure 5.3).

Both the above approaches have employed methods which learn an optimal selection of features based on the data. An alternative approach can use a selection of predefined feature points such as in the face graph used for the EBGM approach. In [22] a set of Gabor jets are obtained for 16 facial points which are automatically extracted using a feature detector. Corresponding features are compared using a similarity measure and the similarity between any two face images is the average of the similarities between their features. A number of training and testing scenarios are evaluated using a data set of 150 identities with pose variations between $\pm 60°$. Good performance is reported where the testing images have small pose disparities compared to the training set, $\sim 95\%$ CRR. While with larger pose differences ($> 25°$) and expressions changes between training and testing examples the performance is worse, $\sim 63\%$ CRR. These results are good, although they only indicate performance for each scenario with partitioned sets of train and test samples based on pose angle. For example, a CRR of $\sim 95\%$ is obtained when comparing test samples with poses of $15°$ with training samples of $0°$. Therefore the testing set in each scenario is small and hence the reported recognition results are high. A similar testing scenario is employed in [60]. In contrast, the testing scenario used in Section 5.4 for the MRFKs involves learning a random selection of samples which may not always contain examples of pose variations. Recognition of the remaining test samples which may include unseen pose, expression and illumination examples is therefore more difficult. The

MRFK representation does however produce good results, 85% CRR on *Set A* with pose and expression variations.

In [116] a feature-based approach is proposed which employs a similar kernel representation for features as outlined in Sections 2.4 and 4.1. A data set of high quality computer graphic renderings of laser-scanned faces from 100 identities is used. The data set consists, for each identity, of 13 views spaced 15° from left to right profile. The test set is chosen so that its views are between training views. In this way classifiers have to recognise faces under a maximum rotation of 15°. Good results are achieved in this case as the pose variation space is highly specified and the test samples are spaced within the training samples, 98.8% CRR using a summed correlation between best matched features and 100% CRR using an SVM classifier. Again this approach is based on template matching and requires a large data set of examples across the variation space. It therefore may not generalise very well with unseen variations or with larger pose difference from any in the training set. Although a similar kernel-based feature representation is employed in the MRFK, all variations are learned using a single subspace representation model. This means the model can smoothly and continuously cover the parameter space and furthermore requires only one classifier. This is shown in Section 5.5, good CRR results are achieved on *Set A* and *Set B* considering that only three random samples are chosen for learning the large variation space.

A final comparison is now made with a state of the art holistic approach employing discriminant analysis. In [120, 121] a comparison is made between the holistic approaches using KPCA, KDA and their Completed KFD (CKFD) on data set *Set B* (described in Section 5.3). The approaches employing KPCA and KDA achieved error rates of 70.57% and 26.57% respectively. Their improved CKFD approach achieved a minimum of 18.02% which they later improved to 11.62% in [121]. The MRFK representation outperforms the holist KDA approach (shown in the result in Section 5.5). The feature-based kernel representation is therefore a better representation and is directly comparable to the holist kernel representation. KDA is used in both approaches to learn the subspace model. Although the improved holisitc CKFD model does perform better than the MRFK-KDA model, by 10%, this may be due to the improved discriminant representation. By employing the CKFD on the MRFK representation an improvement may be gained. This is left for future work.

In the above a number of studies have been discussed. It is evident that a direct and easy comparison of results is not always possible. However it can be seen that the methods presented here yield results at least comparable with other studies. They also demonstrate the outstanding challenge of face recognition.

Chapter 6

Conclusion and Future Work

In this thesis a face recognition problem has been considered. In order to develop a successful face recognition system a number of considerations were outlined in Chapter 1. It turns out that the range of variations in face images is large and it is these variations that pose the greatest obstacles in developing a successful system. In dealing with some of these obstacles a number of desirable requirements for such a system were discussed. Traits such as robustness, generalisability and compactness motivated a number of novel approaches to the face recognition problem. As a first step a non-linear kernel representation is used at the localised image level. In this way complex and non-linear appearance based image variations can be adequately represented. By taking advantage of the 'kernel trick' discussed in Chapter 3, learning a non-linear representation of the input data is reduced to learning a linear representation in a non-linear feature space. These localised kernel representations are called feature kernels as they represent fixed features or feature regions across the learning data. By employing discriminant analysis in the feature space, intrinsic or within-class variations (such as pose and expression) are used to construct a suitable subspace model. It was shown in Chapter 3 that feature kernels of internal facial regions and especially facial landmarks are more useful for building such subspace models, i.e. those features which are most discriminant. This is achieved without affording any a priori information about faces to the learning algorithm. Using a modular approach it was also shown that a collection of feature kernel subspaces can be used for recognition similar to the modular eigenfeatures approach of Pentland et al. [94]. The feature kernel can therefore be considered as a suitable non-linear appearance-based representation of features that can produce discriminant subspace representations.

A major requirement of the proposed system was that it should be compact and generalisable. In Chapter 4 it was shown how feature kernels can be combined to form a compact representation. Only one subspace representation is learned for any combination of feature kernels unlike the plurality of modular feature kernel subspaces. This also means that the learned subspace can generalise well to variations between feature sets. A modular approach is rigid and classifies a face by considering each feature individually. The combined feature kernel subspace is more general and learns a combined representation of the features. Hence classification considers the whole representation of the selected features. A good selection of features is therefore important for such a representation.

Typically, systems which use any localised information also require a suitable selection of feature points to represent. Most common approaches to this problem involve the manual selection of feature positions deemed as being useful. Another approach is to choose feature points that form a rectangular grid over the entire facial region. One of the requirements of the proposed system is the ability to automatically learn an optimal selection of feature points from the data. This was shown in Chapter 4 where a feature selection algorithm is used to find optimal combinations of feature kernels to construct a face kernel model. Learning directly from the data in this case makes the feature kernel model flexible and can be applied to other object learning tasks. When applied to face images it was shown that the selection of features by the model closely resembles an intuitive selection and empirically recognised as being the most salient and relevant in face recognition. This suggests that the face kernel model may be considered as a biologically plausible model.

Section 4.5 showed that good classification results are achievable using the face kernel model. Classification error of the model steadily decreases as more features are combined in the representation. A multiresolution model was also proposed in Chapter 4 in order to enhance the flexibility and robustness of the face kernel model. By using the flexibility afforded by the kernel representation face kernels from multiresolution images are easily combined to form multiresolution face kernels (MRFKs). In the same manner kernels of whole face images can also be combined with local kernels and the hybrid MRFK (HMRFK) was also proposed.

In Chapter 5, modifications to the proposed feature selection algorithm are described. It was found that feature selection based on the subspace discriminant criterion was very process intensive and worsened when large data sets are used. The computational cost is reduced by dividing a data set into a number of smaller batches, and extracting face kernels for each batch. Additionally employing a kernel criterion (Section 5.2) in place of the subspace criterion further reduces computational load. By applying both these modifications a reduction of approximately 800 times in processing time is achieved without degrading the quality and performance of the extracted face kernel models. In the remainder of Chapter 5, the MRFK and HMRFK models are evaluated on a large data set. It was show that with a suitable selection of feature sizes and image resolutions the face kernel can deliver improved recognition performance over a holistic approach. Error rates of 22% and 20.5% are archived for the MRFK and HMRFK respectively on the data set. As the scope of variations in the data set is large and the number of training examples used is relatively small these results are very promising. Using more examples per subject will tend to improve results as the discriminant representation learns the feature space of all the non-linear variations. It was also shown that the face kernel models performed equally well with a number of similar local feature-base systems. Further modifications to the face kernel model may help achieve better results. These are discussed further in Section 6.1.

6.0.2 Contributions

The contributions to face models for automatic recognition, as introduced in Chapter 1, are now briefly reiterated:

Non-linear discriminant feature representations. Non-linear analysis is applied locally to face image regions. A kernel matrix is constructed from the image region of all samples in the ensemble. As face images in the ensemble are normalised with respect to orientation, the kernel matrices therefore represent feature regions common to all samples. *Feature kernels* are therefore data-driven representations of localised image features in a non-linear feature space.

Learn optimal feature placements. Commonly used feature node placements are relaxed by learning optimal placements. Feature placements are found by learning optimal combinations of salient feature kernels to form *face kernels*. This process is again data-driven with no *a priori* knowledge of face structure being supplied.

Structural framework based on features. A representation of an ensemble of faces is achieved using combinations of *feature kernels* to construct *face kernels*. The *face kernel* inherently incorporates local feature representations and facial structure based on the specific set of *feature kernels* with which it is constructed.

Unified multi-observer representation. A *feature kernel* can be considered as an observer of the face ensemble. Each observer is based on a different image region or feature. The *face kernel* unifies the multi-observer representations into a single representation. A *face kernel* is used to construct one non-linear discriminant subspace requiring only one classifier.

Unified multi-resolution representation. *Face kernels* from varying resolutions can also be considered as multi-observers which can also be easily combined. *Multiresolution face kernels* therefore unify representations of multiple features at varying resolutions.

6.1 Suggested Future Work

As seen in Chapter 5 the MRFK and HMRFK representations obtain a good performance on the data set used. In order to enhance these results a number of proposals for future work are briefly discussed. The first and most easily applicable modification is to use an improved model learning technique. As suggested at the end of Section 5.6.1 the recent CKFD technique can be directly applied to face kernel learning. As this approach improves the discrimination of the learned subspace over that using KFD an improved performance is expected. Other more involved enhancements dealing with the structure of the face kernel and the way the learning data is process are discussed below.

6.1.1 Batch Processing Strategies

The batch processing technique employed in Section 5.1 arbitrarily divides the data set into smaller sections. Although good results are achieved using this approach other segmentation techniques may be more useful. It may be possible to create smaller groups based on their similarity to each other. A clustering algorithm can be employed in a pre-processing stage to create a number of batches which have similar faces. Face kernel extraction then takes place as normal on these batches. This can be used to enhance the specificity of the extracted face kernels and learned model as discriminant analysis will learn to distinguish between similar and hence harder examples of faces. A second approach may use a selective sampling of the entire data set in an one vs. some strategy. In this case a two-class problem is formulated for each batch where one class is the identity being learned and the second class is a representative sample of the remaining data set. Such an approach is very similar to sparse-greedy learning strategies applied in SVM learning [111]. The two-class Kernel Fisher Discriminant approach of Mika et al. [81, 82] can be directly applied in this case. The two-class batch formulation also means that a personalised face kernel is extracted which will distinguish the learned identity from the rest of the data set. This will hence increase the specificity of the extracted face kernels.

6.1.2 Cumulative Multiresolution Feature Selection

In Section 4.3 the multiresolution structure of the MRFK was described. At each resolution level an independent face kernel is extracted. This allows for a flexible combination of the face kernels in the testing phase when no predefined structure has been chosen. A number of combinations may be evaluated to find the best face kernels for the specific data set or classification task combination. However the combination of these independent face kernels may not be optimal with respect to the kernel criterion. An alternative approach requires the prior definition of the MRFK structure to be used before the training phase, i.e. which resolution levels to use and how many features will construct each face kernel. Face kernel extraction then proceeds by incorporating face kernels from previous levels as a fixed feature kernel in the feature selection process. Therefore a MRFK at resolution level 2 will include the face kernel at level 1, similarly the MRFK at level 3 will include face kernels form levels 2 and 1. This ensures that optimal features are extracted when combining the face kernels at increasing resolution levels.

6.1.3 Improving Feature Selection

Although tailoring the kernel matrix representation based on the kernel criterion, and combinations of feature kernels, provides good results, it is not fully understood how its structure affects the quality of the constructed subspace. This is seen in Section 5.5.3 with the HMRFK error performance. Combinations of face kernels of increasing order with a holistic kernel does not achieve the expected steady reduction in error rates seen with the MRFK. More research is needed to understand the nature of the constructed kernels and their counterpart non-linear feature spaces. Recent advances in kernel methods have applied learning strategies to find optimal combinations of different kernels and their parameters on the same data [29, 108]. A similar approach may also be applied using the local kernel representation.

Appendix A

Discriminant Criterion Plots

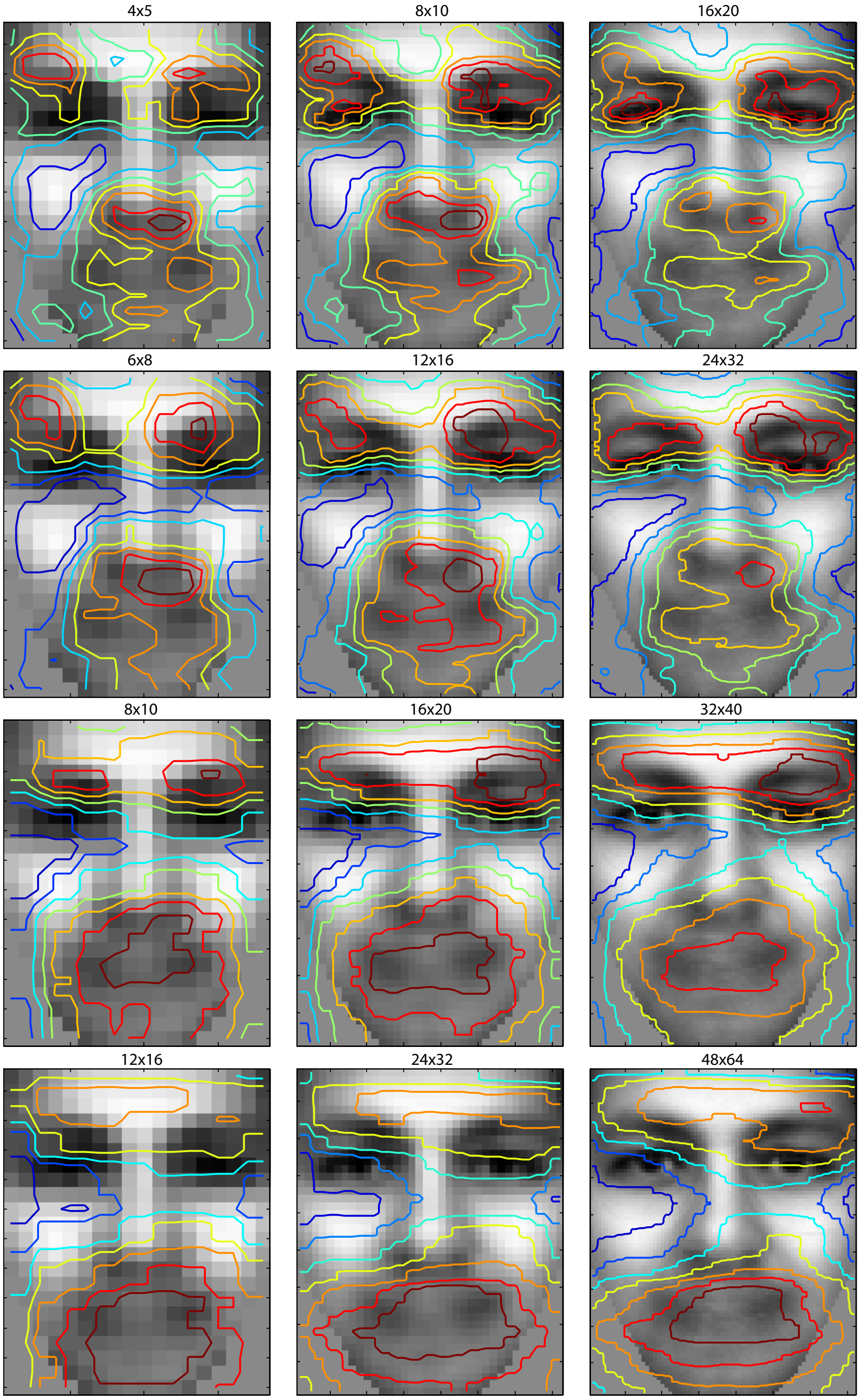


Figure A.1: Kernel criterion responses with varying feature sizes and at varying resolutions. Each column from left to right is based on image resolutions of 22×18, 44×36 and 88×72 respectively. Feature sizes are indicated above each image.

Bibliography

[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[2] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation And Remote Control*, 25:821–837, 1964.

[3] S. Akamatsu, T. Sasaki, H. Fukamachi, and Y. Suenaga. A robust face identification scheme - KL expansion of an invariant feature space. In *SPIE Intelligent Robots And Computer Vision X: Algorithms And Techniques*, volume 1607, pages 71–84, 1991.

[4] F. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20:1361–1369, 1999.

[5] N. Arad, N. Dyn, D. Reisfeld, and Y. Yeshurun. Image warping by radial basis functions: applications to facial expressions. *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, 56(2):161–172, 1994. ISSN 1049-9652.

[6] O. Arandjelović and R. Cipolla. Face recognition from face motion manifolds using robust kernel resistor-average distance. *IEEE Workshop on Face Processing in Video*, 5:88++, June 2004.

[7] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, November 2002.

[8] M. S. Bartlett, H. M. Lades, and T. Sejnowski. Independant component representation for face recognition. In *SPIE Symposium On Electronic Imaging: Science And Technology*, pages 528–539, 1998.

[9] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(1):2385–2404, 2000.

[10] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[11] M. Bischel and A. Pentland. Human face recognition and the face image set's topology. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 59(2):254–261, 1994.

[12] W. W. Bledsoe. The model method in facial recognition. Technical Report PRI:15, Panoramic Research Inc., Palo Alto, CA, 1964.

[13] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory (COLT 92)*, pages 144–152, New York, NY, USA, 1992. ACM Press.

[14] V. Bruce. *Recognizing Faces*. Lawrence Erlbaum Associates, London, UK, 1988.

[15] V. Bruce and A. Young. *In The Eye Of The Beholder: The Science Of Face Perception*. Oxford University Press, Oxford, UK, 1998. ISBN 0198524390.

[16] V. Bruce, P. J. B. Hancock, and A. M. Burton. Human face perception and identification. In H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, editors, *Face Recognition: From Theory To Applications*, pages 51–72. Springer-Verlag, Berlin, Germany, 1998.

[17] V. Bruce, Z. Henderson, K. Greenwood, P. Hancock, A. M. Burton, and P. Miller. Verification of face identities from images captured on video. *Journal Of Experimental Psychology: Applied*, 5:339–360, 1999.

[18] J. S. Bruner and R. Tagiuri. The perception of people. In G. Lindzey, editor, *Handbook Of Social Psychology*, volume 2, pages 634–654. Addison-Wesley, Reading, MA, 1954.

[19] J. Buhmann, M. Lades, and C. van der Malsburg. Size and distortion invariant object recognition by hierarchical graph matching. In *International Joint Conference on Neural Networks*, pages 411–416, San Diego, CA, 1990.

[20] A. M. Burton, S. W., M. Cowan, and V. Bruce. Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science*, 10:243–248, 1999.

[21] A. M. Burton, P. Miller, V. Bruce, P. Hancock, and Z. Henderson. Human and automatic face recognition: A comparison across image formats. *Vision Research*, 41(21):3185–3195, 2001.

[22] P. Campadelli, R. Lanzarotti, and C. Savazzi. A feature-based face recognition system. In *12th International Conference on Image Analysis and Processing*, pages 68–73, 2003.

[23] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[24] L.-F. Chen, H.-Y. M. Liao, J.-C. Lin, and C.-C. Han. Why recognition in a statistics-based face recognition system should be based on the pure face portion: A probabilistic decision-based proof. *Pattern Recognition*, 34(7): 1393–1403, 2001.

[25] J.-T. Chien and C.-C. Wu. Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1644–1649, December 2002.

[26] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–366, 1994.

[27] R. Courant and D. Hilbert. *Methods Of Mathematical Physics*. Publishers Inc., New York, 1953.

[28] I. Craw, N. Costen, T. Kato, G. Robertson, and S. Akamatsu. Automatic face recognition: Combining configuration and texture. In *International Workshop on Automatic Face and Gesture Recognition*, pages 53–58, Zurich, 1995.

[29] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In G. T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 367–373, 2002.

[30] H. D. Ellis. Introduction to aspects of face processing: Ten questions in need of answers. In H. Ellis, M. Jeeves, F. Newcombe, and A. Young, editors, *Aspects Of Face Processing*, pages 3–13. Kluwer Academic Publishers Group, Dordrecht, The Netherlands, 1986.

[31] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America A: Optics Image Science and Vision*, 14(8):1724–1733, 1997.

[32] M. J. Farah, J. W. Tanaka, and H. M. Drain. What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, 21:628–634, 1995.

[33] R. A. Fisher. The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[34] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *13th International Conference on Machine Learning*, pages 325–332, San Francisco, CA, 1996.

[35] I. Gauthier and N. K. Logothetis. Is face recognition so unique after all? *Journal of Cognitive Neuroscience*, 17:125–142, 2000.

[36] I. Gauthier, M. Behrmann, and M. J. Tarr. Can face recognition really be dissociated from object recognition. *Journal of Cognitive Neuroscience*, 11: 349–370, 1999.

[37] B. Gökberk, L. Akarun, and E. Alpaydin. Feature selection for pose invariant face recognition. In *16th International Conference on Pattern Recognition (ICPR2002)*, volume 4, pages 306–309, Quebec City, Canada, 2002.

[38] B. Gokberk, M. O. Irfanoglu, L. Akarun, and E. Alpaydin. Optimal Gabor kernel location selection for face recognition. In *IEEE International Conference On Image Processing (ICIP2003)*, volume 1, pages 677–680, Barcelona, Spain, 2003.

[39] S. Gong, S. J. McKenna, and A. Psarrou. *Dynamic Vision: From Images to Face Recognition*. Imperial College Press, London, UK, 2000. ISBN 1860941818.

[40] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, Reading, Ma, 1992.

[41] R. Gottumukkal and V. K. Asari. An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters*, 25(4):429–436, 2004.

[42] M. A. Grudin. On internal representations in face recognition systems. *Pattern Recognition*, 33(7):1161–1177, 2000.

[43] H. Gupta, A. K. Agrawal, T. Pruthi, C. Shekhar, and R. Chellappa. An experimental evaluation of linear and kernel-based methods for face recognition. In *Sixth IEEE Workshop on Apllications of Computer Vision (WACV'02)*, pages 13–18, Orlando, FL, USA, 2002.

[44] J. Hochberg and R. E. Galper. Recognition of faces: I. an exploratory study. *Psychonomic Science*, 9:619–620, 1967.

[45] M. J. Jones and P. Viola. Face recognition using boosted local features. Technical Report TR2003-25, MITSUBISHI ELECTRIC RESEARCH LABORATORIES, 2003.

[46] P. Kalocsaia, C. V. D. Malsburg, and J. Hornc. Face recognition by statistical analysis of feature detectors. *Image And Vision Computing*, 18(4): 273–278, 2000.

[47] T. Kanade. *Computer Recognition Of Human Faces*. Birkhauser, Basel and Stuttgart, 1973.

[48] M. D. Kelly. Visual identification of people by computer. Techical Report AI-130, Stanford AI Project, Stanford, CA, 1970.

[49] R. Kemp, C. Mcmanus, and T. Pigott. Sensitivity to the displacement of facial features in negative and inverted images. *Perception*, 19(4):531–543, 1990.

[50] K. I. Kim, K. Jung, and J. H. Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2):40–42, February 2002.

[51] K. I. Kim, J. H. Kim, and K. Jung. Face recognition using support vector machines with local correlation kernels. *International Journal Of Pattern Recognition And Artificial Intelligence*, 16(1):97–111, 2002.

[52] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[53] N. Krüger. An algorithm for the learning of weights in discrimination functions using a priori constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):764–768, 1997.

[54] N. Krüger, M. Pötzsch, and C. von der Malsburg. Determination of face position and pose with a learned representation based on labelled graphs. *Image And Vision Computing*, 15(8):665673, August 1997.

[55] L. I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[56] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transaction Computers*, 42(3):300–311, 1993.

[57] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image And Vision Computing*, 13 (5):393–401, June 1995.

[58] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:743–755, 1997.

[59] T.-W. Lee. *Independent Component Analysis: Theory And Applications*. Kluwer Academic, 1998. ISBN 0792382617.

[60] J. Li, G. Poulton, Y. Guo, and R.-Y. Qiao. Face recognition based on multiple region features. In C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, editors, *Seventh International Conference on Digital Image Computing: Techniques and Applications*, pages 69–78, Sydney, Australia, December 2003. Macquarie University.

[61] J. Li, S. Zhou, and C. Shekhar. A comparison of subspace analysis for face recognition. In *28th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 121–124, Hong Kong, China, April 2003.

[62] S. Z. Li, Q. D. Fu, L. Gu, B. Scholkopf, Y. M. Cheng, and H. J. Zhang. Kernel machine based learning for multi-view face detection and pose estimation. In *8th IEEE International Conference on Computer Vision (ICCV'01)*, volume 2, Vancouver, Canada, July 2001.

[63] S. Z. Li, X. Lv, and H. Zhang. View-subspace analysis of multi-view face patterns. In *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)*, pages 125++, Washington, DC, USA, 2001. IEEE Computer Society.

[64] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces in a nonlinear discriminating space. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 258–263, Kauai, HI, USA, December 2001.

[65] S. H. Lin, S. Y. Kung, and L. J. Lin. Face recognition/detection by probabilistic decision based neural network. *IEEE Transactions on Neural Networks*, 8:114–132, 1997.

[66] C. Liu and H. Wechsler. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Transactions on Image Processcing*, 9(1): 132–137, 2000.

[67] C. Liu and H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):570–582, 2000.

[68] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ICA) for face recognition. In *2nd International Conference On Audio and Video-Based Biometric Person Authentication*, pages 22–24, Washington D.C., March 1999.

[69] Q. Liu, H. Lu, , and S. Ma. Improving kernel Fisher discriminant analysis for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):42–49, January 2004.

[70] Q. S. Liu, R. Huang, H. Q. Lu, and S. D. Ma. Face recognition using kernel based Fisher discriminant analysis. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 197–201, Washington, DC, May 2002.

[71] W. Liu, Y. Wang, S. Z. Li, and T. Tan. Null space-based kernel Fisher discriminant analysis for face recognition. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 369–374, Seoul, Korea, 2004.

[72] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):623–627, June 2000.

[73] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Face recognition using LDA-based algorithms. *IEEE Transactions on Neural Networks*, 14 (1):195–200, January 2003.

[74] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 14(1):117–126, January 2003.

[75] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181–191, 2005.

[76] T. Marill and D. M. Green. On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9:11–17, 1963.

[77] A. M. Martínez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[78] T. Maurer and C. van der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In *2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pages 176–181, 1996.

[79] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions Royal Society London*, A209:415–446, 1909.

[80] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, Madison, WI, August 1999.

[81] S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the Kernel Fisher algorithm. In T. K. Leen, T. G. Diettrich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 591–597. MIT Press, 2001.

[82] S. Mika, A. Smola, , and B. Scholköpf. An improved training algorithm for Kernel Fisher discriminants. In T. Jaakkaola and T. Richardson, editors, *8th International Conference on Artifical Intelligence and Statistics (AISTATS '01)*, pages 98–104. Morgan Kaufmann, 2001.

[83] B. Moghaddam. Principal manifolds and bayesian subspaces for visual recognition. In *7th IEEE International Conference on Computer Vision*, pages 1131–1136, 1999.

[84] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[85] H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face recognition algorithms. *Perception*, 30(3):301–321, 2001.

[86] Y. Moses, Y. Adini, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. In *European Conference On Computer Vision*, pages 286–296, 1994.

[87] C. O'Donnell and V. Bruce. Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, 30(6):755–764, 2001.

[88] K. Okada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. van der Malsburg. *Face Recognition: From Theory to Application*, pages 186–205. Springer-Verlang, Berlin, Germany, 1998.

[89] A. O'Toole and S. Edelman. Face distinctiveness in recognition across viewpoint: an analysis of the statistical structure of face spaces. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 10–15, 1996.

[90] Z. Pan and H. Bolouri. High speed face recognition based on discrete cosine transforms and neural networks. Technical report, niversity of Hertfordshire, 1999.

[91] Z. Pan, A. G. Rust, and H. Bolouri. Image redundancy reduction for neural network classification using discrete cosine transforms. In *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)*, volume 3, pages 3149+, 2000.

[92] O. Pascalis, S. D. Schonen, J. Morton, C. Deruelle, and M. Fabre-Grenet. Mother's face recognition by neonates: A replication and an extension. *Infant Behavior And Development*, 18:79–86, 1995.

[93] P. S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[94] A. Pentland, B. Moghaddam, T. Starner, and M. Turk. View-based and modular eigenspaces for face recognition. In *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 1994.

[95] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. Technical Report Nistir 6264, National Institute Of Standards And Technology, 1998.

[96] P. J. Phillips, H. Wechsler, and P. J. R. Jeffrey S. Huang. The FERET database and evaluation procedure for face recognition algorithms. *Image And Vision Computing*, 16(5):295–306, 1998.

[97] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[98] L. Portinale and L. Saitta. Feature selection. Technical report, IST Project MiningMart, 2002.

[99] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Patter Recognition Letters*, 15(11):1119–1125, 1994.

[100] G. Rhodes, S. Brake, and A. P. Atkinson. What's lost in inverted faces? *Cognition*, 47:25–57, 1993.

[101] S. A. Rizvi, P. J. Phillips, and H. Moon. The FERET verification testing protocol for face recognition algorithms. In *3rd. International Conference on Face and Gesture Recognition (FG '98)*, pages 48+, Washington, DC, USA, 1998. IEEE Computer Society.

[102] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[103] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, 25(1):65–77, 1992.

[104] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[105] B. Schölkopf and A. S. K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[106] J. H. Searcy and J. C. Bartlett. Inversion and processing of component and spatial-relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4):904–915, 1996.

[107] G. Shakhnarovich and B. Moghaddam. *Handbook of Face Recognition*, chapter Face Recognition in Subspaces. Springer-Verlag, 2004.

[108] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. ISBN 0521813972.

[109] T. Sim and S. Zhang. Exploring face space. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 5, page 84, Washington, DC, USA, 2004. IEEE Computer Society.

[110] L. Sirovich and M. Kirby. A low-dimensional procedure for the characterization of human face. *Journal Of The Optical Society Of America A*, 4(3): 519–524, 1987.

[111] A. J. Smola and P. Bartlett. Sparse greedy Gaussian pocess regression. In T. K. Leen, T. G. Diettrich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 619–625. MIT Press, 2001.

[112] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:831–836, 1996.

[113] P. Thompson. Margaret Thatcher - a new illusion. *Perception*, 9:483–484, 1980.

[114] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal Of Congnitive Neuroscience*, 3(1):71–86, 1991.

[115] V. Vapnik. *The Nature Of Statistical Learning Theory*. Springer Verlag, New York, 1995. ISBN 0387987800.

[116] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: The kernel recipe. In *9th International Conference On Computer Vision*, volume 1, pages 257–264, Nice, France, 2003. IEEE Press 2003.

[117] H. Wechsler, P. Phillips, V. Bruce, F. Soulie, and T. Huang. *Face Recognition: From Theory To Applications*. Springer-Verlag, Berlin, 1998. ISBN 3540644105.

[118] A. W. Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computing*, 20:1100–1103, 1971.

[119] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[120] J. Yang, A. F. Frangi, and J. yu Yang. A new kernel Fisher discriminant algorithm with application to face recognition. *Neurocomputing*, 56(1):415–421, 2004.

[121] J. Yang, A. F. Frangi, J. yu Yang, D. Zhang, and Z. Jin. KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):230–244, February 2005.

[122] M. H. Yang. Kernel eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods. In *Fifth International Conference On Automatic Face And Gesture Recognition*, pages 215–220, Washinton D.C., USA, May 2002.

[123] P. Yang, S. Shan, W. Gao, S. Z. Li, and D. Zhang. Face recognition using ada-boosted gabor features. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004)*, pages 356–361, Seoul, Korea, May 2004.

[124] R. K. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 81:141–151, 1969.

[125] A. W. Young, D. Hellawell, and D. C. Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

[126] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34(12):2067–2070, 2001.

[127] L. Zhang, S. Z. Li, Z. Y. Qu, and X. Huang. Boosting local feature based classifiers for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 5, Washington, D.C., USA, 2004.

[128] S. Zhao and R.-R. Grigat. Multiblock-fusion scheme for face recognition. In *17th International Conference on Pattern Recognition (ICPR 2004)*, pages 309–312, 2004.

[129] W. Zhao. *Robust Image Based 3D Face Recognition*. PhD thesis, University Of Maryland, 1999.

[130] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *International Conference On Automatic Face And Gesture Recognition*, pages 336–341, 1998.

[131] W. Zhao, R. Chellappa, and P. J. Phillips. Subspace linear discriminant analysis for face recognition. Technical Report CAR-TR-914, Center For Automation Research, University Of Maryland, 1999.

[132] D. Zongker and A. Jain. Algorithms for feature selection: An evaluation. In *International Conference On Pattern Recognition (ICPR '96)*, pages 18–22, Vienna, Austria, August 1996. IEEE.