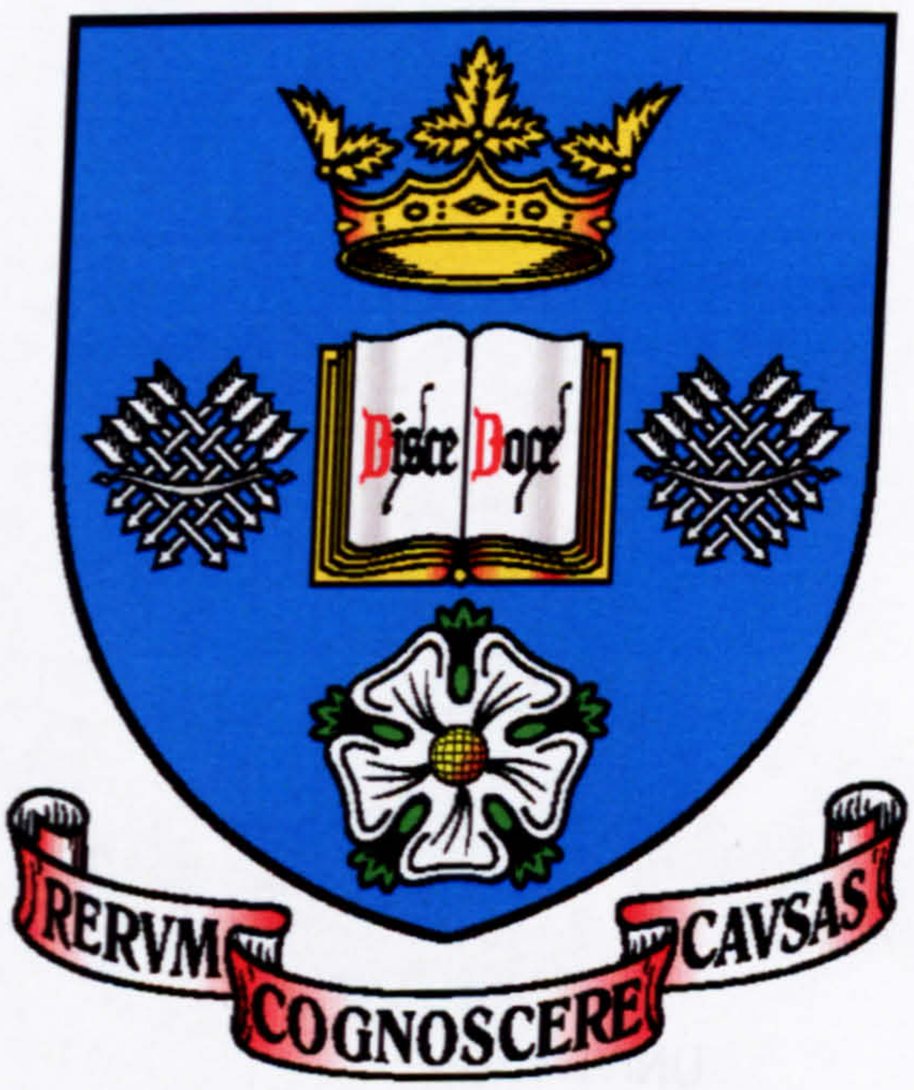
Automatic Face Recognition Using Stereo Images

Anjali Bharatkumar Samani



A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
at
The University of Sheffield
Sheffield, UK
January 2006

To my Dad...

the one person to whom this PhD meant the world and the one person who could not see me accomplish it... To him and to my guru late Mahant Shree Shantidasji Maharaj and late Mahant Shree Ramswarupdasji Maharaj - my guardian angels and my guiding light.

Abstract

Face recognition is an important pattern recognition problem, in the study of both natural and artificial learning problems. Compared to other biometrics, it is non-intrusive, non-invasive and requires no participation from the subjects. As a result, it has many applications varying from human-computer-interaction to access control and law-enforcement to crowd surveillance.

In typical optical image based face recognition systems, the systematic variability arising from representing the three-dimensional (3D) shape of a face by a two-dimensional (2D) illumination intensity matrix is treated as random variability. Multiple examples of the face displaying varying pose and expressions are captured in different imaging conditions. The imaging environment, pose and expressions are strictly controlled and the images undergo rigorous normalisation and pre-processing. This may be implemented in a partially or a fully automated system. Although these systems report high classification accuracies (>90%), they lack versatility and tend to fail when deployed outside laboratory conditions.

Recently, more sophisticated 3D face recognition systems harnessing the depth information have emerged. These systems usually employ specialist equipment such as laser scanners and structured light projectors. Although more accurate than 2D optical image based recognition, these systems are equally difficult to implement in a non-co-operative environment.

Existing face recognition systems, both 2D and 3D, detract from the main advantages of face recognition and fail to fully exploit its non-intrusive capacity. This is either because they rely too much on subject co-operation, which is not always available, or because they cannot cope with noisy data.

The main objective of this work was to investigate the role of depth information in face recognition in a noisy environment. A stereo-based system, inspired by the human binocular vision, was devised using a pair of manually calibrated digital off-the-shelf cameras in a stereo setup to compute depth information. Depth values extracted from 2D intensity images using stereoscopy are extremely noisy, and as a result this approach for face recognition is rare. This was cofirmed by the results of our experimental work. Noise in the set of correspondences, camera calibration and triangulation led to inaccurate depth reconstruction, which in turn led to poor classifier accuracy for both 3D surface matching and $2\frac{1}{2}$D depth maps.

Recognition experiments are performed on the Sheffield Dataset, consisting 692 images of 22 individuals with varying pose, illumination and expressions.

Psychology literature elucidated that although depth information is crucial to the way humans recognise faces, this information is perceived through *disparity* information rather than actual depth in the form of familiar structure and texture information seen in 2D images. Hence, disparity information can be thought of as neural proxy for depth. Computationally too, depth and disparity are proportional up to camera parameters. Disparity values of sub-pixel accuracy are computed by matching the stereo image pairs to solve the ill-posed Correspondence Problem.

A $2\frac{1}{2}$D image representation based on disparity values is proposed in this work, and is shown to give a higher classifier accuracy than the equivalent depth based $2\frac{1}{2}$D images obtained in a noisy environment. Disparity images encode the horizontal and vertical displacement in pixels between the two images. Also proposed is a *composite image* representation, incorporating both 2D texture and 3D shape information. This representation is shown to result in better classifier performance than either 2D or 3D representations individually. This representation also captures the systematic variability arising from the transformation of the facial surface from the 3D world-space to the 2D image-space. The performance of the baseline classifier in this work is lower than the accuracies typically reported in the literature. However, this is primarily because image capture is not strictly controlled and the images themselves undergo no normalisation or pre-processing.

It is concluded that depth information in the form of disparity values plays a crucial role in both human and machine recognition of faces. Disparity information significantly enhances the performance of a face recognition system when it operates in a noisy stereoscopic environment, and particularly when it is combined with texture information. It is also noted however, that in the presence of noisy or inaccurate 3D data, using 2D intensity images and established 2D face recognition algorithms results in more accurate recognition. Reconstructing depth from disparity in noisy environments leads to a loss of discriminatory information, which explains the lack of stereo-based systems in 3D face recognition literature.

Acknowledgements

There are many people whose help and contribution have been vital to the accomplishment of this PhD. First of all, I would like to thank my supervisors, Dr. Joab Winkler and Professor Mahesan Niranjan for providing me the opportunity and funding for this PhD. Their help and guidance has been invaluable. To Niranjan, especially for the loan of his machine and office during the last 12 months.

I would also like to extend a very special thank-you to a few people whom I have never met, but whose willingness to help and answer questions both complex and trivial, time and time again, has greatly enriched my understanding of certain subjects covered in this work. Gabriella Catellano - without her timely help with deciphering and coding of Magarey's algorithm, I would have given up. Dr. Kingsbury, for sending me some of his wavelets code and Magarey's PhD thesis, both of which aided my understanding of the subject. P. Anandan, for his help with aspects of the work in his paper (Anandan 1989). Jiri Walder, for his help with efficiently coding Pan's algorithm. Daniel Hebert, not only for allowing me access to the "Mesh Toolbox" for Spin-Images, but also for his significant help and technical support.

A big thank-you to Dave Abbott and Julian Briggs for all the technical support throughout my PhD, especially when I accidentally deleted important files - like my entire thesis! To Neil Lawrence for the loan of his PC to install the Mesh Toolbox. To my lab-mates: Renata da Silva Camargo, Hongyu Li, Sujimarn Suwnnaroj, Tisanai Krisanamathukul, Tonatiuh Pena and Nat King, for their moral support. To James Carmichael for all the encouragement, but more importantly for putting up with my moaning and complaining!

To everyone who volunteered (or were volunteered!) to feature in the Sheffield Database: Renata, Joab, Stuart, Bala, James, Tonatiuh, Tisanai, Karen, Monika, Lucy, Gillian, Sarah, Andre, Shrifah, Simon, Heidi, John, Francesco, Stephanie, Mike and Ekta. I am particularly grateful to Bala for agreeing to let me use his face images so extensively in all my work.

To my family for their love and support through everything. My parents, for instilling in me a passion for higher education and for giving me the best opportunities, regardless of their own circumstances. To my mum, Chandrika, who never gave up on my education, despite many failures. To my dad, Bharat, for teaching me to be independent and reaching for the stars, for supporting me in all my decisions (like this PhD!) whether he agreed with them or not. To my gran, for her constant belief in the power of positive thinking, love and blessings.

To my big sister Aarti, for being one of my best friends and worst critics - I would never have improved without some of incessant nagging! And for her love and support, for always looking out for my best interests and most importantly, for letting me turn her living/dining room into my work-space. I can only imagine how difficult it was for her to come home and be greeted with mountains of paper all over her tidy space! To my younger brother and sister, Munindra and Radhika, for their love, the laughs and the 'pick-me-up' hugs! And last but not least, to Deepan, for his encouragement, support and unshakeable belief in my ability to finish this PhD. He will never realise how much his patience and understanding have meant.

To Salim Vanak, for his generosity and all the help, advice and round-the-clock technical support - I learnt a tremendous amount from him. In the last few months of my PhD I became a semi-permanent guest at his place and came to rely greatly on his friendship and moral support. Thanks for some very candid advice and enlightening discussions, especially on ROC curves!! To Trudy Glasgow for her impeccable hospitality and for always making me feel so welcome in her flat - even at the shortest notice. To Gordon Booker and Paul Johnson for some of the most brilliant chats and late night cups of tea. The laughs I had with them (especially doing crosswords with Gordon!) and the way in which they always put life into perspective really helped my sanity and made the PhD a really enjoyable experience. To Pranav Patel, for his selfless and unconditional friendship and for keeping me going and making me smile when I thought I could no longer. For that I am truly indebted to him.

A special thank-you also to Reza Gahssemieh and Roger Wilson at Nomura International Plc. for giving me the opportunity to do an internship in Quantitative Analysis under their supervision. I am particularly grateful to them for their patience and understanding, and for allowing me to use a significant proportion of my time and the resources at Nomura for my PhD work.

Contents

CHAPTER 1

Introduction

Face recognition is an important pattern recognition problem in the study of both natural and artificial learning systems. As a pattern, the face is a challenging object to recognise. Anatomically, it is rigid enough so that all faces have the same structure. However, each face is unique due to the shape, size and placement of features (e.g. separation of eyes), gender and race. The problem is made further difficult due to variations that arise as a consequence of changing expressions, illumination, accessories, cosmetics and aging.

Humans demonstrate an impressive ability to recognise faces and are not easily deceived by superficial or cosmetic alterations. Once a face has been learnt, people can recognise it accurately despite not having seen it for years. Often a brief look at a face is enough for correct recall (e.g. witnesses to crime recognising their perpetrators from mug-shots or police line-ups). The ability to recognise faces accurately is essential for both, social interaction and human survival - an infant immediately responds to face shapes at birth and can recognise his or her mother's face from a stranger's at just 45 hours (Voth 2003). It is known that the human brain has specialised subsystems of neural circuits for face recognition rather than general object recognition (Würtz 2002). Brain imaging studies show a great deal of activity in an area of the temporal lobe known as the fusiform gyrus (George et al. 1999), an area also known to cause prosopagnosia (an inability or difficulty in recognising familiar faces) when damaged.

Over the last 15 years or so, face recognition has become an active area of research in computer vision, neuroscience and psychology (Phillips et al. 2000). Since the terrorist attacks of September 11th 2001, biometrics in general and face recognition in particular, have transformed into a lucrative market. There is an increased interest in the field from both government and non-government organisations. In the UK, the first passports equipped with facial biometrics are due to be introduced in February 2006[1]. In the USA, major funding initiatives are being undertaken by Federal agencies (e.g. HumanId project at Defence Advanced Research Projects Agency (DARPA)) to further research in this area[2]. Face recognition tech-

[1]UK Passport Service Press Release "PASSPORTS ARE CHANGING: BIOMETRIC INFORMATION CAMPAIGN LAUNCHED IN MANCHESTER", 12th September 2005, www.ukpa.gov.uk/press_120905.asp

[2]National Institute of Standards and Technology (NIST), "IT Performance: HumanID - Ranking Algorithms for Face Recognition", 21st June 2002, www.itl.nist.gov/div898/itperf/humanid.htm

nology has also been recommended by the International Civil Aviation Organization (ICAO) as the most suitable biometric for Machine Readable Travel Documents (MRTD)[3].

In "traditional" biometric applications such as access-control, face recognition encounters stiff competition from other more accurate biometrics such as iris and fingerprint recognition. In a co-operative scenario where accuracy is paramount and the search set small, face recognition may not be the most appropriate means of identification. However, it is an intrinsic part of identification documents such as driving licenses and passports[3]. The database for these documents is much larger in comparison and subject co-operation is not always readily available. This covert or clandestine capacity of face recognition is its major advantage over other biometrics.

As a means of personal identification, it has the advantage of being remote, quick and convenient. But most important of all, it is totally non-intrusive and can be accomplished without subject participation or knowledge. This latter property has also initiated much debate and controversy with regards to privacy and civil liberties. However, advocates of these arguments often forget that manual face recognition systems are already in operation and have been for many decades. Such systems rely on police or security officials' ability to identify "wanted" criminals or to allow access to restricted areas to "known" individuals. Manual monitoring of CCTV images is widely accepted in most public places. Although human ability to recognise faces surpasses that of most automated systems', this ability varies from individual to individual. Constant performance levels cannot be guaranteed across operators. Human face recognition performance is also limited by the number of faces an individual can accurately remember and identify. In addition, human operators are also susceptible to fatigue, stress and other distractions, hence the motivation for an automated face recognition system.

1.1 Automated Face Recognition

Although the history of automated face recognition dates back to the 1960's, it remains an unsolved problem and still offers a great challenge to computer-vision and pattern-recognition researchers (Li & Jain 2005). The difficulty arises because even though faces are generally thought to be unique, statistically they are quite similar. This, combined with the numerous possible variations in the images of the same face arising due to changes in pose, expressions, illumination, etc. makes this seemingly simple task non-trivial. A great deal of cognitive effort is required to identify faces. Neurologically, in addition to the fusiform gyrus, large, distributed regions of the brain are involved in accomplishing this. Computationally, sophisticated and robust algorithms are required to classify faces - often there is greater variation in images of the same individual from different viewpoints than in the images of different individuals from the same viewpoint (Adini et al. 1997). Face recognition is effectively a within-category discrimination and is a more difficult problem than general object recognition, which involves identifying objects with largely distinct shapes.

Traditionally, face recognition systems have operated in the two-dimensional (2D) space. Most systems use digitised gray-scale photographs of faces (*intensity images*), though some research involving colour images is also found in literature (Torres et al. 1999). Intensity

[3] "World Face Recognition Biometrics Markets", 6th October 2003, www.marketresearch.com, Pub ID: MC935321

images measure the amount of light reflecting from a surface and hitting a photosensitive device (camera). The reflected light enters the camera lens and hits the image plane, where image of the surface being photographed is captured and stored. Thus, they represent an affine projection of an underlying three-dimensional (3D) object on the 2D camera image plane.

Data for 2D face recognition systems is easily available since specialist equipment (other than cameras) is not required. There are also numerous publicly available datasets of 2D face images captured under conditions with varying degrees of control (see (Tolba et al. 2005) for a comprehensive list). The images usually display uniform variation in pose, expressions and illumination. It is usually advisable to use these datasets for evaluating performance of new face recognition systems against existing ones.

The main disadvantage of 2D image based face recognition systems is that they do not exploit the systematic transformations of optical structure that occur when an object is viewed over multiple vantage points. This systematic variability is treated as random, by collecting a number of images per individual, captured with different pose, illumination and expressions. This leads to poor recognition accuracy when the test images deviate significantly from the training images. However, recognition rates in excess of 90% are common for test and training images collected in strictly controlled laboratory conditions.

More sophisticated 3D face recognition algorithms have emerged in the recent years, taking advantage of the depth information lost in 2D images. These systems are inherently invariant to changes in pose and illumination since the 3D shape of the facial surface captures the anatomical structure of the face and are independent of the imaging environment. Depth information can be obtained using specialist equipment such as laser scanners or structured light projectors. Both these methods are very accurate, though laser scanners are considerably more expensive. From a pattern recognition perspective, 3D systems allow better classification due to the availability of additional information. However, from a biometrics point of view, there is a tradeoff between the non-intrusiveness and the accuracy. Such an identification system can only be operated in a co-operative scenario. And since more accurate biometrics exist for this, the 3D technology does not particularly widen the scope of face recognition applications.

3D face recognition is usually approached in one of three ways. The first approach is the *model-based approach*, in which a generic 3D face model is created using the face models of all the subjects in the database. Models may be represented as a shaded model or a wire-frame. The generic model is manipulated to customise it to each individual in the database, either by locating fiducial points on the surface of the model (manually or automatically) or by fusing information from 2D images (usually frontal and profile). The colour or texture information from the 2D intensity images is then overlaid on the customised 3D model to give it a realistic appearance. The depth and texture information are often used in conjunction to achieve better classifier accuracy. Model-based approaches are heavily reliant on pre- and post-processing to achieve uniformity across all the models in the database and are computationally intensive.

The second approach reconstructs the facial surface from depth values for every single individual in the database. The depth values may be obtained from a variety of cues such as stereo images, range images, laser scans, etc. These approaches do not, in general, use the texture information. They rely on the premise that the depth data is already fairly accurate either due to the nature of the data acquisition process, or because the data has been

rigorously filtered to eliminate noise. Surface matching techniques are used for identification.

An alternative approach encodes the 3D depth information directly in a 2D image by replacing the intensity values with the depth values, so that the new pixel values correspond to the surface geometry of the 3D object (e.g. range images, depth maps and surface profiles). Such a representation captures the depth information of a scene, whilst still enabling all the existing 2D image processing and face recognition techniques to be used. There is some debate in the literature as to whether the representation of depth information in this latter approach should be referred to as "3D" or as "$2\frac{1}{2}$D" (2.5D) images. For example, Heseltine (Heseltine et al. 2004*d*,*b*,*c*) refers to his depth maps as 3D images, whereas Fromherz (Fromherz 1996) states that a 3D image is an image or a model that encodes the geometry of an object from 0° to 360°.

In this work, 3D images refer to models that can be viewed in the 3D space either as a point cloud, a polygonal mesh (unshaded wire-frame) or a surface (shaded wire-frame) and do **not** consist of the whole head. Since this dissertation concerns face recognition, rather than general object recognition, it is fair to say that the data corresponding to the back of the head can be discarded without any loss of information. In addition, this part of the human head is subject to constant change due to variations in hair styles and length, and would only serve to confuse the system. The terms "3D model" and "3D image" are used interchangeably and refer to the same entity. $2\frac{1}{2}$D images on the other hand, refer to 2D images encoding depth information.

1.2 Objective

The objective for this work was to investigate the usefulness of depth information in improving face recognition accuracy in the presence of noise.

1.3 Inspiration: Human Binocular Vision

Humans, in most cases, recognise faces using stereo vision. Results of the experiments by Hill et al (Hill et al. 1997) showed that 3D shape information is fundamental for face recognition across rotations in depth, although superficial texture or shading information may also be useful in reducing the viewpoint dependence of the recognition process. These findings are also echoed by Liu et al (Liu et al. 2000).

There are many different aspects of optical stimulation that are known to provide perceptually salient information about 3D form (Todd 2002). Cues such as texture gradients, contour configurations, and patterns of shading are usually available simultaneously within individual static images. However, by far the most important cue is defined by systematic transformations among multiple images, including the disparity between each eye's view in binocular vision and the optical deformations that occur when objects are observed in motion.

In human binocular vision system, each eye receives a slightly different, but overlapping view of the same scene. It is interesting to note that this overlap reduces the size of the overall visual field relative to what would be otherwise possible if the two eyes faced in opposite directions, as is the case with many other animals. For the ecology of human observers, this cost is apparently outweighed by the useful information that is provided by the disparities between each eye's view in the region of overlap (Todd 2002).

Although depth information plays an important role in face recognition, review of some of the psychology literature showed that this depth information can be perceived without familiar structure or texture information (Nakayama 1996). This notion was further reinforced by experiments using random dot stereograms invented by Julesz. He found that *binocular disparity alone is sufficient to mediate perceived depth.* An example of neural mechanisms for such processing by the perceptual visual system is the existence of disparity tuned cells in the striate cortex of cats (Barlow et al. 1967).

1.4 Implementation

Computationally, this process of binocular vision can be reproduced using stereoscopic depth extraction methods. These methods involve computing the geometry of the imaged scene by matching corresponding points in two images. The images capture a static scene from different viewpoints using either a number of still cameras (*Passive Stereo*) or using a single, dynamic camera (*Active Stereo*). The latter is also equivalent to imaging a dynamic scene using a single still camera. Points in the 2D images corresponding to the same scene point are matched (manually or automatically) and their positions (co-ordinates) in the 2D image planes are obtained. Because the images are captured from slightly different viewpoints, there is a *disparity* or a difference in the position of the features between the two images. This disparity is used in conjunction with the camera parameters (focal length, principal point, location in the 3D space, etc.) and the position of the scene features in the 3D space are calculated using triangulation. Stereoscopy is extremely noisy primarily because the task of matching two images and computing disparities is ill-posed and as yet unsolved - a problem known as the *Correspondence Problem.* This is compounded by errors in the estimation of camera parameters and triangulation. Consequently, stereoscopic methods are rarely used for 3D face recognition.

3D face recognition using stereoscopic methods, although noisy, retains all the advantages of face recognition as a biometric. Faces are imaged using ordinary cameras and subject cooperation is not required. The problem is also of great interest from pattern recognition, machine learning and computer vision perspectives. In this work, two digital off-the-shelf cameras are used to capture face images in a partially controlled environment. Subjects are instructed to display varying pose and expressions but the exact degree of variation is not strictly enforced. Such an approach, although not totally non-intrusive, can be extended to operate in an uncontrolled environment if it is sufficiently accurate. Indeed, this is the long-term goal of this work. An accurate face recognition system using stereoscopy has many applications including crowd surveillance and law enforcement. It is also much more in line with how we humans recognise faces.

Experiments in this work showed that stereoscopic depth extraction is noisy and computationally expensive, and unless performed accurately, detracts from rather than adds to the discriminatory information in face recognition. Binocular disparity encodes depth information. Physiologically, disparity is the neural proxy for depth, and mathematically, the two are proportional up to camera parameters. This crucial piece of information is used in this work to produce $2\frac{1}{2}$D images. While existing techniques use the depth values to represent relief information in the $2\frac{1}{2}$D images, this work uses *disparity* information. $2\frac{1}{2}$D disparity images show how much each pixel has moved by between the two images. Much of the noise is thus eliminated by by-passing the noisy camera calibration and triangulation processes,

whilst still retaining all the benefits of face recognition as a biometric. This leads to increased accuracy when compared with $2\frac{1}{2}$D depth images.

Humans use all available information (Todd 2002), and in particular shape and texture (Liu et al. 2000) to identify known individuals. It is now known that a combination of shape and texture information, collected using specialist equipment (such as laser scanners and structured light projectors) in controlled environments results in greater accuracy in automated face recognition systems (Bowyer et al. 2004). With this in mind, a simple yet effective method of combining 3D shape and 2D texture information in the form of a *Composite Image* is adopted, with extremely promising results. Composite images are formed by appending the $2\frac{1}{2}$D disparity image with the 2D image. Thus the information from the two modalities is processed simultaneously, much like in the human visual system. Existing approaches process the depth and texture channels separately and fuse the results at a later stage (Phillips et al. 2005). This approach is also more suitable for the long term goals of this work. It is intended to replace passive stereo (still cameras) with active stereo (moving cameras/scene) initially, and later, embed the recognition module in a fully automated surveillance system incorporating face detection, face tracking and possibly multi-modal recognition.

3D reconstructions are obtained using image matching and noisy camera parameters and triangulation. A signature-based 3D surface matching algorithm, Johnson's spin-image matching (Johnson 1997), is used to perform recognition in the 3D space.

Face recognition in 2D, $2\frac{1}{2}$D and composite spaces is performed using Turk and Pentland's Eigenfaces algorithm (Turk & Pentland 1991*a*) and Spies and Ricketts' Fourier space based nearest neighbours algorithm (Spies & Ricketts 2000). Eigenfaces technique derives a face space that is sensitive to the statistical structure of the faces in the training. It is thought that humans also perform identification in a similar manner. Valentine's (Valentine 1991) face space theory states that the human memory for faces can be thought of metaphorically as a multi-dimensional face space. The average face is at the centre of the space, where it is "crowded" with many faces that are at close proximity to the average face. Distinctive faces are located in the sparser parts of the space, away from the average.

1.5 Thesis Overview

The remainder of the thesis is structured as follows:

Chapter 2: Face recognition literature is reviewed and the main techniques in 2D, $2\frac{1}{2}$D and 3D face recognition are described.

Chapter 3: An overview of the 3D reconstruction process using stereo images is covered in this chapter, along with the merits and drawbacks of the technique.

Chapter 4: A description of the Sheffield Dataset, its main features and the data acquisition process are outlined. The camera calibration process in outlined and the camera matrices are presented.

Chapter 5: A discussion of the Correspondence Problem and a brief literature review are presented. Two feature detection and feature matching algorithms are described and applied to face images. Results and analysis are presented at the end of the chapter.

Chapter 6: Two wavelets-based image matching algorithms are explained. They are used to match face images and the results are analysed both qualitatively and quantitatively.

Chapter 7: Spin-image representation and recognition algorithm for matching 3D objects is detailed in this chapter, along with the results of applying it to recognise 3D face models.

Chapter 8: A mathematical exposition of the Principal Component Analysis based Eigenfaces algorithm and a Fourier space based nearest-neighbours algorithm is presented. These algorithms are used for face recognition in 2D, $2\frac{1}{2}$D and composite spaces.

Chapter 9: This chapter presents the results and analysis of 2D face recognition using the Sheffield Dataset.

Chapter 10: The results of face recognition experiments in the $2\frac{1}{2}$D and composite spaces are analysed in this chapter.

Chapter 11: Conclusions, contributions and suggestions for future research are put forward.

Appendices A and B: An outline of the wavelets theory and details of the Magarey-Kingsbury wavelets respectively.

Appendices C, D and E: Detailed description of two image matching algorithms and the surface matching algorithm used in this work.

CHAPTER 2

Face Recognition Literature Review

2.1 Introduction

In the recent years, face recognition has received significant attention. Some of the reasons for this have already been outlined in Chapter 1. In addition, the technology to carry out the complex tasks in face recognition has only just started becoming more readily available. This, combined with the wider range of commercial and law enforcement applications has also contributed to the increased interest in face recognition.(Zhao et al. 2000).

The literature in this field spans many decades, both in psychology and computer science. In computer science, it also covers many different sub-areas such as face detection, face tracking, feature extraction, etc. The review in this chapter will concentrate only on the major face recognition techniques since it is impossible to cover all the techniques reported in the literature.

Early face recognition techniques relied on manual definition of geometry-dependent features to be used for recognition. These feature values depended on the detection of geometric facial features, including the distance and angles between points such as eye corners, mouth extremities, nostrils and chin top (Weng & Swets 1999). The features defined for the face profiles typically consist of a set of characteristic points on the profile such as the notch between the brow and the nose and the tip of the nose. For example, Kaya and Kobayashi (Kaya & Kobayashi 1972) used Euclidean distances between manually identified points in the images to characterise the faces.

Although this manual definition of features is intuitively understandable, the number of features measurable in this way is small and the reliability of each feature measurement is difficult to estimate (Weng & Swets 1999). In addition, there has been an increasing demand to develop systems that are completely automatic and require no human input. These systems have many applications in fields ranging from graphics and human-computer-interaction (HCI) to law enforcement and access control.

A face recognition system is required to perform *identification*, *verification* or a combination of both these tasks depending on the application. Identification is a multi-class task, where the input image of an unknown individual is matched against a database of known

individuals and an identity label is assigned to it. Verification is an easier, two-class task in which the claimed identity of an individual is confirmed or rejected by the classifier. Alternatively, an appropriate error message is provided if the individual does not exist in the the database. The main issue with the verification problem is setting the appropriate threshold values, based on which the system decides whether or not the input image and the claimed identity match. These values are typically determined empirically, and are dictated by the dataset.

Automatic face recognition techniques can be grouped in many ways, depending on the criteria chosen to solve the problem (Weng & Swets 1999), for example:

Sensing Modality: This refers to the inputs the system accepts. For example, 2D intensity images, colour images, infra-red images, 3D range images or some combination of these.

Temporal Content: This refers to whether the inputs are of a static or a dynamic nature. Static images are those taken at a particular point in time using a digital camera for instance. Dynamic images are time-varying and are produced using a CCTV camera for example. A dynamic system may be an "all inclusive" system and may facilitate face detection, face tracking and face identification.

Geometry & Viewing Angle: Geometry refers to the space in which the system operates: 2D or 3D. 3D systems are inherently view and pose independent, but a 2D system is usually designed for frontal views, profile views, general views or a combination of all of these.

Computational Tools: This refers to the actual technique used to perform recognition. Examples include programmed knowledge rules, statistical decision rules, neural networks, genetic algorithms, etc. Although these techniques and their variants were initially used by themselves, they are now often used in conjunction with each other. These methods are known as *hybrid* methods and they take advantage of the useful primitives from the constituent methods. Techniques can also be divided into feature-based methods and template-based methods. Feature-based methods first compute a set of geometrical features and use these individual features to match faces. Template-based methods use a single template to represent the entire face and use holistic matching techniques. A comparison of the feature and template matching techniques can be found in (Brunelli & Poggio 1993).

2D face recognition techniques are reviewed in Section 2.2. However, most of these techniques apply equally well to $2\frac{1}{2}$D and 3D input data as well. Evaluation methodologies for 2D recognition techniques is discussed in Section 2.3. This is followed by a review of techniques using depth information to recognise faces. Techniques relying on shape information alone, and those relying on both shape and texture are covered.

2.2 2D Face Recognition

This section will provide a general overview of some of the major classification techniques in 2D face recognition. As mentioned earlier, since the subject of face recognition has been around for a long time, there are many other techniques to be found in literature. It is not

possible to cover them all in detail. Good surveys of face recognition literature can be found in (Fromherz 1998, Weng & Swets 1999, Zhao et al. 2000, Lu 2003, Zhao et al. 2003, Kong et al. 2005, Tolba et al. 2005).

2.2.1 Eigenfaces

One of the most popular methods for face recognition, Eigenfaces was pioneered by Turk and Pentland in 1991. It is based on Kirby and Sirovich's (Kirby & Sirovich 1990) proposition to use Principal Component Analysis (PCA) for face analysis and representation. This approach treats face recognition as a problem in the 2D space rather than requiring the recovery of the 3D geometry. It takes advantage of the fact that faces are normally upright and thus may be described by a small set of characteristic views.

The basic idea behind principal component analysis (PCA) is to take advantage of the redundancy that exists in the training set, so that it may be represented in a more compact way. Face images are represented as 1D vectors and are projected onto a feature space that spans the significant variations among known face images. PCA determines an orthogonal space, *Eigenspace*, by computing the eigenvectors of the covariance matrix of the set of vectors. This Eigenspace is an orthogonal basis with the axes ordered according to their overall variance (Troje & Vetter 1996), and the basis vectors in this space can be used to represent all the faces. The basis vectors or the principal components of the set of faces are known as *Eigenfaces* because they are the eigenvectors (principal components) of the set of faces; they do not necessarily correspond to physical features on the face (Turk & Pentland 1991*a*). How descriptive these Eigenfaces are depends on the initial set of training faces that are used to determine the Eigenspace.

The Eigenface approach has many advantages. It is a relatively simple technique that is easy to implement and has a very good runtime performance (Barrett 1998). Generating the Eigenfaces for the database is computationally intensive (proportional to the size of the images and the database), however, it is undertaken only when the database is updated. Recognition rates of over 90% are reported on a large database (3000 subjects, 8000 images) of full-frontal images ("mug shots") with strictly controlled scale, illumination and pose (Bichsel & Pentland 1994). This technique is insensitive to some forms of noise - small occlusions, as long as the topological structure remains unchanged, blurring and minor changes to the background (Zhang 2003).

Among its drawbacks is that it is extremely sensitive to deviations from the initial training set. It is heavily reliant on pre-processing and normalisation and is known to perform poorly in the presence of illumination changes. The degradation in the performance is also significant when there is a marked difference in the pose and expressions (Sirovich & Kirby 1987, Barrett 1998, Phillips et al. 2000). This is because PCA essentially selects a subspace which retains most of this variation, and consequently the similarity in the face space is not necessarily determined by the identity (Shakhnarovic & Moghaddam 2004). The system is also easily fooled by short term changes such as variations in facial hair or hairstyle (Suthankar 1997). Long term changes such as those due to aging can, in theory, be handled with ease assuming new training images are added to the database regularly. To optimise the classifier performance, the images have to be normalised for head location, lighting, contrast, rotation and scale, and for the geometry of the face in order to obtain the kind of results reported in the literature.

Eigenfaces is usually the benchmark algorithm in 2D face recognition (Phillips & Newton 2002) and has been combined with many other techniques such as neural networks, wavelets, support vector machines, Bayesian methods, etc. to form hybrid methods. An independent comparative study of different Eigenface-based approaches can be found in (del Solar & Navarrete 2005), while a comparison of the different distance measures appears in (Yambor et al. 2002).

In (Pentland et al. 1994), the technique has been extended to Eigenfeatures corresponding to face components such as eyes, nose and mouth. A modular Eigenspace composed of Eigeneyes, Eigennose and Eigenmouth is used. This has the advantage of being relatively insensitive to appearance changes than the standard Eigenfaces method. Approximately 95% accuracy on the FERET database of 7,562 images of approximately 3,000 individuals (Pentland et al. 1994, Tolba et al. 2005) is achieved. Pentland et. al also explores the idea of view-based Eigenspaces, in which different views (front and profile) of the face are grouped together to produce a single Eigenspace, corresponding to the chosen view. This approach also performs better than the original suggested in (Turk & Pentland 1991*a*). In (Penev & Atick 1996), PCA is combined with Local Feature Analysis (LFA), another biologically motivated technique which extracts topographic information from the principal components of the faces. The combination leads to better performance than the use of PCA by itself, however, performance results for LFA alone are not provided. LFA is claimed to be used in the "FaceIt" commercial system of Identix[1] (formerly known as Visionics) (Zhao et al. 2003). Moghaddam et. al (Moghaddam & Pentland 1997) extend the standard technique to a Bayesian approach. However, this approach requires the estimation of probability distributions in a high dimensional space from a limited number of training images per class.

Eigenfaces is used extensively in this project. It is thought to have been motivated by the way humans process faces for recognition (Valentine 1991) and can easily be used with dynamic inputs such as CCTV images. Such images are small and are not suitable for use with most feature based methods. Appearance based methods like Eigenfaces are appropriate for such applications (Zhao et al. 2003) and its use is in line with the long term goals of this work. A detailed mathematical exposition of the technique can be found in Chapter 8.

2.2.2 Linear Discriminant Analysis (LDA)

The LDA algorithm uses the PCA subspace projection as a first step in processing the image data (Zhang 2003). Fisher's Linear Discriminants are defined in the K dimensional sub-space defined by the first K principal components. $c-1$ basis vectors are defined for c classes.

LDA determines a subspace in which the between-class scatter is maximised while keeping the within-class scatter constant (Weng & Swets 1999). The face subspace obtained using LDA optimally discriminates face classes (classes are most linearly separable) in the training set (Weng & Swets 1999, Martínez & Kak 2001, Shakhnarovic & Moghaddam 2004) by using class-specific information very effectively. PCA on the other hand constructs the face space without using any information from the face classes, and as a result, the LDA procedure is known to work in cases where PCA has failed (e.g. in images with varying illumination and facial expressions) (Swets & Weng 1996*b*).

For all images in all of the classes, two measures are defined: the *within-class* scatter

[1] www.identix.com

matrix S_w and a *between-class* scatter matrix S_b given by (Martínez & Kak 2001):

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{x}_{ij} - \mu_j)(\mathbf{x}_{ij} - \mu_j)^\top, \qquad S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^\top \tag{2.1}$$

where $\mathbf{x}_{ij}$ is the i^{th} image of the j^{th} class, c is the number of classes, N_j is the number of samples in class j, μ_j is the mean of class j and μ is the mean of all classes.

The goal is to maximise the between-class measure while minimising the within-class measure. A commonly used approach is to maximise the ratio of determinants of S_w and S_b, i.e. ratio $\frac{\det|S_b|}{\det|S_w|}$ (Swets & Weng 1996*b*, Zhao et al. 2000, Martínez & Kak 2001).

Based on Fisher's Linear Discriminant (FLD) (Fisher 1936), this technique is reported to yield better results than several competing methods (Belhumeur et al. 1997), especially for large databases (Swets & Weng 1996*b*,*a*, Belhumeur et al. 1997, Etemad & Chellappa 1997, Zhao et al. 1998, Zhao 1999, Zhao, Chellappa & Phillips 1999). The main drawback of this technique is that it is computationally intensive (more so than PCA). Calculation of the scatter matrices increases both complexity as well as the processing time, while the technique's ability to better classify images means that the dimension of projection in the face space is not as compact as that of Eigenfaces, resulting in increased storage requirements. The technique is known to fail if both scatter matrices are singular (Li et al. 2004). In terms of input images, the classical LDA does not cope well with images that differ significantly from the training images, and in particular when the backgrounds in the training and the test images are different (Zhao, R.Chellappa & Phillips 1999).

PCA and LDA both tackle the face recognition problem using a dimensionality-reduction approach. LDA is generally reported to give better results than PCA. However, in (Martínez & Kak 2001), an experimental comparison of the two techniques showed that when the number of training images available per class is small (typically 1-2) or when the training data do not uniformly sample the underlying distribution, PCA outperforms LDA. They conclude that in practical applications such as face recognition, since the underlying distribution for the different face classes is not known in advance, it is generally difficult to ascertain whether PCA or LDA is best suited. Zhao et. al strike a compromise between the two techniques and combine them in "Subspace LDA" (Zhao, R.Chellappa & Phillips 1999) and report superior results to both PCA and LDA on the FERET database. A detailed review of the different variants of the LDA approach can be found in (Zhao et al. 2003).

2.2.3 Support Vector Machines

Support Vector Machines (SVM's) were proposed in 1998 by Vapnik (Vapnik 1998) and are used for a variety of pattern recognition problems including face recognition. SVM's perform pattern recognition between two classes by finding a decision surface that has maximum distance to the closest points in the training set which are termed support vectors (Heisele et al. 2001). They have been used for face recognition and verification by many researchers including Phillips (Phillips 1998), Kwong (Kwong & Gong 1999), Guo (Guo et al. 2000) and Heisele (Heisele et al. 2001), and they report better performance than the Eigenfaces technique. SVM's are reported to yield better results on the FERET (see (Zhao et al. 2000) for details on FERET) database when tested against the PCA methods. SVM's have been employed for the 3D object recognition tasks and have reported good performance (Blanz et al. 1996, Pontil & Verri 1998, Roobaert & Hulle 1999).

SVM's are formulated to solve a classical two class pattern recognition problem, but are usually adapted in some way to deal with the multi-class face recognition problem (Phillips 1998). Two common approaches are "one-against-all" and "one-against-one" (Guo et al. 2000, Heisele et al. 2001). A "one-against-all" strategy can be used to classify between each class and all the remaining, however, this is reported to give ambiguous results. In a "one-against-one" strategy, classification is reduced to a series of two-class problems. The results of these problems can be combined to give a final answer using a bottom-up binary tree, for example.

An alternative approach in (Phillips 1998), treats face recognition as a K class problem, where K is the number of known individuals. The problem is then formulated as a problem in the *difference space*. This models the *dissimilarities* rather than similarities between face images. From the training set $T = \{t_1, \ldots t_M\}$ of faces of K individuals, 2 classes are generated. The first is the *within-class differences set* C_1, which are the dissimilarities between the faces of the same person (i.e. the dissimilarities arising due variation in pose, expression, illumination, etc.). The second is *between-class differences set*, which are the dissimilarities between the faces of different people. C_1 and C_2 are the inputs to the SVM algorithm, which outputs a decision surface. Phillips modifies the interpretation of the decision surface of the SVM and proposes a similarity metric. Faces are recognised by setting threshold values for the similarity metric. This idea of a difference space to classify faces is also explored by Moghaddam et al (Moghaddam et al. 1998) but in a Bayesian framework.

The main drawback of SVM's is that for large optimisation problems, they are computationally intensive and the memory requirement grows with square of the training vectors (Yang & Ahuja 2000). A comprehensive survey of the use of SVM's in face recognition can be found in (Tolba et al. 2005).

2.2.4 Elastic Bunch Graph Matching (EBGM)

This method was proposed by Wiskott et al in Wiskott et al. (1997). Faces are stored as grids (graphs) with the characteristic facial features attached to the nodes of the graphs. Nodes are positioned at fiducial points (such as eyes, tip of the nose, etc.) (Lu 2003) and edges are labelled with 2D distance vectors. The features are obtained by convolutions of the faces with Gabor wavelets computed at node locations (Fromherz 1998). The finite wavelet set (encompassing both phase and magnitude information) at a particular grid point forms a feature vector called a *jet*. The image is characterised by these set of jets, which comprise of a relatively small set of numbers by which two images may be compared (Barrett 1998). In order to accommodate different scales, translations, facial expressions and poses, Lades and von Malsburg discovered that the grid could be elastically distorted (within constraints), in order to find the best match between two images (Barrett 1998).

Individual faces in the database are represented by simple labelled graphs. All face graphs are stacked together to obtain a comprehensive representation of the training set. The individual graphs are required to have the same structure so that the nodes refer to the same fiducial points (Lu 2003). All jets referring to the same fiducial point (e.g. left eye) are bundled together in a bunch, from which one can select any jet as an alternative description (Lu 2003). Hence, a face bunch graph (FBG) is a collection of individual face model graphs combined into a stack-like structure. Each node corresponds to a certain facial feature and contains the feature jets of faces from the training set. This allows the system to locate the

fiducial points in a single matching process and eliminates the need for matching each model graph individually (Kong et al. 2005).

Test faces are matched by maximising some similarity measure between the test face graph and the FBG. The similarity measure is usually taken as the average of the best possible match between the new image and any face stored within the FBG minus a topographical term, which accounts for distortion between the image grid and the FBG (Lu 2003). Pose invariance is a consequence of the elastic deformation of the graphs.

The elastic graph method is robust with respect to variations in pose, size, and facial expression and can deal with different lighting conditions (Fromherz 1998, Zhang 2003). This is mostly due to the fact that Gabor features are largely insensitive to lighting, face-position and expressions (Zhang et al. 1997). Small changes in pose about any axis amount to local scale and rotation transformation, which the jets can cope with effectively (Barrett 1998). Changes in expressions are slightly more difficult to deal with, but the grid distortions will to some extent track these changes. Trials using the FERET database have resulted in high recognition rates and the algorithm was ranked among the three most accurate (Kong et al. 2005).

The algorithm is complex and difficult to implement as a large number of grid placements have to be done manually. It requires a large number of convolution images for better performance. Careful pre-processing and accurate feature point location are also a crucial requirement. Further, high resolution images (e.g. 128×128) are required, which largely restricts the applications of this algorithm (Kong et al. 2005).

2.2.5 Neural Networks

One of the first examples of neural networks in face recognition is Kohonen's associative map (Kohonen 1988). Even when the input images were very noisy or had portions missing, an accurate recall capability was achieved on a small set of face images. Neural networks are generally very complex and difficult to train to perform face recognition tasks (Barrett 1998, Zhang et al. 1997), even when the image size is fairly small. For example, if the image size is 128×128, a simple back-propagation (BP) network would require 16,384 inputs (Zhang et al. 1997). Difficulties also arise when the number of classes increases and when only single training image per class is available since multiple images are necessary in order to determine the optimal parameter settings for training the systems (Tolba et al. 2005).

Cottrell and Fleming propose using two BP networks in order to reduce complexity (Cottrell & Fleming 1990). The first network operates in an auto-associative mode, while the second one operates in the classification mode. The auto-associative network has n inputs, n outputs and p hidden units, with $p \ll n$. The network takes a face vector $\mathbf{x}$ as input and is trained to produce an output $\mathbf{y}$ that is a best approximation of $\mathbf{x}$ (Zhang et al. 1997). The output $\mathbf{h}$ of the hidden layer is a compressed version of $\mathbf{x}$, or a feature vector and is used as the input to the classification network. In (Zhang et al. 1997), the authors found that the performance of such a network is comparable to the Eigenfaces algorithm, but the implementation costs are much higher.

Multi-layer perceptron neural networks and radial basis function networks (Howell & Buxton 1996) have also been used for face recognition. A back-propagation training algorithm for multi-layer perceptron may be sufficient for a low dimensionality with a small number of classes (Weng & Swets 1999).

In (Lawrence et al. 1997), a hybrid neural network that combines local image sampling, a self-organising map (SOM) neural network and a convolutional neural network is presented. The proposed method is capable of rapid classification, requires only fast, approximate normalisation and preprocessing, and is reported to consistently exhibit better classification performance than the Eigenfaces approach (96.2% accuracy on the ORL database of 400 images of 40 individuals). Invariance to minor changes in the image samples and rapid classification are achieved through SOM, which quantises the samples into a topological space where the inputs that are nearby in the original space are also nearby in the output space. The convolutional network extracts successively larger features in a hierarchical set of layers and provides partial invariance to translation, rotation, scale and deformation (Tolba et al. 2005). The main drawback of this technique is that training the network takes a long time.

In (Brunelli & Poggio 1992) a neural network approach to gender classification is described. Two Hyper Basis Function (HyperBF) networks (Poggio & Girosi 1990) were trained, one for each gender. This approach has been extended to face recognition using one HyperBF per person. Neural networks have also been combined with statistical methods. For example, Lin et al (Lin et al. 1997) used a probabilistic decision based network (PDBNN) for face detection and recognition. The PDBNN does not have a fully connected network topology. Instead, it divides the network into K subnets, each of which is dedicated to recognise one person in the database. Gaussian activation functions act as neurons and the output of each subnet is the weighted sum of the neuron outputs (Tolba et al. 2005).

2.2.6 Hidden Markov Models

Hidden Markov Models (HMM) are a set of statistical models used to characterise the statistical properties of a signal (Nefian & III 1998). While Hidden Markov Models (HMM) have been used in speech recognition for over three decades, and were also promoted for gesture recognition, relatively little work has been done on applying HMM to face recognition (Samaria 1993, Achermann & Bunke 1996).

HMM's generally work on sequences of coherent 1D signals (feature vectors), while an image usually is represented by a simple 2D matrix. To overcome this, a sliding window is applied to the image. The window covers the entire width of the image, and is moved from the top to the bottom of the image. The brightness values of the windows are passed to the HMM process as 1D feature vectors. Successive windows overlap to avoid cutting off significant facial features and to bring the missing context information into the sequence of feature vectors. The human face can be divided in horizontal regions like forehead, eyes, nose, mouth, etc. that are recognisable even when observed in isolation. Thus, the face is modelled as linear left-right HMM model of five states, namely forehead, eyes, nose, mouth and chin (Fromherz 1998). Accuracy of 87% is achieved on the ORL database of 400 images of 40 individuals.

A pseudo 2D HMM is presented in (Samaria & Harter 1994). Each face image is represented by a 1D vector series of pixel observations. Each observation vector is a block of L lines, with an overlap of M lines between successive observations. A test image is first sampled to a 1D observation sequence and then matched against every HMM in the training set. The match with the highest likelihood is considered the best match and corresponds to the identity of the test image. A recognition rate of 95% was reported in the preliminary experiments on the ORL database of 400 images of 40 individuals (Tolba et al. 2005).

2.2.7 Miscellaneous

Other popular techniques for face recognition not covered here include Active Shape Models (Cootes et al. 1995, Lanitis et al. 1997, Cootes et al. 2000, 2001), Evolutionary Pursuits (Liu & Wechsler 2000), Independent Component Analysis (ICA) (Draper et al. 2003), non-negative matrix factorisation (Lee & Seung 1999). None of these techniques achieve perfect results by themselves. But when combined with other techniques, many of these limitations can be overcome (Tolba et al. 2005).

2.3 Evaluation

The evaluation of face recognition systems has become very important in the recent years as the number of systems available both commercially and otherwise has grown exponentially. As a result, it is necessary to have a common and consistent evaluation protocol in order to assess the quality of these systems. The main goals of *Facial Recognition Technology (FERET)*[2] database and evaluation methodology are (Phillips et al. 2000, Zhang 2003):

- Measuring the performance of face recognition technologies in a framework that models real-world settings and uses a large database,
- advancing face recognition technologies, and
- collecting a database of facial images to support algorithm development and evaluation.

The FERET evaluations provide a comprehensive picture of the state-of-the-art in face recognition from still images (Phillips et al. 2000). Algorithms' identification and verification abilities can be evaluated by testing them on different versions of algorithms, scenarios and categories of images (e.g. lighting change, people wearing glasses or not, time elapsed between training and test image, etc.).

The FERET database has made it possible for researchers to develop algorithms on a common database and to report results in the literature (Phillips et al. 2000). This allows an objective assessment of the algorithms and their relative merits and drawbacks since they are tested using identical normalisation procedures, scoring methods and images. Image collection for the database started in September 1993, and the first test was carried out in August 1994.

FERET tests evaluate *fully automated* and *partially automated* face recognition systems. Fully automated systems localise and normalise the face in a given image and perform automatic face recognition. Partially automated systems only perform the recognition task. The database consists of 1,199 individuals and a total of 14,126 (Phillips et al. 2000) images taken in a variety of settings and time scales (2 days to 2 years). The dataset is partitioned further into various categories based on the images contained in them. For example, there are categories consisting of duplicate images, images taken under the same/different lighting conditions, images taken within 5 minutes/2 days/2 years of each other, etc. The results of the identification tests are presented in (Phillips et al. 2000), while the verification results are published in (Rizvi et al. 1998).

[2] www.itl.nist.gov/ad/humanid/feret

2.3.1 Face Recognition Vendor Tests

Based on FERET, the Face Recognition Vendor Tests (FRVT)[3] provide independent government evaluations of commercially available and prototype face recognition technologies. So far, two tests have been performed - in 2000 (FRVT 2000) and in 2002 (FRVT 2002). FRVT 2002 featured a high computational intensity test (HCint), which consisted of 121,589 full frontal images of 37,437 people. The medium computational intensity test (MCint) consisted of two parts: still and video. The still portion of the test compared different categories of still images (e.g. varying illumination and pose). The video portion tested face recognition accuracies using dynamic images. It should be noted that these tests aim to assess the performance of the state-of-the-art technology and identify areas of future research and are not "buyer's guides" as such. The main findings of the test are summarised below:

- The recognition rate for indoor images is much higher than outdoor ones, even with noticeable changes in lighting. Thus, face recognition from outdoor imagery remains a research challenge area.
- Identification performance decreases linearly as the database size increases logarithmically.
- Identification rates were higher for males than females. They were also higher for older people (38-42 year olds) than for younger people (18-22 year olds). Hence, demographic information should be accounted for when assessing face recognition performance.
- Performance using video sequences was the same as using still images (using FRVT 2002 datasets).
- The use of morphable models significantly improve non-frontal face recognition.

2.3.2 The CSU Face Identification Evaluation System

Colorado State University provides algorithm evaluation facility for researchers developing 2D face recognition algorithms[4]. It provides 4 baseline algorithms:

1. PCA or Eigenfaces algorithm (standard implementation as described in (Turk & Pentland 1991*a*)
2. A combination of PCA and LDA algorithm
3. A Bayesian Intrapersonal/Extrapersonal Image Difference Classifier
4. An EBGM algorithm that uses localised landmark features represented by Gabor jets

The site does not provide any data as most of the work is based on the FERET data. Normalisation code for the data is made available. As with the FERET database, the normalisation involves aligning the eyes using the co-ordinates of the centres of the eyes. The images are scaled so that the distance between the eyes is constant and cropped so that they are all the same size. Canonical images are obtained by applying a mask that zeroes the all pixels that

[3] www.frvt.org

[4] "Evaluation of Face Recognition Algorithms:" www.cs.colostate.edu/evalfacerec

do not fall in an oval that contains the typical face. This removes hair, clothes, ears and anything else in the background. Histogram equalisation is applied to smooth the distribution of the grey values for the non-masked pixels. Finally, the pixel values are normalised so that the non-masked pixels have a mean of zero and a standard deviation of one.

Performance results of the benchmark algorithms using the FERET database are also provided on the website.

2.3.3 Face Recognition Grand Challenge

FERET, FRVT and the CSU Face Identification Evaluation System all address the problem of comparing and evaluating *2D* face recognition systems. However, the recent advances in 3D face recognition technologies have meant that a common data corpus and evaluation framework are necessary to measure the performance of these systems. The Face Recognition Grand Challenge (FRGC) aims to determine the merit of techniques that perform face recognition using 3D scans, high-resolution still images (single and multiple) and multi-modal images (2D and 3D) (Phillips et al. 2005). To this end, it provides a data corpus of 50,000 high-resolution images and 3D scans, a set of evaluation experiments and an infrastructure that supports an objective comparison among different approaches (Phillips et al. 2005).

Images of 200 subjects are collected at the University of Notre Dame, in a series of photo sessions spanning between 2 and 10 months. Four controlled still images, two uncontrolled still images and one 3D image is collected in each session. The controlled images are taken indoors with two different lighting and facial expressions. The uncontrolled images are taken in varying illumination and in non-laboratory settings: hallways, atria or outdoors. The 3D images are taken in controlled setting using a structured light sensor, which captures both range and corresponding texture information. The high-resolution still images measure either 1704×2272 pixels or 1200×1600 pixels. The 3D range images are 640×480 range sampling and a registered colour image of the same size. As with the FERET database, the images are normalised so that the average distance between the centres of the eyes is 68 pixels with a standard deviation 8.7 pixels.

The challenge consists of 6 experiments to gauge the performance of algorithms when dealing with a number of face recognition problems. These include recognition in controlled and uncontrolled environments, under varying illumination conditions, under varying expressions, multi-image and multi-modal recognition and finally, recognising 2D images using 3D or multi-modal training data.

Turk and Pentland's PCA based Eigenfaces algorithm (Turk & Pentland 1991*a*) is used as the baseline algorithm. For the multi-modal recognition, PCA is performed on the range and texture channels separately, and the similarity scores are fused to give a final similarity score.

The FRGC data corpus is the only one of its kind and allows researchers to test their 3D and multi-modal algorithms using a common dataset, similar to that of the FERET or the FRVT. However, for this work, such a dataset was available too late. Besides, this work investigates 3D face recognition from binocular stereo data, which is not available in the FRGC data corpus.

2.4 $2\frac{1}{2}$D and 3D Face Recognition

Early face recognition techniques functioned in the 2D space, and this has been the case until recently. Technology in this field is now mature and a certain ceiling has been reached in terms of performance. Hence the advent of 3D face recognition. 3D face recognition has been around for about a decade and is still in its infancy. However, the constant demand for robust, accurate recognition systems, and the advances in technology is causing a rapid shift in focus from 2D to the 3D space. Much of the research in 3D face recognition used to be done commercially, and as a result, the literature on the subject is sparse in comparison. Intuitively, it can be seen that the 3D approach would have many advantages over the 2D. If a practical and accurate 3D recognition system can be developed, the financial gains are tremendous. Research information on 3D technology has only started to become available in the public domain in the last 5-8 years.

In this work, 3D systems refer to model- and surface-based approaches while $2\frac{1}{2}$D refers to systems using 2D images that encode depth information. This section will review some of the main approaches to face recognition using depth information. The techniques will be divided on the basis of modality, i.e. those using only shape information and multi-modal systems using both shape and texture information.

2.4.1 Shape based systems

These systems use face data captured with specialist equipment such as laser scanners, range cameras or structured light projectors. Although the colour and texture information may be available, its use is generally avoided and identification is on the basis of depth information alone. Stereo-based shape-only systems require accurate data capture and precise camera matrices.

Gordon extracts facial features merely to perform normalisation and define template regions, which are later used for combined recognition of frontal and profile regions in a classical template matching process (Gordon 1995, Fromherz 1998). Features are extracted from one frontal and one profile image using tangency constraints and heuristic knowledge about head structure. These features are used for normalisation, and later to define template regions to be used for face recognition. In (Gordon 1991, 1992), she proposes a technique in which the face is first segmented using surface descriptors based on curvature features and knowledge about the structure of the face. Features such as width and separation of eyes, and height and width of nose are extracted and thus each face is represented as a point in the feature space. Faces are matched using a nearest-neighbours type approach and an accuracy of up to 100% is reported on a dataset consisting of 3 views of 8 faces (24 images) captured using a rotating laser scanner.

Hesher et al (Hesher et al. 2003) explore PCA based approaches using different numbers of eigenvectors and image sizes for range images. As in 2D PCA analysis, the range images undergo rigorous normalisation. In the 3D domain, this takes the form of feature alignment. Images are aligned (rotated and translated) using the tip and the bridge of the nose. This is followed by projection to a lower dimensional space and recognition using PCA. The system is tested using a dataset of 37 subjects, with 6 images per subject (varying expressions). Invariance to changes in pose and illumination are not tested and the expressions are controlled. Recognition rates of over 80% and around 90% are reported when the classifier is trained using single and multiple images respectively.

In (Heseltine et al. 2004*b*), a combination of PCA and LDA is applied to surface representations of 3D models to generate Fishersurfaces. This extends the earlier work (Heseltine et al. 2004*c*) using only PCA to produce Eigensurfaces. To generate Fishersurfaces, PCA is used to reduce the dimensionality of the LDA scatter matrices, prior to computing the eigenvectors of the matrix of ratios of the dimensionality-reduced scatter matrices. These eigenvectors are then used to project a face surface vector into a face space of smaller dimensions, as in PCA. The vector representing each face in this reduced dimensionality space (analogous to the vector of "weights" in Eigenfaces) is known as the "face-key". Face-keys are compared using either Euclidean or cosine distance measures, and acceptance/rejection in verification tasks is determined by applying threshold values. Fishersurfaces perform better than the Eigensurfaces and cosine distance results in lower error rates than the Euclidean distance. The lowest Equal Error Rate (EER) of 11.3% is reported for the horizontal surface gradient representation. Identification experiments using this approach are not conducted. This work is extended in (Heseltine et al. 2004*d*), where different surface representations are combined to improve the accuracy of the classifiers. The authors conclude that although some surface representations do not perform so well when used for recognition, they may contain highly discriminatory components that could complement other surface spaces. The EER is reduced to 9.3% when cosine distance metric is used in a space combining 184 dimensions from 16 different surface spaces. A database of 3D images collected at the University of York is used in these works. A stereo based 3D camera using structured light projection captures the 2D and 3D data. The database consists images of 280 people (1770 images, on average 6 images per person), captured under partially controlled imaging conditions. Lighting is not strictly controlled and no effort is made to enforce precise orientation. Earlier works of Heseltine et al (Heseltine et al. 2002, 2004*a*) focus on pre-processing techniques most suitable for Eigenfaces approach and on combining the outputs of these techniques using Fisher's Linear Discriminant, both for 2D images.

Medioni and Waupotitsch (Medioni & Waupotitsch 2003) and Uchida et al (Uchida et al. 2005) both use a pair of stereo cameras to determine the 3D shape of the face. Stereo based systems are very rare in face recognition due to the low quality and low accuracy of the captured 3D information (Uchida et al. 2005). Both these approaches use calibrated cameras to obtain a metric reconstruction of the facial surface from the disparity values. The surfaces are then matched using Iterative Closest Point (ICP) algorithm. Medioni's system is tested on a dataset of 100 people, with 7 images per subject. Images are captured both indoors and outdoors in a variety of lighting conditions. It is compared against Identix FaceIt System - one of the top three in the FRVT 2000 tests, and is reported to give significantly better results. Uchida et. al's system is tested on a much smaller dataset of 18 subjects. Details about the nature of the dataset and the performance of the system are not presented in the paper.

Yoda et al (Yoda & Sakaue 2003) also developed a stereo based identification system. Their system however, is not a dedicated face recognition system. It processes faces, hand gestures and a number of inanimate objects and is used for human computer interactions. The cameras are mounted on a workstation and the system attempts to identify the objects and individuals most frequently encountered in vicinity. It operates in a strictly controlled cooperative scenario with a small number of subjects to identify. Learning involves extracting regions of interest from depth maps and then using higher order auto-correlation functions and hierarchical LDA to perform recognition.

Chen et al (Chen et al. 2003) use a combination of PCA and wavelet transforms for face recognition using $2\frac{1}{2}$D range images. They use photometric stereo to reconstruct the depth maps from three images, each with a different direction of illumination. A total of 101 subjects are imaged with 6 different viewing directions. Wavelet decomposition using Daubechies-1 (Db-1) wavelet is performed on the depth maps. The dimensionality of the decomposed Approximation images (output of the low-pass channel) is reduced and the images are recognised using PCA. Four different classifiers are tested for use with the PCA algorithm. A moderate decrease in the recognition time is observed for the $2\frac{1}{2}$D images when compared with the 2D images. The classifier combining LDA with the Nearest Feature Line approach performs the best with an accuracy of 95.1% for $2\frac{1}{2}$D data and 89.4% for the 2D data.

2.4.2 Shape and Texture based systems

Systems combining shape and texture may or may not use dedicated equipment to capture depth information. Depth information used in these systems may be captured using a pair of still cameras in a stereo set-up or using a single dynamic data source such as a webcam or CCTV images, with either the camera or the subject in motion.

Model-based approaches construct a generic 3D model of the human face that is able to capture the facial variations in pose, illumination and expressions. The model-based scheme usually contains three steps: (a) Constructing the model; (b) Fitting the model to the given face image and (c) Using the parameters of the fitted model as the feature vector to calculate the similarity between the test face model and the training models in the database to perform the recognition (Lu 2003). Alternatively, the final step can be replaced by 2D recognition system where the classifier is trained using numerous images synthetically generated from the morphable model.

The morphable model approach was first proposed by Blanz et. al (Blanz & Vetter 1999, Blanz et al. 2002, Blanz & Vetter 2003). A 3D morphable face model that parametrises the shape and texture of an individual's face from a single image is used. The model represents shapes and textures of faces as vectors in a high-dimensional face space. Thus, any combination of shape and texture can result in a face (Blanz & Vetter 1999, Blanz et al. 2002, Blanz & Vetter 2003). However, not all of these are realistic. A probability density function of all the faces in the training set is used to limit the generation of unrealistic faces (Blanz & Vetter 1999, Blanz et al. 2002, Blanz & Vetter 2003). Identification is based on these shape and texture parameters, which are independent of the imaging conditions such as illumination and viewpoint (Lu 2003, Tolba et al. 2005).

The training set models are laser scanned and stored in cylindrical co-ordinates relative to a vertical axis (Lu 2003). Texture is represented by the RGB colour values at each of the 3D points. Model parameters are calculated by applying PCA to both shape and texture parameters individually and then concatenating the two. This captures class-specific information about the faces. The 3D model is deformed to obtain the "best fit" between its 2D projection and the new 2D image. Given a single test image of a person, the algorithm automatically estimates the 3D shape, texture, colour, illumination and all relevant 3D scene parameters. Illumination is not restricted to Lambertian reflection, but takes into account specular reflections and cast shadows, which have considerable influence on the appearance of human skin (Tolba et al. 2005).

Weyrauch et. al. (Weyrauch et al. 2004) use the morphable model based approach in the training phase to generate the synthetic face images with varying pose and illumination, and then use a component-based approach to face recognition by first extracting the facial features (using SVM's). They report a better performance than an equivalent global approach and attribute it to the fact that the individual components of the face vary much less than the entire face when the pose and illumination vary.

The main advantage of the 3D morphable model is that it avoids the need to generate an intermediate morphable model for the test image since the 2D image can be matched directly to the 3D model. The system does however, require manual initialisation of pose. Estimation of the parameters from a single image is a computationally intensive task and does take a few minutes to process. In terms of accuracy, the system ranks among the top three algorithms in the recent FERET tests.

Beumier and Acheroy (Beumier & Acheroy 1998, 2000) combine shape and texture information, obtained using structured light projections. Their system is designed with a co-operative scenario in mind and adopts a global surface matching approach to establish geometrical correspondence between two surfaces prior to matching. Surface matching is implemented using parallel profiles 1cm apart. An individual is represented using a central and lateral profile, and identification is based on a matching score obtained by minimising some distance measure based on curvature values of the profiles. Vertical symmetry of the face is used to normalise the surfaces. Although an accuracy of over 90% is reported, the system is known to give poor results or fail completely for individuals with glasses and bushy beards!

In (Bichsel 1995), a modular framework for shape from multiple views and varying illumination is presented. A $2\frac{1}{2}$D depth map and a corresponding texture map are both estimated in an iterative process. This allows, within limitations, the computation of additional views of the head. The model estimation also includes a probability estimation module which computes the probability of a specific set of shape and texture parameters. According to (Fromherz 1998), this can also be used as a recognition module. This, however, is not mentioned by the authors in (Bichsel 1995) as the article deals mostly with estimating head models.

In his PhD thesis (Fromherz 1996), Fromherz adopts a similar approach. He proposes the use of shape from multiple views and visual cues embedded in a framework that allows consistent integration of additional visual cues (Fromherz 1998). A head model is reconstructed and represented as a set of depth maps. Recognition is performed by adjusting the orientation of a particular depth map, together with its associated texture map. This is done iteratively until a match is found. Matches are sought using a simple template matching technique.

Bronstein et al developed a 3D face recognition system that is considered the state-of-art due to its unique ability to distinguish between identical twins. Their system is invariant to facial expressions (Bronstein, Bronstein & Kimmel 2003) but sufficient information about invariance to pose and illumination is not available. They capitalise on the fact that the class of transformations that the facial surface can undergo is not arbitrary. They note that empirical observations demonstrate that facial expressions can be modelled as *isometric* or *length preserving* transformations that do not stretch or tear the surface, or more rigorously, preserve the surface *metric* (Bronstein, Bronstein & Kimmel 2003). Their method first computes geodesic distances between sampled points on the facial surface. Based on

these distances, it produces an isometric invariant representation of facial surface (Bronstein, Bronstein, Kimmel & Spira 2003, Bronstein, Bronstein & Kimmel 2003, Kimmel & Sapiro 2003, Bronstein et al. 2005), which is unique to each individual in the database. The representation is multi-modal in that the recognition is performed using Eigen-decomposition of textures mapped into a lower dimensional Euclidean space and canonical surface representations (Bowyer et al. 2004). Depth and texture information is obtained using a range camera (with coded light - similar to structured light, but more accurate) and the images are pre-processed extensively. The authors report that the registration process is quick but the need for specialist data-capture equipment and lack of information on algorithm's performance under varying pose and illumination limits its applications and versatility. They also report the system's failure when the faces are significantly deformed (e.g. inflated cheeks). Their representation enables distinction between identical twins even without the texture information and the system outperforms both, the 2D Eigenfaces approach and the straightforward incorporation of range images into the Eigenfaces framework.

In (Chang & Bowyer 2005), a dataset of around 200 individuals with multiple images captured over a six to thirteen week period is used to perform various Eigenfaces based experiments for 2D and 3D images. Range scanner is used to obtain a 640 by 480 sampling of range data and a registered colour image of the same size. A standard implementation of Eigenfaces algorithm with the Mahalanobis cosine distance is used. The images are heavily normalised and pre-processed, both automatically and manually. For multi-modal recognition, scores from each modality are combined using a confidence-weighted variation of sum-of-distances rule. The weights are estimated based on the distribution of the top three matches in the 2D and the 3D spaces. A larger distance between the first and second ranked matches implies a greater certainty that the first ranked match is correct. They conclude that recognition using 2D and 3D images leads to similar performance when considered individually. Using multiple 2D images is better than using single 2D images for training. However, combining both 2D and 3D leads to a better performance than either modality by itself and multiple 2D images. This work and their earlier work (Chang et al. 2003) are two of the largest experimental studies reported in the literature either for 3D or for multi-modal face recognition in terms of the number of subjects, the number of training and test images and the time lapse between the training and test image capture.

Tsalakanidou et. al (Tsalakanidou et al. 2003) use the Eigenfaces approach for recognising faces using a combination of depth and colour images. The use of colour images for multi-modal recognition has not been observed before. Experiments are conducted for 2D colour images, 3D range images and a combination of the two. For the 2D images, a Euclidean distance between the test and training image is computed for each of the Y, U and V colour channels. The test image is assigned to the class k for which the smallest product of the Euclidean distances is obtained. A similar approach is adopted for the multi-modal approach. Accuracies as high as 99% are reported on a dataset of 40 individuals from the XM2VTS[5] dataset.

Lu and Jain (Lu et al. 2004*a*,*b*, Lu & Jain 2005*a*,*b*, Lu & K.Jain 2005) present a number of approaches of combining 3D range and 2D texture images to perform automatic face recognition. Feature points are identified either automatically or manually prior to surface registration and alignment between the training and test images. Two methods are presented for automatic feature location. The first uses maximum and minimum local curvature (Lu

[5]http://xm2vtsdb.ee.surrey.ac.uk/

et al. 2004*a*,*b*) to yield a shape index at each point, which helps to identify spherical cups, caps and saddle points on the surface of the face. The second approach automatically locates the tip of the nose (Lu & K.Jain 2005) using cross profile analysis on the segmented range map, based on the shape of the nose. Once the tip of the nose has been identified, other features such as the corners of the mouth and the eyes can be located with ease. Feature points on the facial surface are used to align and register the test and training facial surfaces. Iterative Closest Point (ICP) algorithm is used for matching the facial surfaces, and a 3D matching score is devised using point-to-plane matching. The score of the 3D matching is integrated with the 2D matching score using the sum rule. The 2D matching score is obtained in a number of ways. Hierarchical LDA is applied to a set of potential matches from the training set, short-listed on the basis of the ICP score, in (Lu & Jain 2005*b*). A combination of thin-plate splines and SVM's are used in (Lu & Jain 2005*a*), while (Lu et al. 2004*a*) uses a combination of registration error from ICP, shape index from feature identification and texture matching. In (Lu et al. 2004*b*), the ICP matching score is combined with the cross-correlation between the shape index vectors of the sets of control points in the test and training images. On a database of 100 subjects, the highest accuracy is obtained by the system presented in (Lu & Jain 2005*a*), combining ICP, thin-plate splines and SVM's.

2.5 Summary

After over 30 years of research and development, basic 2D face recognition has reached a level of maturity and many commercial systems are available for various applications (Zhao et al. 2003). Evaluation methodologies are available and both academic and commercial systems are tested using common datasets of FERET and FRVT. Image-based, holistic methods such as Eigenfaces and Fisherfaces dominated 2D face recognition. However, these approaches are susceptible to changes in pose and illumination and can lead to inaccurate identification if the test images vary significantly from the training images. The advent of the 3D techniques that utilise depth information aims to tackle the geometric sensitivity of these systems.

3D face recognition systems that use shape information only rely on accurate depth information. For these systems, the shape of the face is usually measured using dedicated hardware. A handful of systems use stereo images to extract depth information. Shape from stereo is generally avoided for face recognition due to low quality and accuracy of the 3D data. Existing stereo-based systems require careful camera calibration to accurately reconstruct the shape information. Surface matching techniques such as ICP are used for matching, making these systems computationally intensive and not necessarily practical for large databases or for deployment outside laboratory conditions. Further, to the author's knowledge, none of the stereo-based face recognition systems utilise texture information.

Multi-modal systems combining 3D shape information with the 2D texture information have been found to be more successful than either modality individually. Consequently, much of the upcoming research concentrates on this area. Most 3D and multi-modal systems report accuracies in excess of 90%. However, the lack of 3D and multi-modal test sets (such as the 2D FERET dataset) until recently has made it difficult to compare and evaluate these techniques effectively. The Face Recognition Grand Challenge (FRGC), which has been running since 2004, addresses this issue in part by providing a large multi-modal dataset for evaluating 3D and multi-modal systems. However, such a dataset is not available for comparing stereo-based systems.

CHAPTER 3

Three Dimensional Reconstruction

3.1 Introduction

The problem of static 3D face recognition can be divided into three major tasks:

1. Extraction of the 3D structure of the face.
2. Effective representation of the 3D image so that it aids the recognition process.
3. The actual recognition of the 3D image.

The 3D structure of a scene can be extracted in many different ways, depending on the mode of data acquisition (see section 3.2). Stereoscopy is the most generic method for surface reconstruction from 2D images. A general stereoscopic system takes at least two different views of the scene to be reconstructed. These views are obtained either by a single moving imaging sensor or by several sensors at different locations (Rottensteiner 2001). 3D Reconstruction from a pair of stereo images is a non-trivial task and much of the research for this work has concentrated on finding appropriate algorithms to accomplish this. Mathematically, it draws on simple results from classical geometry. However, extracting the appropriate information from stereo images is challenging. This is mainly due to the presence of noise and ambiguity in the data. Despite much of the theory being established for decades, a computational solution applicable to all classes of images remains elusive.

This chapter starts with a brief review of depth extraction methods and the relative merits and drawbacks of stereoscopy. This is followed by the mathematics of 3D reconstruction from stereo. The strength and weaknesses of the calibrated and uncalibrated camera set-ups is also discussed.

3.2 Brief Overview of Depth Extraction Methods

Depth of a scene can be computed in many ways depending on the mode of data acquisition. The modes vary according to the restrictions imposed by availability of hardware, finances

and most importantly, on the purpose of the computer vision system being developed. Some methods of depth extraction include:

Disparity from Stereo: Also known as *Passive Stereo*, this method requires at least two images of the scene, taken from slightly different locations. Also required are the relative positions of the cameras to each other and to the scene. *Disparity*, or the shift in the position of objects in the two (or more) images of the scene is used to compute depth. Special equipment, operation personnel and storage are not required. Off-the-shelf digital cameras are adequate.

Disparity from Motion: Also known as *Active Stereo*, this method uses the geometry of motion to compute depth. This geometry is essentially the same as the geometry used for stereo. However, establishing the relative positions of the cameras is far more difficult now. This method is heavily reliant on pre-processing to detect and track objects in different frames. Again, special equipment, operation personnel and storage are not required. Although off-the-shelf camcorders or webcams are adequate, high-resolution images are preferable for accurate reconstruction.

Texture Gradient: This is best understood by thinking of grass in a field. In close up, all the blades of the grass can be seen quite distinctly, but as the distance from the actual blades increases, it all merges into a green texture. This has limited use by itself and is usually used in conjunction with other techniques such as disparity from stereo and disparity from motion. It also requires detailed knowledge of the scene.

Occlusion: This method uses prior knowledge of the positions of various objects in the scene to establish depth. For example, if there is a tree in front of a house, then a part of the house is occluded by the tree and hence the house is at a greater depth than the tree. Again, detailed knowledge of the scene is required and the technique is rarely used by itself.

Structured Light: This technique is one of the most well-known and widely used methods of depth extraction. It is highly accurate and cost-effective, both in monetary and computational terms. A light stripe pattern (similar to a bar code) is constructed and projected on to a 3D scene. The projections provide light stripe features over the object surfaces. These can be captured to calculate the 3D position of a surface point where the stripe edge is detected (Gühring et al. 2000). This method of depth extraction requires subject co-operation and some specialist equipment. Depth values are usually more accurate than all of the above methods.

Laser Scans: Laser scanning works by projecting a coloured stripe onto a laser illuminated object. Depth is extracted by analysing the changing shape of the stripe as the scanner moves across the object. This method is perhaps the most accurate and the most expensive (Levoy 2000). In addition, the system requires careful storage and maintenance as well as trained personnel to operate it. Subject participation and co-operation are also necessary.

Often some of these techniques are combined to obtain more robust measurements (Fromherz 1996, Choudhury et al. 1999). However, computational as well as monetary costs are high and both sophisticated hardware and software are required. Other types of images used for

depth extraction include Magnetic Resonance Image (MRI) scans for medical images and infra-red images (Kong et al. 2005).

This work employs only one method of depth extraction - *Shape-from-Stereo.* The use of stereo vision defines the scope of the project sufficiently without limiting avenues for further research. A system based on stereo can be mostly non-intrusive and requires little input or participation from the subjects or the operators. In addition, it processes the data in a way that allows the results to be easily verified by human experts, thus limiting the number of false alarms.

Depth extraction methods that require prior knowledge of the scene were immediately rejected. Although occlusion and texture provide valuable depth information, these techniques need to be used in conjunction with other depth extraction methods. They can be introduced at a later stage to enhance the information already available from a stereo based system.

Passive stereo was chosen over active stereo in order to minimise the degree of pre-processing required. It also serves to isolate the effects of using depth information. Active stereo requires tracking of the object to be reconstructed in each of the frames. The tracked object then has to be aligned and matched so that the depth may be extracted. All these procedures add to the computational overheads and may require increased human input. Tracking, in such systems is often performed by marking feature points on the object (Park et al. 2002). This is obviously impractical if a large subject database is to be built, and requires too much human interference. Additional processing renders shape-from-motion techniques more prone to errors than shape-from-stereo techniques. A static system that is efficient and accurate can, in time, be developed further to process dynamic inputs too.

In the long term, it is envisaged that the system will encompass not only images from the static domain but also the dynamic domain (using CCTV images, for example). A system that can recognise faces efficiently and accurately, taking as input dynamic images, has many applications in areas ranging from crowd surveillance and access control to computer games and human-computer-interaction. Hence it is important to be able to extract and process the useful information from simple (and noisy) sources such as images. Depth extraction using structured light and laser scans were also turned down with this long term goal in mind. These techniques are far more accurate than stereo vision, but lack versatility and are confined to use in laboratory conditions.

3.3 Stereo Vision: Advantages and Disadvantages

It may be argued that static stereo can be as intrusive as laser scans or light-stripe projections, since extensive subject participation is required to capture the images with varying pose and expressions. However, it is essential to impose some degree of control on the data acquisition process to allow effective comparison between the 2D and the 3D recognition systems. Further, this degree of intrusiveness can be justified by the project's long term goals of handling dynamic image inputs. A problem as complex as face recognition needs a great deal of simplification in order to curtail the number of variables. These can be introduced gradually to develop a sophisticated system. Stereo imaging using simple off-the-shelf digital cameras does not feature much in the literature on 3D face recognition. This was a further motivation to investigate it in this work.

Reconstruction from 2D images using stereo is a very active and challenging area of re-

search. Two or more cameras are used to capture the scene. The cameras are calibrated so that their internal parameters and positions, both relative to each other and to the scene (external parameters) are known. There are many established practices for calibrating cameras (Tsai 1986, Wilson 1994, Heikkila & Silven 1997, Bouguet 1999, Zhang 2000) and this in itself is an interesting area of research.

Extracting depth information from multiple stereo images is very generic and can be applied in many different situations, both in static and dynamic face recognition applications (e.g. "mug-shot" identification in the static domain and automated crowd surveillance in the dynamic domain). In addition, having two or more separate cameras allows greater control in terms of data acquisition: placement of cameras, resolution of images and a uniform dataset for the training and testing of images in many dimensions. Cost effectiveness is an added advantage of this approach.

When using stereo images to compute depth of a scene, the accuracy of depth information is directly proportional to the number of images used to compute it. However, the number of images used is also directly proportional to the complexity of the calculations, the execution time and the processing power required. This, in addition to financial constraints led to the decision of starting with the simplest two camera setup. Although it is possible to obtain any number of images with a single camera, the camera has to be moved to a different position for each scene capture. This requires re-calibrating the cameras each time they are moved. This can have the undesirable effect of reducing accuracy as errors are introduced each time the cameras are calibrated. It is also a very time consuming process. Slightest errors in the calibration parameters can amplify significantly during the depth extraction process. Hence, two cameras in a stereo configuration were used.

Although the use of more cameras can yield greater accuracy, it can make it extremely difficult to evaluate the matching and the reconstruction algorithms. It may not be possible to establish whether the errors in reconstruction are due to the calibration process or due to the actual algorithm. If the cameras are calibrated once and then not moved throughout the data collection, then the errors can be assumed to be constant across the subjects. Each reconstruction is affected equally and the different algorithms can be compared effectively. In theory, it is possible to extend the results obtained from two cameras to any number of cameras. How well this works in practice; the various tradeoffs between accuracy, complexity, time and processing power requirements; optimisation of the various parameters (including the number of cameras); reconstruction and recognition in the presence of significant noise levels (caused by moving the actual cameras, for example) are all areas of potential future research.

Processing the data obtained from digital images is particularly challenging and computationally more intensive compared with data obtained from laser scans and structured light projections. Although the *process* of data acquisition is simple and easy, the *data* itself is not always so accurate. It is very susceptible to subtle changes in the environment. For example, even though the cameras may be positioned as close as is physically possible (parallel camera model), the scene they capture may still contain slightly varying ambient illumination. These changes are not necessarily observable by eye, but are recorded by the digital cameras in the form of marginally different pixel values. Another source of error, as mentioned above, is the calibration process. Calibration is done manually, so there is an element of human error. Cameras themselves are also a source of errors. Although two cameras of the same make and model have the same technical specification in theory, this is rarely the case in

practice. Inaccuracies in the internal parameters of the cameras may be introduced during the manufacturing process. The variations are usually minor, and again not discernible by the human eye. But they can result in significant errors during the reconstruction.

Face recognition is a challenging biometric as it is, due to the changeable nature of the face. The task is made further difficult by choosing to execute it in a partially-controlled environment using a noisy and a potentially unstable method of depth computation. However, if successful, the applications of such a system would be limitless.

3.4 3D from Images

Figure 3.1: An image of a scene (Pollefeys 2000).

A 2D image such as Figure 3.1 conveys a lot of information about the 3D scene that it represents (Pollefeys 2000).

However, there is not enough information in the image to reconstruct the scene. This is due to the nature of image formation process which consists of a projection from a 3D scene onto a 2D image. During this process the depth information is lost, as illustrated by Figure 3.2.

The point in the 3D space corresponding to a specific image point is constrained to be in the associated line of sight. From a single image it is not possible to determine which point on this line corresponds to the image point. If two (or more) images are available, then the 3D point can be retrieved as the intersection of the two lines of sight. This process is called *triangulation* (Figure 3.3).

A number of variables are needed in order to reconstruct a 3D scene from two or more 2D images:

1. The corresponding image points
2. Relative pose of the camera for the different views
3. Relation between the image points and the corresponding line of sight

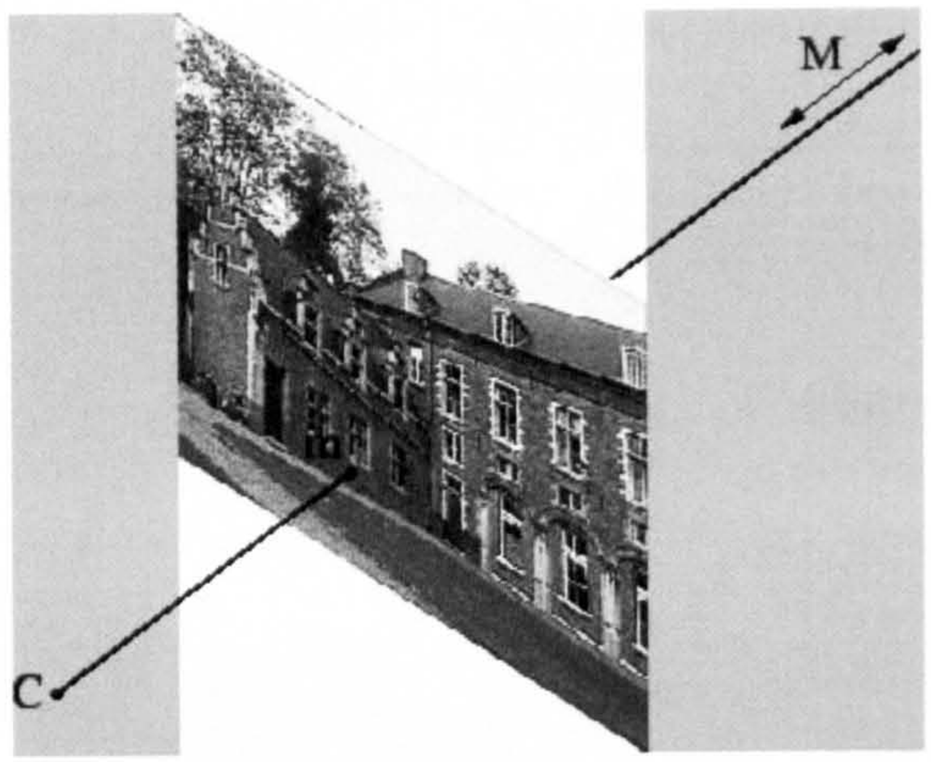


Figure 3.2: Back-projection of a point **m** viewed through a camera **C**, along the line of sight. The 3D point **M** corresponding to the 2D point **m** lies somewhere along the line of sight. However, its exact location cannot be determined from one image alone. (Pollefeys 2000).

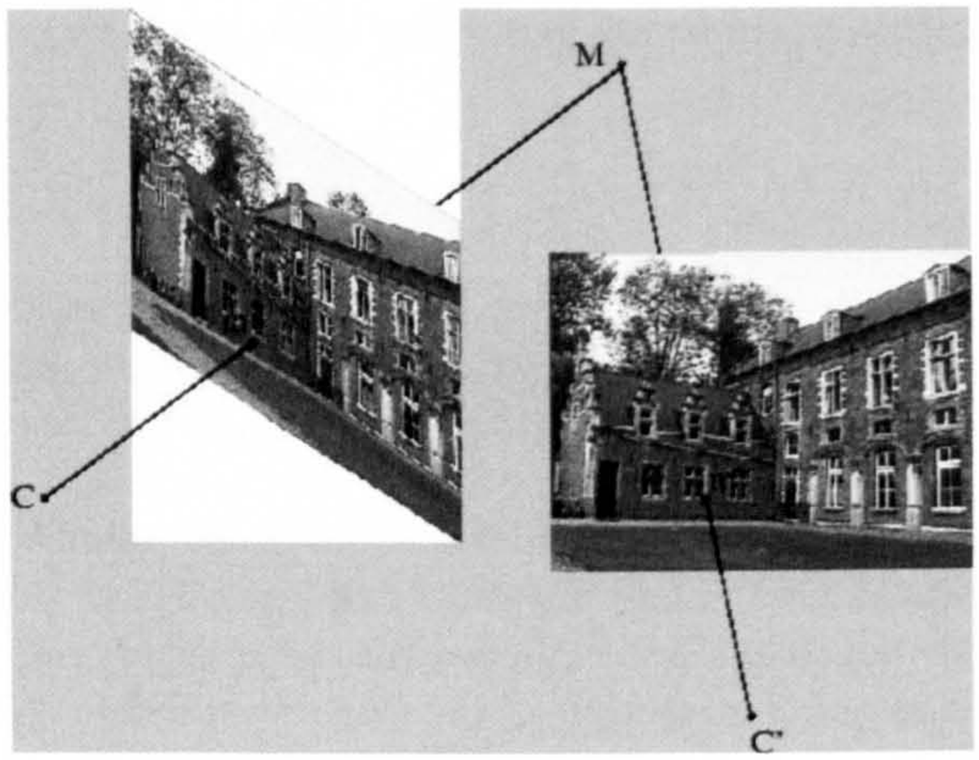


Figure 3.3: Reconstruction of a 3D point **M** through triangulation. When the same scene is imaged using two different cameras **C** and **C′**, it is possible to determine the exact location of a point in the 3D space, *given* that the point is visible in both the images *and* its exact location in the two 2D images is known. (Pollefeys 2000).

The relation between an image point and its line of sight is given by the camera model (e.g. a pinhole camera) and the calibration parameters. Parameters such as the focal length of the lens and the distance between the lens and the film or the image plane are called the *intrinsic* camera parameters, while the position and orientation of the camera are in general called *extrinsic* parameters.

Depth can be estimated more accurately by utilising additional cues such as texture, occlusion and shading (see Section 3.2). However, often despite having all the above information, scenes are still difficult to reconstruct. This is either due to their complex nature or due to the way in which they have been imaged. For example, there may be many discontinuities or reflections within the scene, or the camera model might not satisfy the assumption of a pinhole camera. The images used in this work were taken explicitly for the purpose of reconstruction and recognition. Therefore, every effort has been made to ensure that such situations do not occur. However, sometimes this is unavoidable (e.g. some reflection is encountered in images of subjects with glasses). These images have not been excluded from the dataset as it was imperative to keep the dataset as realistic as possible.

3.5 Stereo Vision - An Overview

Stereo Vision refers to the process in which a scene is projected on both the right and the left eye, i.e. on two image frames (Fromherz 1996). Reconstruction of 3D scenes is performed using multiple images, all taken from slightly different angles. A minimum of two images is required. The theory can however, be easily extended to more than two views.

The basic principle behind depth from stereo is to use the *disparity* between the different views of the same scene to extract depth. Disparity is the shift or the difference between the same scene in the two images. This principle is best understood by holding up a finger in front of one's eyes and alternately closing each eye. In each view, the finger appears to shift by a certain amount - this shift is called disparity.

The process of reconstruction from stereo can be divided into three stages.

1. Feature Extraction
2. Feature Matching or Image Matching
3. Triangulation (depth extraction)

Quite often, the first two stages are combined, depending on the method of establishing correspondences between the two images. Feature extraction and triangulation rely on existing techniques in computer vision and geometry respectively. Feature matching or image matching is the most difficult and time consuming stage of stereo vision, and hence much of the research in stereo vision focuses on this field (Fromherz 1996).

First, a simple approach to the mathematics of the stereo imaging is presented in order to understand the tasks involved in reconstructing a scene from stereo. For detailed mathematics, the reader is referred to (Hartley & Zisserman 2000, Pollefeys 2000). Consider the optical setup in Figure 3.4 (Marshall 1994).

The figure shows two cameras with their optical axes parallel and separated by a distance d. The perpendicular distance between the lens centre and the image plane is the focal length, f. Note that normally both d and f are fixed. The line connecting the camera lens centres is called the *baseline*.

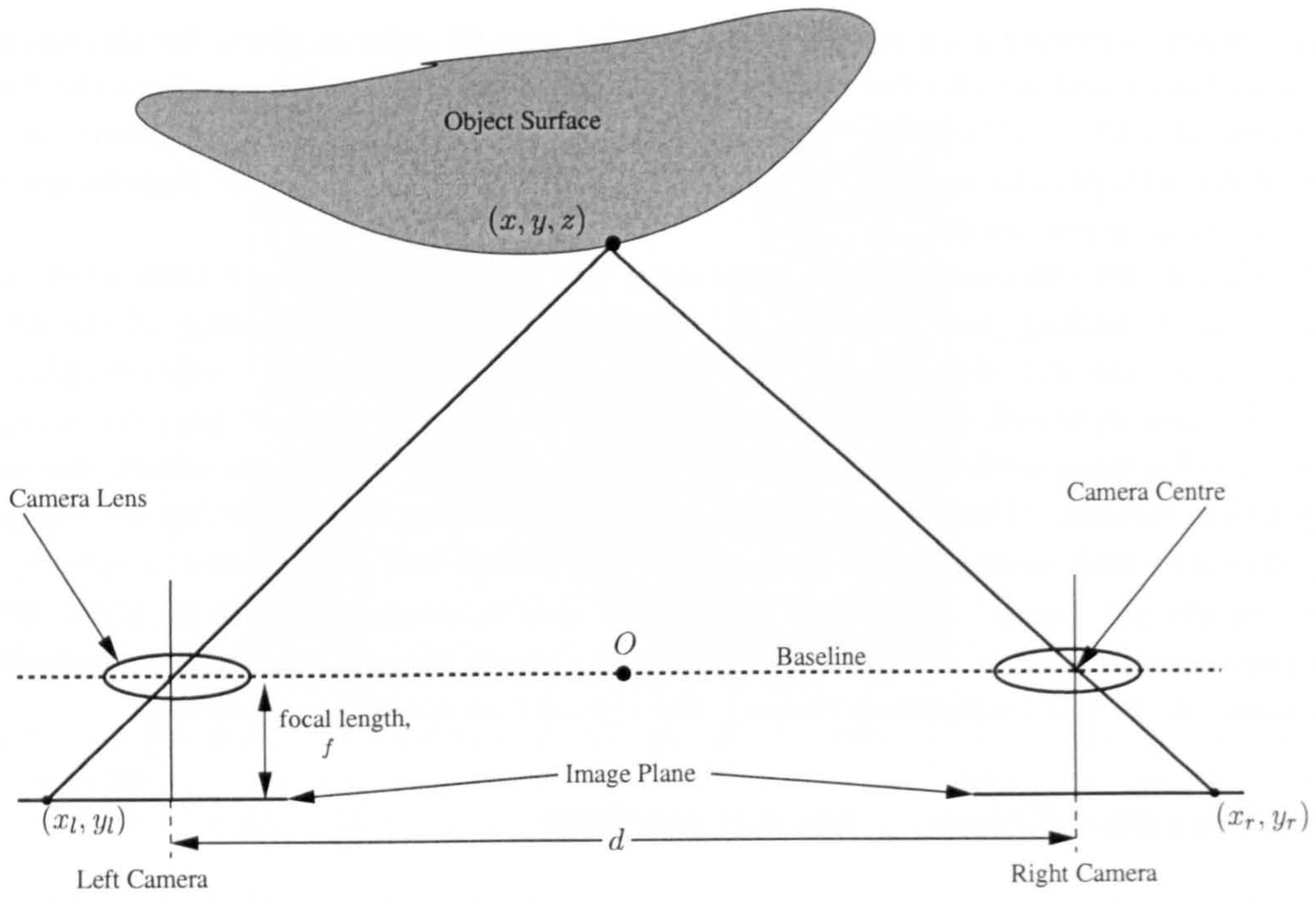


Figure 3.4: Simple triangulation with calibrated cameras and noise free matching points in the two images

Let the baseline be perpendicular to the optical axes of the cameras. Let the x axis of the three-dimensional world co-ordinate system be parallel to the baseline. Let the origin O of this system be mid-way between the lens centres. Consider the point (x, y, z) on an object. Note that this point is in, 3D *world* co-ordinates. Let this point have *image* co-ordinates (x_l, y_l) and (x_r, y_r) in the left and right image planes of the respective cameras. Then by similar triangles:

$$\frac{x_l}{f} = \frac{x + \frac{d}{2}}{z}, \qquad \frac{x_r}{f} = \frac{x - \frac{d}{2}}{z}, \qquad \frac{y_l}{f} = \frac{y_r}{f} = \frac{y}{z} \tag{3.1}$$

Solving for (x, y, z) gives:

$$x = \frac{d(x_l + x_r)}{2(x_l - x_r)}, \qquad y = \frac{d(y_l + y_r)}{2(x_l - x_r)}, \qquad z = \frac{df}{x_l - x_r} \tag{3.2}$$

The quantity $(x_l - x_r)$, which appears in each of the above equations is called the disparity. Disparity is measured in pixel differences and is proportional to the distance between the two cameras. There are several practical problems with this setup. Firstly, the reconstruction accuracy diminishes as the object gets further away from the cameras. As the distance from the cameras increases, the scale and the quality of the object's image deteriorates, subject to the camera resolution. Secondly, as the camera separation d increases, difficulties arise in correlating the two camera images. As the disparity increases, the portion of the scene that is visible in both the images decreases, thus decreasing the proportion of the scene that can be reconstructed. In such situations, employing more than two cameras can be useful. And finally, since disparity is proportional to the camera separation, if there is a fixed error

in determining the disparity then the accuracy of depth values will decrease as d increases (depth values are also proportional to the disparity).

3.6 Two-View Geometry

This section briefly covers the geometry of two perspective views. These views may be acquired simultaneously as in a stereo rig, or acquired sequentially, for example by a camera moving relative to the scene. These two situations are geometrically equivalent and are not differentiated.

Each of the two views has an associated camera matrix, $\mathbf{C}$ and $\mathbf{C}'$. A point $\mathbf{X}$ in 3D space is imaged as $\mathbf{x} = \mathbf{C}\mathbf{X}$ in the first image and as $\mathbf{x}' = \mathbf{C}'\mathbf{X}$ in the second image. Image points $\mathbf{x}$ and $\mathbf{x}'$ are said to be *corresponding points* or *matching points* because they are the image of the same 3D point. Establishing correspondence points between two images of the same scene is an extremely difficult task and forms a part of the *Correspondence Problem* (Section 3.6.3).

In order to reconstruct the scene, three pieces of information are required:

1. Epipolar Geometry
2. Camera Geometry
3. Scene Geometry

3.6.1 Epipolar Geometry

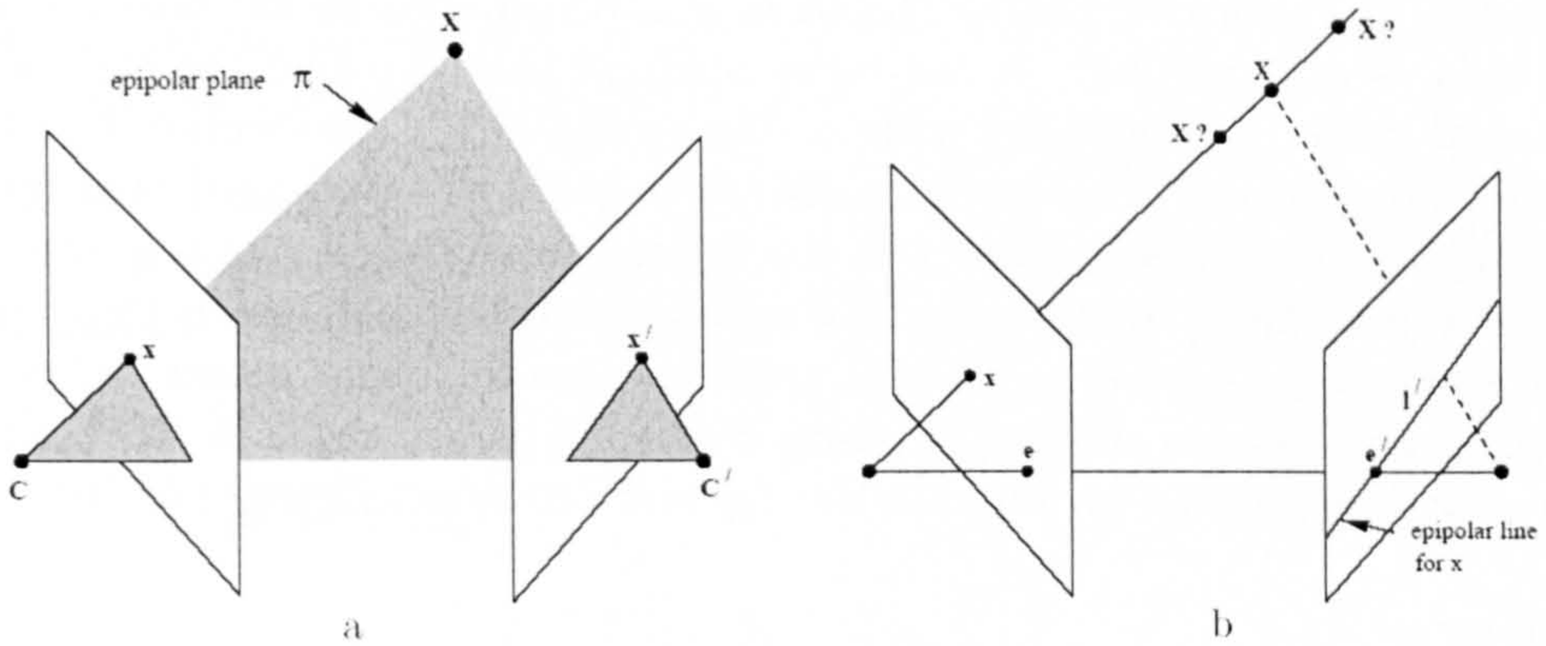


Figure 3.5: **Epipolar Geometry:** (a) $\mathbf{C}$ and $\mathbf{C}'$ are the two camera centres in the left and right images respectively. The camera centres, the 3D point $\mathbf{X}$, and its images $\mathbf{x}$ and $\mathbf{x}'$ lie in a common plane π. (b) An image point $\mathbf{x}$ back projects to a point in 3D space defined by the left camera centre, $\mathbf{C}$, and $\mathbf{x}$. The ray is imaged as a line $\mathbf{l}'$ in the right view. The 3D point $\mathbf{X}$ which projects to $\mathbf{x}$ must lie on this ray, so the image of $\mathbf{X}$ in the right view must lie on $\mathbf{l}'$.(Hartley & Zisserman 2000).

The *epipolar geometry* is the intrinsic projective geometry between two views. It is independent of scene structure, and only depends on the cameras' internal parameters and

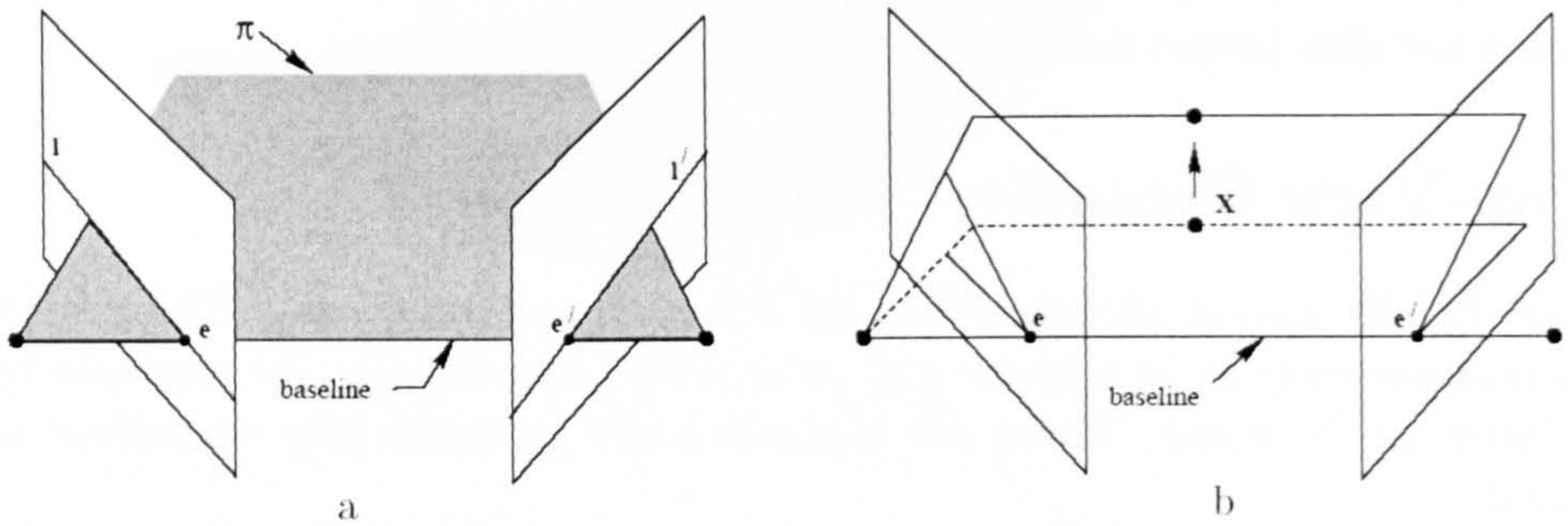


Figure 3.6: **Epipolar Geometry:** (a) The camera baseline intersects each image plane at the two epipoles **e** and **e**′. Any plane π containing the baseline is an epipolar plane, and intersects the image planes in corresponding epipolar lines **l** and **l**′. (b) As the position of the 3D point **X** varies, the epipolar planes "rotate" about the baseline. This family of planes is known as an epipolar pencil. All epipolar lines intersect at the epipole. (Hartley & Zisserman 2000).

relative pose. It is essentially the geometry of the intersection of the image planes with the pencil of planes having the baseline as axis (see Figure 3.6). The geometry is usually motivated by considering the search for corresponding points in stereo images.

In Figure 3.5, **X** is a 3D point that is imaged in two views at image points **x** and **x**′. The image points, the 3D point and the camera centres are coplanar. The rays back-projected from **x** and **x**′ intersect at **X**. These rays are coplanar and lie in the same plane π as **x**, **x**′ and **X** (Hartley & Zisserman 2000). The plane π is determined by the baseline and the ray projected by **x** (see Figure 3.5). It can be seen from Figure 3.6 that the ray corresponding to the point **x**′ lies in π. Hence the point **x**′ lies on the line of intersection **l**′ of the plane π and the right image plane. This line **l**′ is the image of the ray back projected from **x** in the right image plane. Therefore, a *point* **x** in one image generates a *line* in the second image on which its corresponding point must lie. The search for correspondences is thus reduced from a region to a line. This *epipolar constraint* arises because the image points, 3D point and the optical centres are coplanar, for image points corresponding to the same 3D point. Detailed exposition of the subject can be found in Hartley and Zisserman (Hartley & Zisserman 2000).

Terminology

The following terms and their descriptions are taken from (Zisserman 1997).

Epipole The point of intersection of the line joining the optical centres (the baseline) with the image plane. The epipole is the image in one camera of the optical centre of the other camera (denoted by **e** and **e**′ in Figures 3.5 and 3.6).

Epipolar Plane The plane defined by a 3D point and the optical centres. Or equivalently, by an image point and the optical centres (denoted by π in Figures 3.5 and 3.6).

Epipolar Line The straight line of intersection of the epipolar plane with the image plane. It is the image in one camera of a ray through the optical centre and image point in the

other camera. All epipolar lines (denoted by $\mathbf{l}$ and $\mathbf{l}'$ in Figures 3.5 and 3.6) intersect at the epipole.

Many image matching techniques take advantage of the epipolar constraint to establish more robust correspondences between the two images (Zhang et al. 1995, Xu 1997). Rectification is a process by which epipolar lines in the two images become collinear and are aligned to a scan line, i.e. epipolar lines become the rows of the images on the rectified image plane (Wu 2004). The rectified images are produced by a re-projection operation (a homography), such that the rectified plane is parallel to the baseline. As a result, the intersection of the epipolar plane with the rectified image plane is a line, thus reducing the disparity to horizontal displacement on both the rectified images (Wu 2004). Rectification is not adopted in this work.

3.6.2 Camera Geometry

A camera projects a 3D scene space onto a 2D image plane. It is a mapping between the 3D world (object space) and a 2D image. The camera mapping is represented by a 3×4 matrix which maps from homogeneous co-ordinates of a point in 3D space to homogeneous co-ordinates of the imaged point on the image plane (Hartley & Zisserman 2000).

$$\mathbf{C} = \mathbf{KD} = \begin{pmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \left(\mathbf{R}_{3\times3} \middle| \mathbf{T}_3 \right) \tag{3.3}$$

where $\mathbf{R}$ and $\mathbf{T}$ represent the orientation and position of the camera, and $\mathbf{K}$ its intrinsic parameters (Sturm & Quan 1995):

- u_0 and v_0 are the co-ordinates of the principal point, the point where the optical axis of the lens meets the image plane.
- α_u and α_v are the horizontal and vertical scale factors, or equivalently, the focal length in pixels. If the pixels are square then $\alpha_u = \alpha_v$, otherwise α_u and α_v correspond to the width and the height of a pixel as units of length.
- s is the skew factor.

So, for a pair of stereo images,

$$\mathbf{x} = \mathbf{CX} \quad \text{and} \quad \mathbf{x}' = \mathbf{C}'\mathbf{X} \tag{3.4}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are the co-ordinates in the two images of a 3D space point $\mathbf{X}$, and $\mathbf{C}$ and $\mathbf{C}'$ are both 3×4 camera matrices.

The matrix has 11 degrees of freedom and encapsulates the *intrinsic* and the *extrinsic* parameters of the camera. The intrinsic parameters consist of the scale factor in the x and y directions, the skew angle (to take into account non-rectangular pixels), the focal length measured in width and height of the pixels and the co-ordinates of the principal point (u_0, v_0). The extrinsic parameters refer to the location and the orientation of the cameras relative to each other and the subject.

In a stereo setup, the cameras can either be *calibrated* or *uncalibrated.* In the calibrated setup, both intrinsic and extrinsic camera parameters are known or set *a priori.* In the

uncalibrated setup, no information regarding the scene or the cameras is known directly. The camera matrices have to be estimated from whatever information is available. In most cases, the only available information is set of potential correspondences between the two images.

Calibrating cameras is a simple but tedious task. It generally involves viewing some kind of calibration object of known geometry and identifying certain landmark points. The calibration object usually consists of chess-board like grid, or a white board with equidistant black points, or similar. Based on this information, the intrinsic and the extrinsic camera parameters are computed (see Section 4.4).

The reader is referred to (Hartley & Zisserman 2000, Pollefeys 2000, Wu 2004) for further details on Camera Geometry.

3.6.3 The Correspondence Problem

Although the principle of passive stereo vision is extremely simple, finding a way of matching points in the two images accurately presents a major hurdle. The *Correspondence Problem* refers to the search in two or more 2D images for pairs of points that are projections of the same point in the scene. So, in order to reconstruct a scene point appearing in say the left image, the corresponding scene point has to be found in the right image too. That is, given the image co-ordinates of a point in the left image, the image co-ordinates of the exact *same* point the right image need to be established in order to calculate the 3D co-ordinates of the point. The centre of the 3D world co-ordinate system is usually taken to be the optical centre of one of the cameras (usually left), i.e. the position of the point in the 3D space is calculated *relative* to some global frame of reference (Owens 1997).

To solve the correspondence problem without any human input is extremely difficult and requires mathematically robust and sophisticated algorithms. The computer has no concept of features - the only information about the image that the machine has is the image matrix. Each element of the image matrix represents a pixel intensity value. For every point with a particular intensity value, there could be any number of points in the image with the same intensity value. In addition, there is no guarantee that the image point being searched for is actually *present* in the second image. Consequently, the correspondence problem is extremely difficult, and developing a mathematically sound and robust solution for it is a challenging and as yet an unsolved task.

The correspondence problem has been a significant part of this work and is described in detail in Chapter 5, along with a review of the existing techniques for solving it.

3.6.4 Triangulation

Triangulation addresses the problem of finding the position of a point in 3D space, given its position in two images taken using cameras with known calibration and pose (Hartley & Sturm 1997). Hartley and Sturm present a good review of the existing techniques for triangulation and the merits and drawbacks of each in (Hartley & Sturm 1997).

In the absence of noise, this is a relatively trivial problem. However, noise in the camera parameters and in the list of correspondences means that robust algorithms need to be devised. It is generally assumed that there are errors only in the measured image co-ordinates (i.e. correspondences), not in the camera matrices $\mathbf{C}$ and $\mathbf{C}'$ (Hartley & Zisserman 2000).

The presence of noise means that simply back-projecting rays using the correspondences will fail as the rays, in general, will not intersect.

The problem of triangulation has been approached in many ways (Lengagne et al. 2000, Carlsson & Weinshall 1998, Hartley & Sturm 1997, Hartley 1994, Shashua 1994). The choice of technique will depend on whether the set of cameras being used is calibrated or uncalibrated. In general, more algorithms exist for the latter case. These take into account factors such as projective invariance, robust estimation of additional parameters such as camera matrices and noise filtering to allow accurate scene reconstruction. Details of the triangulation algorithm used in this project are as under:

Given a set of correspondences and the camera matrices (computed either by manual calibration or by computing the epipolar geometry directly from the matches), the structure $\mathbf{X}$ of the imaged scene may be recovered by triangulation relatively easily (Hartley & Zisserman 2000). However, obtaining an optimal solution can be costly, both in terms of processing time and complexity. This is because the optimal solution would minimise the re-projection error of the 3D points, i.e. minimise the sum of squares of Euclidean distance between the observed point in each image and the re-projection using the camera matrices and putative 3D structure (Torr 2002), i.e.

$$\min_{\mathbf{X}} \quad e_u(\mathbf{x}, \mathbf{C_1X})^2 + e_u(\mathbf{x}', \mathbf{C_2X})^2 \tag{3.5}$$

where $e_u(\mathbf{a}, \mathbf{b})$ is the Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$. This is equivalent to finding $(\hat{x}, \hat{y}, \hat{x}', \hat{y}')$ such that

$$\sum e = (x - \hat{x})^2 + (y - \hat{y})^2 + (\hat{x}' - x')^2 + (\hat{y}' - y')^2 \tag{3.6}$$

is a minimum and $(\hat{x}, \hat{y}, \hat{x}', \hat{y}')$ satisfies

$$\hat{\mathbf{x}}^\top \mathbf{F} \hat{\mathbf{x}}' = 0 \tag{3.7}$$

where $\hat{\mathbf{x}} = (\hat{x}, \hat{y}, 1)^\top$ and $\hat{\mathbf{x}}' = (\hat{x}', \hat{y}', 1)^\top$ are the re-projected image points. $\mathbf{F}$ is the 3×3 Fundamental Matrix of rank 2. It encapsulates the intrinsic geometry of the scene.

This is computationally expensive and a simpler Singular Value Decomposition (SVD) based linear method may be used to estimate $\mathbf{X}$ from $\mathbf{x}$ (see (Hartley & Zisserman 2000) for a detailed mathematical exposition of the algorithm). Combine the equations $\mathbf{x} = \mathbf{CX}$ and $\mathbf{x}' = \mathbf{C}'\mathbf{X}$ to form $\mathbf{AX} = \mathbf{0}$, which is linear in $\mathbf{X}$ (Hartley & Zisserman 2000).

Eliminating the homogeneous scale factor using vector cross-products gives three equations for each image point. Two of these are linearly independent. For example, expanding $\mathbf{x} \times (\mathbf{CX}) = \mathbf{0}$ for the first image gives

$$x(\mathbf{c}^{3\top}\mathbf{X}) - (\mathbf{c}^{1\top}\mathbf{X}) = 0 \tag{3.8}$$

$$y(\mathbf{c}^{3\top}\mathbf{X}) - (\mathbf{c}^{2\top}\mathbf{X}) = 0 \tag{3.9}$$

$$x(\mathbf{c}^{2\top}\mathbf{X}) - y(\mathbf{c}^{1\top}\mathbf{X}) = 0 \tag{3.10}$$

where $\mathbf{c}^{i\top}$ are the rows of $\mathbf{C}$. Let $\mathbf{c}'^{1-3}$ be the three rows of $\mathbf{C}'$. Similar equations can be written for the second image as well. These equations are linear in the components of $\mathbf{X}$.

Thus an equation of the form $\mathbf{AX} = \mathbf{0}$ may be written with

$$A = \begin{bmatrix} x\mathbf{c}^{3\top} & - & \mathbf{c}^{1\top} \\ y\mathbf{c}^{3\top} & - & \mathbf{c}^{2\top} \\ x'\mathbf{c}'^{3\top} & - & \mathbf{c}'^{1\top} \\ y'\mathbf{c}'^{3\top} & - & \mathbf{c}'^{2\top} \end{bmatrix} \tag{3.11}$$

Two equations from each image are included, resulting in four equations in four homogeneous unknowns. These equations are however, redundant since the solution of $\mathbf{X}$ can only be determined up to scale.

The solution of $\mathbf{X}$ is obtained using the SVD of $\mathbf{A}$ and coincides with the unit singular vector corresponding to the smallest singular value. Specifically, if $\mathbf{A} = \mathbf{UDV}^\top$ and $\mathbf{D}$ is the diagonal matrix with positive entries arranged in non-increasing order along the main diagonal, then $\mathbf{X}$ is the last column of $\mathbf{V}$.

Note that this method of triangulation is not suitable for projective reconstruction (it is not projective-invariant). Although it does not employ any form of geometric error minimisation or re-projection error minimisation, it still gives acceptable results. Furthermore, it has the advantage that it generalises easily to triangulation when more than two views of the point are available (Hartley & Zisserman 2000).

3.7 Reconstruction Ambiguity due to Uncalibrated Cameras

Without some knowledge of a scene's placement with respect to a 3D co-ordinate frame, it is generally not possible to reconstruct the absolute position or orientation of a scene from two views, or in fact, any number of views (Hartley & Zisserman 2000). This is independent of any knowledge which may be available about the internal parameters of the cameras or their relative placement.

If nothing is known of the calibration of either camera, nor the placement of one camera with respect to the other, then the ambiguity of reconstruction is expressed by an arbitrary projective transformation.

Informally, the basic theorem of projective reconstruction from uncalibrated cameras may be stated as follows:

> If a set of point correspondences in two views determines the fundamental matrix uniquely, then the scene and cameras may be reconstructed from these correspondences alone, and any two such reconstructions from these correspondences are projectively equivalent.

Techniques to change the transformation from projective to affine and then affine to metric exist (Sturm & Quan 1995, Horaud & Csurka 1998, Ruf et al. 1998). This can be achieved in two ways. The *Direct* method involves computing the homography from five or more ground points with known Euclidean positions (Hartley & Zisserman 2000). The transformation is simple and easy. The system however, is not practical or realistic. For example, if the cameras are displaced slightly then one or more of the ground points might not be in view any longer. This would hinder the metric reconstruction, unless the number of ground points is considerably greater than five, so that no matter how much the cameras are perturbed, at least five of the points are always in view. For the purposes of a face recognition system, these fixed points would be in the background (the subject would be in the foreground) and may well be occluded by the subject. Another problem with this method is that determining the Euclidean co-ordinates of points in 3-space is not always that simple.

The *Indirect* or the *Stratified* method involves computing the plane at infinity first for the affine reconstruction, and then computing the image of the absolute conic for the metric reconstruction (Hartley & Zisserman 2000). These are again non-trivial tasks requiring 3D scene information such as parallel lines in the image. It is not always possible to find

these in the image of a face. Thus, going from a projective to an affine and then to a metric reconstruction poses an added challenge to the task of 3D reconstruction from point correspondences alone.

3.8 Calibrated Vs. Uncalibrated Setup in a noisy environment

A calibrated set up refers to one in which both intrinsic and the extrinsic parameters of the cameras are known or set *a priori*. In an uncalibrated setup these parameters have to be inferred from the correspondences between the two images.

Using 2D intensity images introduces noise in the data in a number of ways. Camera images are very sensitive even to the slightest variation in illumination. In digital images, scene information (such as features) is contained in the pixel values and the ratio of these values to each other. Very small changes in illumination can cause significant changes in the pixel values. If the object is not illuminated uniformly, some parts of it appear darker or lighter than other parts. This causes the ratio of pixel values to be altered drastically and images of the same object can appear vastly different.

Cameras can introduce noise in the data in two ways. Unless they are manufactured with utmost precision and accuracy, the intrinsic parameters of the cameras can vary slightly around their documented values. As a result of these inaccuracies, no two commercially manufactured cameras are going to be exactly the same. These variations in the camera parameters can distort the images slightly. These distortions may not necessarily be visible to the human eye, but are easily picked up by the computer in the form of changed pixel intensity values. In addition, optical and mechanical misalignments in the lens system of the camera can cause the position of the principal point to vary quite significantly as the zoom position and the lens focus change (Burner 1995, Wilson & Shafer 1993). This kind of noise can cause the matching algorithms to perform poorly and may result in an increased number of mismatches among the correspondences.

The other source of camera noise manifests in the camera matrices. How the noise is introduced in these depends mostly on whether the cameras being used are calibrated or uncalibrated. Both of these generate noise in different ways.

When using an uncalibrated rig, it is in theory possible to calculate the camera matrices accurately from the set of correspondences between the two images alone (Hartley & Zisserman 2000, Shashua 1994, Hartley et al. 1992). In practice, however, the correspondences themselves are riddled with mismatches. The mismatches are a consequence of the ill-posed nature of the correspondence problem and distortions in the images caused by imprecise cameras. For each point in the left image, there can be many potential matches in the right image, which may or may not contain the correct match. A choice of incorrect matches can lead to incorrect camera matrices, and therefore incorrect reconstruction of the scene. Robust algorithms are required to not only establish potential correspondences, but also to eliminate the incorrect ones. In addition, without any prior knowledge of the scene or the cameras setup, the scene can only be reconstructed projectively. This means, that different sets of perfectly valid correspondences can lead to perfectly valid and equivalent projective reconstructions of the same scene. These reconstructions can however look very different as the ratio of areas, lengths and angles is not preserved (Hartley & Zisserman 2000). This

can make it very difficult to differentiate between noise and projective ambiguity. The computational load of the algorithms increases significantly and the system is more likely to be numerically unstable (Faugeras et al. 1992).

The calibrated rig also has its downfalls. It is not always practical or easy to obtain the camera parameters. If a stereo-based system is to be implemented in a real-world setting then it is highly likely to be subject to minor camera displacements. In such a situation, unless the cameras are calibrated again, the reconstruction algorithm at best might produce very erroneous results and at worst result in a failure of the entire algorithm.

The calibrated system has two main advantages over the uncalibrated system. Firstly, the projective ambiguity is eliminated. Unlike algorithms for the uncalibrated setup, the camera matrices and reconstruction in the calibrated setup do not rely on numerous random combinations of correspondences to eliminate noise. As a result, each time a scene is reconstructed using a given set of correspondences and camera matrices, it will look exactly the same. The need to establish projective equivalence is eliminated and the processing time is reduced. Secondly, the assumption that the two cameras used are identical can be discarded. Thus, individual distortions in each camera can be modelled in the camera matrices and the resulting reconstructions also take into account these differences, thus eliminating further noise from the reconstructions. This assumption is common in reconstruction algorithms for uncalibrated cameras (Hartley & Zisserman 2000). Algorithms that treat each camera individually exist, but some knowledge of the scene is required.

The choice of setup depends largely on the application. The uncalibrated setup and the projective approach to structure from motion was introduced to get round the problem of calibrating the cameras precisely (Faugeras 1992, Hartley et al. 1992). Uncalibrated cameras lead to a projective reconstruction, while calibrated cameras result in Euclidean reconstruction. Oliensis and Govindu compare the projective and the Euclidean approaches to reconstruction in (Oliensis & Govindu 1999). Their findings are summarised below.

The projective reconstruction is equivalent to the Euclidean except that the linear camera calibration is treated as unknown and potentially arbitrarily different in each image. The projective approach is unrealistic in that it allows for arbitrary linear calibration errors but neglects potentially significant non-linear camera distortions. Consequently, there is a significant loss in accuracy. This is true even for the most robust and optimal projective reconstruction algorithms. Even if only a few of the calibration parameters are known, the Euclidean reconstruction will be more accurate than the optimal projective reconstruction.

Oliensis and Govindu (Oliensis & Govindu 1999), state that Euclidean structure-from-motion techniques, assuming a single camera of unknown calibration, recover the projective structure more accurately than the projective structure-from-motion approaches do. Thus, inaccurate knowledge of camera parameters may not necessarily justify a projective approach to structure from motion. Their experiments with the optimisation techniques show that the local minimum problem appears less severe for the projective approach than for the Euclidean approach. They are keen to emphasise that this finding does *not* imply that the projective framework is more robust than the Euclidean one, since the projective approach is simply a generalisation of the Euclidean.

Between these two extremes lies the partially calibrated setup. The external camera parameters are subject to noise and perturbations and are often unknown. The internal parameters however are often known, or can be determined if the camera model is known. Both the aspect ratio (α_u/α_v) and skew s of the pixel co-ordinate axes are very stable over

long periods of time and so can be taken as known values in the camera matrices (Sturm 1997). The position of the principal point depends on the zooming position and the lens focus. Thus, there is a certain interdependence between the intrinsic camera parameters, which can be modelled into the self-calibration techniques used in the uncalibrated setup (Sturm 1997). This partial calibration reduces the number of unknowns and can improve the accuracy of the estimated parameters in the uncalibrated setup.

3.9 Summary

This chapter has outlined some of the major aspects of the 3D reconstruction process from stereo images such as the epipolar and the camera geometry, triangulation and the correspondence problem. A minimum of two images of the scene to be reconstructed, acquired from slightly different positions are required. The depth information is encompassed in the disparity, or the difference between these two images. Stereo vision from two images mimics the human binocular vision.

Computing depth from 2D images requires knowledge of the camera geometry and pairs of corresponding points from the images that map to the same point in the 3D space - the epipolar geometry. The 3×4 camera matrix is essential for the reconstruction process and can be obtained either by manually calibrating the cameras or by inferring the camera parameters from the correspondences and the knowledge of the scene. Manually calibrating cameras is more accurate, robust and numerically stable than the uncalibrated approach. In addition, it produces a Euclidean reconstruction rather than the projective reconstruction resulting from uncalibrated cameras. It is possible to obtain a Euclidean reconstruction from the projective reconstruction, but it is a lengthy, computationally intensive and an error-prone procedure. Calibrated setup involves the use of a calibration object and extensive human input, which in itself can be a major source of errors.

In addition to the camera geometry, pairs of points from the two images that map to the same point in the 3D space are required. The idea of the Correspondence Problem is introduced. A detailed exposition of the problem and some of the solutions are presented in Chapter 6. Depth can only be extracted for the points that are visible in both the images and for which correct correspondences have been established. This task is extremely complex and a solution for all classes of images does not exist. Once the camera geometry and the matched points are available, depth can be extracted via triangulation. The linear triangulation procedure used in this work is also briefly described in Section 3.6.4.

CHAPTER 4

Data Acquisition and Processing

4.1 Introduction

It is often found that there is a significant variation in the results reported in literature for the same algorithm. A good example of this is Turk and Pentland's Eigenfaces algorithm (Turk & Pentland 1991*b*). This algorithm is often used as a benchmark and different results are reported depending on the dataset, the pre- and post-processing techniques and the distance measures used (Zhao et al. 2003). A comparative study of this is presented in (del Solar & Navarrete 2005, Yambor et al. 2002).

There has been some research on the subjects of data collection, processing and representation techniques (Craw et al. 1999). Some of the important questions that they attempt to address include:

- The size of the data set, number of training images and the number of test images that should be used in order to test a particular technique.

- The number of images in each class, i.e., the optimum number of images per individual in the dataset.

- The best poses and head orientations to include in the training images to maximise reconstruction and recognition accuracy and minimise redundancy in the dataset (Lee et al. 2004).

- The controls and the restrictions to be imposed on the imaging environment (illumination, clutter, kind of background, etc.) and the image content (head orientations, poses, expressions, scale, etc.).

- The type and degree of pre-processing to apply to images in order to optimise the performance of the techniques being investigated. For example, Heseltine investigates normalisation and convolution methods in (Heseltine et al. 2002, 2004*b*,*c*) and Adini investigates how to best compensate for illumination variation (Adini et al. 1997).

- The best representation for the data to minimise redundancy and storage costs without compromising performance of the system. For example, reduce the image to edges only (Canny 1986), coefficients of Gaussian basis functions (Edelman et al. 1992), etc.

Datasets and collection methods in face recognition has largely been driven by applications or by the particular aspect of face recognition that is being investigated (e.g. pose or illumination invariance). Although there have been some attempts to establish the criteria for ideal dataset (both training and test), lack of sufficient research means that there is no established methodology or best practise for data collection in face recognition. This work aimed to investigate the usefulness of depth information in face recognition. Therefore, it was essential that the dataset was kept as realistic as possible so that the results of the study would also be applicable outside the laboratory conditions. However, some controls also needed to be exercised so that the data could be analysed in a meaningful way without limiting the scope of the research.

This chapter describes in detail the Sheffield Dataset used in this work, how it was compiled and processed prior to being used for reconstruction and recognition. It concludes with a description of the camera calibration process and details of the camera parameters.

4.2 Data Set

The main features of the dataset used in this project (referred to hereafter as the Sheffield Dataset) are:

- Data collected in August 2004. All the images of an individual were captured on the same day. This was to avoid temporal effects since there is some evidence (Chang et al. 2003) that classifier accuracies can vary for images captured with time delay of few days.

- It consists of 22 individuals - 11 males and 11 females, to avoid any kind of gender bias.

- Subjects belong to various ethnic groups including Caucasian, Afro-Carribean, Asian, Oriental and South-American.

- The ages range from early 20's to late 40's.

- 7 out of the 22 subjects wear glasses. 6 of the individuals who wear glasses are photographed both with and without glasses.

- 2 of the females are photographed with a head-scarf. Of these, one is photographed without the head-scarf as well. She is also photographed both with and without the glasses (see Figure 4.1).

- The database consists of a total of 692 images.

Figure 4.2 displays the different classes in the dataset. Approximately 23 images were captured for each individual. These images were grouped in various categories (see Figure 4.3 for a sample of images captured for each individual):

- Frontal: 1 image with no rotation and a neutral expression.

Figure 4.1: One of the subjects in the dataset pictured with and without glasses and, with and without a head-scarf.

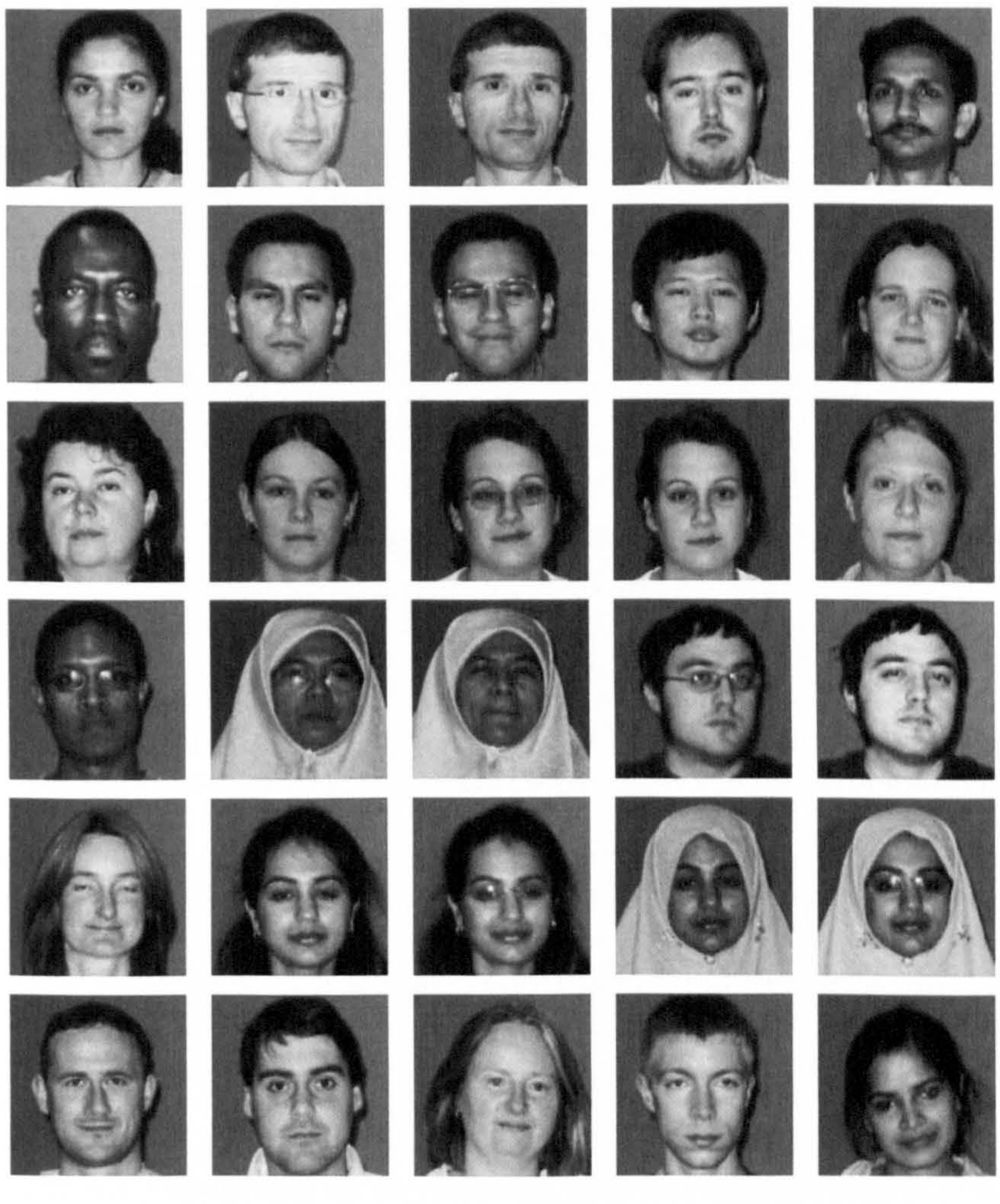


Figure 4.2: The reduced dataset depicting the first image in each of the 30 classes.

- Rotation1 and Rotation2: 7 images with neutral expression and head rotation about the y-axis (axis of symmetry), approximately 12°-15° apart. Subjects were instructed to face reference points placed strategically around the room. The exact degree of rotation was not strictly enforced as a reasonable balance between totally controlled and uncontrolled environments was required. The expression for these images was neutral as they were to be used to test the pose-invariance of the recognition algorithms.

  Rotation1 and Rotation2 contain 3-4 images with rotation about the y-axis. As the degree of rotation from the frontal image increases, the identifiable features (such as eyes and nose) become less visible and the non-identifiable features (such as ears and cheeks) become more visible. Images with fewer identifiable features are harder to classify correctly and are grouped together in Rotation2. The rotation images are partitioned manually, depending on their information content.

  Since none of the subjects' faces are noticeably asymmetrical, only one side of the face is imaged. It is assumed that imaging both the sides of the face in the same manner would serve only to introduce redundancy in the data.

- EyesClosed: 1 frontal image with eyes closed. This is a commonly arising situation in photographs. It is usually a consequence of the flash being too bright or the subjects blinking at the time of image capture. Again, the expression is neutral.

- Smile: 1 frontal image with smile. This is one of the most common expressions that can alter the shape of the eyes, cheeks and the mouth significantly, with the added appearance of the teeth!

- EyesCloseLookUp and EyesCloseLookDown: 4 frontal images with rotation about the x-axis and eyes both open and closed. Face recognition literature that discusses pose variation or different head orientations always refers to rotation about the y-axis as this is the most commonly occurring situation. For a face recognition system to be of value in applications such as crowd surveillance, it has to be able to cope with very generic face images. When the head is tilted up or down, its appearance can alter significantly as more or less of the neck and head may be visible. Little is known about this class of face images. Its effects on recognition rates in the partially-controlled laboratory environment are investigated using these images.

- RotationXYLookUp and RotationXYLookDown: 2 images with rotation about both x and y axes, i.e., looking up or down and not looking directly at the camera. It was left up to the subject to decide which direction they would face. Again such images, although common in most real-life situations, are rare in literature.

- Expression: 5 images with random expressions. During the capture of these images, subjects were given complete freedom of expressions. They chose anything from a simple frown to blowing up the cheeks and yawning! These images test the expression-invariance of the algorithms and they also serve to introduce a degree of randomness into the dataset. The aim of this work is to test the usefulness of depth information for a generic face recognition system, i.e. a face recognition system that can process face images with a wide variation in pose and expressions.

- Lighting: 2 images with "harsh" illumination. It is known that most 2D recognition systems perform poorly in the presence of varying illumination. As a result, these systems are heavily reliant on normalisation and techniques that compensate for varying illumination (Adini et al. 1997). If the change in lighting is uniform, then these techniques work well, as intensity can be equalised. However, if the face is lit very brightly from one particular direction, such that it appears as if the contours of the face have been altered, then normalisation techniques also fail. The subjects' faces are illuminated very brightly from the left and from underneath for these images. The usefulness of depth information in such cases in not known.

Figure 4.3: A sample image class from the dataset D_1.

Every attempt was made to ensure that the dataset is as versatile as possible. Although the total number of individuals in the dataset is quite small compared to some of the other publicly available datasets (such as the FERET database), it has many unique features. For example, it allows the user to test the effects of changes in pose, both, about the x and y axes. This is not commonly found in literature. It also allows the user to test the effects of non-uniform illumination and controlled (such as smiles and eyes open and closed) as well as uncontrolled expressions. In addition, the effects of accessories such as glasses and headscarves can also be investigated. All of these can be explored individually or in conjunction with each other, as is done in this work.

There are numerous datasets that allow the user to investigate the effects of many of these image types on face recognition in 2D and 3D spaces (Tolba et al. 2005). However,

until recently, there was no single dataset that allowed the user to compare *both* 2D and 3D recognition algorithms. Algorithms are generally tested on different datasets and attempts are made to compare the results.

The Sheffield Dataset also serves the stereo matching and reconstruction communities. Stereo matching algorithms that yield dense and accurate disparity fields for face images are not very common since faces are a very complex class of 3D objects - they are rigid yet deformable. Admittedly, the lack of ground truth data can make it difficult to provide results that are quantitatively comparable. However, the availability of the calibration images gives the user some control over the accuracy of the calibration parameters, and hence the reconstruction.

The dataset also has many drawbacks. As mentioned earlier, it is small in comparison with some of the publicly available datasets (such as the FERET database[1] and the Face Database of the Max Planck Institute for Biological Cybernetics[2]). It includes subjects from various ethnic origins, but many more subjects are required to draw concrete conclusions about the performance of the chosen algorithms for these various groups. Lack of ground truth data is an obvious problem for the shape-from-stereo algorithms comparison. It may also be argued that it does not strike a good balance between controlled and uncontrolled environments, or that the lack of ground truth data and accurate reconstructions does not allow a fair comparison of recognition algorithms in two, two-and-a-half, and three dimensional spaces. However, to an extent, this was intentional. The goal of the project is to investigate whether or not having the depth information improves the recognition rates in the presence of noise, and it is felt that the dataset serves this purpose adequately. It is true that it has a long way to go before it can reach the standards of the other databases such as the FERET, but it serves as a good starting point which can be built upon.

4.2.1 Sheffield Data Sub-Sets

The Sheffield Dataset is referred to as D_1 in the remainder of this thesis. This notation is introduced purely for convenience. Two further subsets of D_1 are also used for some of the experiments.

Dataset D_2 is a small subset of the Sheffield Dataset D_1. It consists of 165 images and is compiled using five or six representative images from all 30 classes (see Figure 4.2). Figure 4.4 is an example of the images contained in one of classes in this dataset. This dataset is used for feasibility studies prior to applying a new technique on larger datasets D_1 or D_3. A good performance on this smaller dataset was taken to be an indication of the viability of the approach being tested.

Dataset D_3 is a larger subset of D_1, consisting of 540 images. There are approximately 18 images (see Figure 4.5) in each of the 30 classes. This dataset is formed by removing three image categories from the dataset: Rotation2, RotationXYLookUp and RotationXYLookDown. In some of the images from these categories, very little of the actual face can be seen because the subject has either turned too far away from the camera or because a large part of the face is covered by the hair or the head-scarf.

[1] www.itl.nist.gov/ad/humanid/feret
[2] http://faces.kyb.tuebingen.mpg.de

Figure 4.4: A sample image class from the dataset D_2.

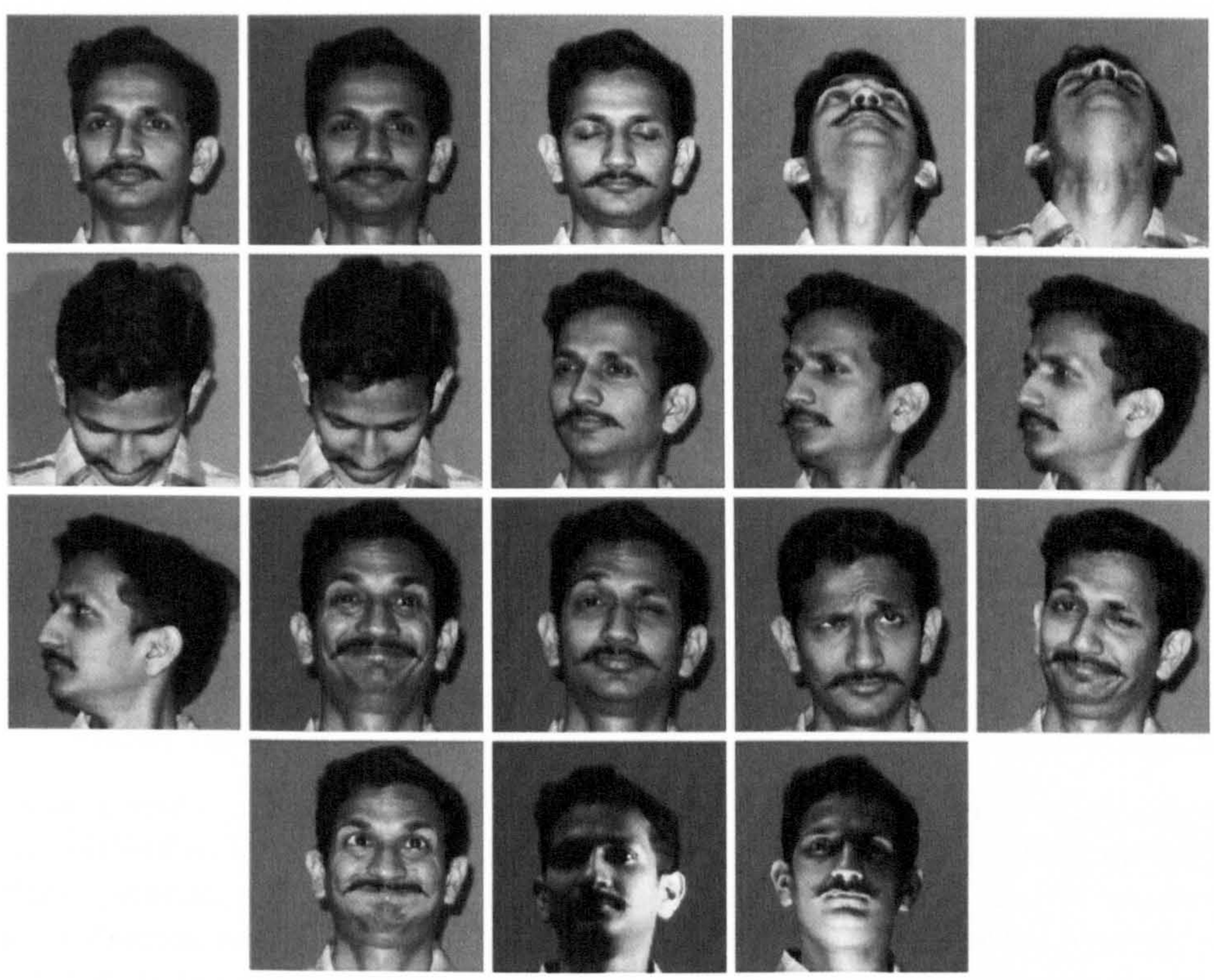


Figure 4.5: A sample image class from the dataset D_3.

4.3 Set-up

The subjects were seated approximately 60cm from the smooth, monotonic background. This distance was arrived at through a process of trial and error. Positioning the subjects at a distance from the matte background ensured adequate dispersal of shadows. Effort was made only to minimise the shadows, not to eliminate them completely. This was again in order to strike a balance between totally controlled and uncontrolled environments. On occasions, due to the subject moving around slightly, shadows were generated. These images were not eliminated from the database as it was felt that they would be useful in testing the robustness of the matching and the reconstruction algorithms. The cameras were placed approximately 300cm from the subject, and were separated by a horizontal baseline of 22cm.

All natural light was barred from the room - the curtains drawn shut. The room was illuminated using fluorescent lights and the subjects were positioned such that their faces were illuminated uniformly from all directions, as far as the eye could tell. Fill-in flash was left on. This ensured good contrast in the images. The flash was switched off completely when taking the "harsh lighting" pictures. The effects of harsh lighting are simulated using an ordinary 40W desk lamp. The lamp was held very close to the left side of the face and very close to the chin to illuminate the face from the side and from underneath. Digital zoom is switched off so that the focal length of the lenses remained constant across all the images. Digital zoom can be left on if a calibration object is also imaged along with the subject to estimate the camera parameters.

A pair of digital cameras (Olympus Camedia C200Z) are used to capture 1280×960 RGB images. These images are then individually cropped so that they contained only the faces and minimal background. Both left and right images are cropped equally so that there is no loss of stereo information. The images are initially cropped to 512×512, after which they are further reduced to 256×256 using bicubic interpolation. Two of the matching algorithms are wavelets based, and so at each level of decomposition the images are decimated by $2^n \times 2^n$. Hence, as a matter of convenience, the images are chosen to be of size $2^n \times 2^n$. The image size is reduced from 512×512 to 256×256 as this speeds up the computation, without any noticeable reduction in the recognition accuracy during the software testing phase. This is also echoed in the findings of Zhang (Zhang 2003). He found that although the facial recognition performance becomes better with improvement of the image quality, it does reach a certain ceiling which cannot be exceeded, irrespective of the image quality.

The images undergo no other form of pre-processing or normalisation. There is significant evidence that normalisation and the application of certain pre-processing techniques improve the performance of face recognition systems. However, the recognition accuracy varies according to the pre-processing technique that is applied. Furthermore, the normalisation and the pre-processing are often dictated by the dataset being used (Heseltine et al. 2002). In an uncontrolled environment such as a crowd surveillance system, the dataset is not necessarily known a priori and as a consequence, all forms of normalisation and pre-processing (except resizing) were avoided. This also limited the number of variables in the system, which was preferable at this early stage of research. Refraining from such pre-processing also allows the discriminatory power of actual recognition algorithm (without the aid of pre-processing techniques) to be analysed.

4.4 Camera Calibration

Jean-Yves Bouguet's "Camera Calibration Toolbox for Matlab"[3] is used to calibrate the two cameras. The calibration process involves photographing a chess-board like calibration object in different orientations (as shown in Figure 4.6) and then verifying the corners identified by the software.

Once the corners have been extracted by the toolbox, the user is asked to identify four points on the corners of the grid. These points act as a frame of reference for the grid, and the order of clicking allows the toolbox to compute the orientation of the calibration object. The structure of the calibration object (i.e. the positions of the corners on grid) is known, so the position of the corners can be expressed in terms of this reference frame. The extracted corners are displayed and the user is asked to verify these. If the putative corners do not match the actual corners on the grid, the distortion and the skew parameters are adjusted to ensure that the two line up. This process is repeated for all the calibration images. Once this is complete, the toolbox automatically computes the intrinsic camera parameters (focal length, principal point, skew and lens distortion) using established camera geometry (Faugeras 1993), along with predicted errors.

Both the cameras are calibrated individually first, and then together. This allows the extrinsic camera parameters $\mathbf{R}$ and $\mathbf{T}$ to be computed. $\mathbf{R}$ and $\mathbf{T}$ represent the relative rotation and translation of the right camera with respect to the left. Consider a point $\mathbf{X}$ in the 3D space with co-ordinate vectors $\mathbf{x}$ and $\mathbf{x}'$ in the left and right camera reference frames respectively. Then, $\mathbf{x}$ and $\mathbf{x}'$ are related to each other through the rigid motion transformation

$$\mathbf{x}' = (\mathbf{R} \times \mathbf{x}) + \mathbf{T}$$

where $\mathbf{R}$ is a 3×3 rotation matrix and $\mathbf{T}$ is a 3×1 translation vector.

The calibration process is not covered in detail here and the interested reader is directed to Jean-Yves Bouguet's webpage[3].

4.4.1 Camera Matrices

Recall from Section 3.6.2 the camera matrix (equation 3.3):

$$\mathbf{C} = \mathbf{KD} = \begin{pmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \left(\mathbf{R}_{3\times3} \middle| \mathbf{T}_3 \right) \tag{4.1}$$

When there are two cameras, as in a stereo setup, the left and right camera matrices are given by:

$$\mathbf{C} = \mathbf{K}_1 \left[\mathbf{I} \middle| \mathbf{0}_{3\times1} \right] \qquad \mathbf{C}' = \mathbf{K}_2 \left[\mathbf{R} \middle| \mathbf{T} \right] \tag{4.2}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the intrinsic camera parameters for the left and the right cameras respectively, and $\mathbf{R}$ and $\mathbf{T}$ are the extrinsic parameters that give the position of the right camera with respect to the left.

Matrix $\mathbf{K}$ encodes the intrinsic camera parameters and has the form given in equation 4.1. (α_u, α_v) are the scale factors in the x and y directions (or equivalently, the width and

[3] www.vision.caltech.edu/bouguetj/calib_doc/

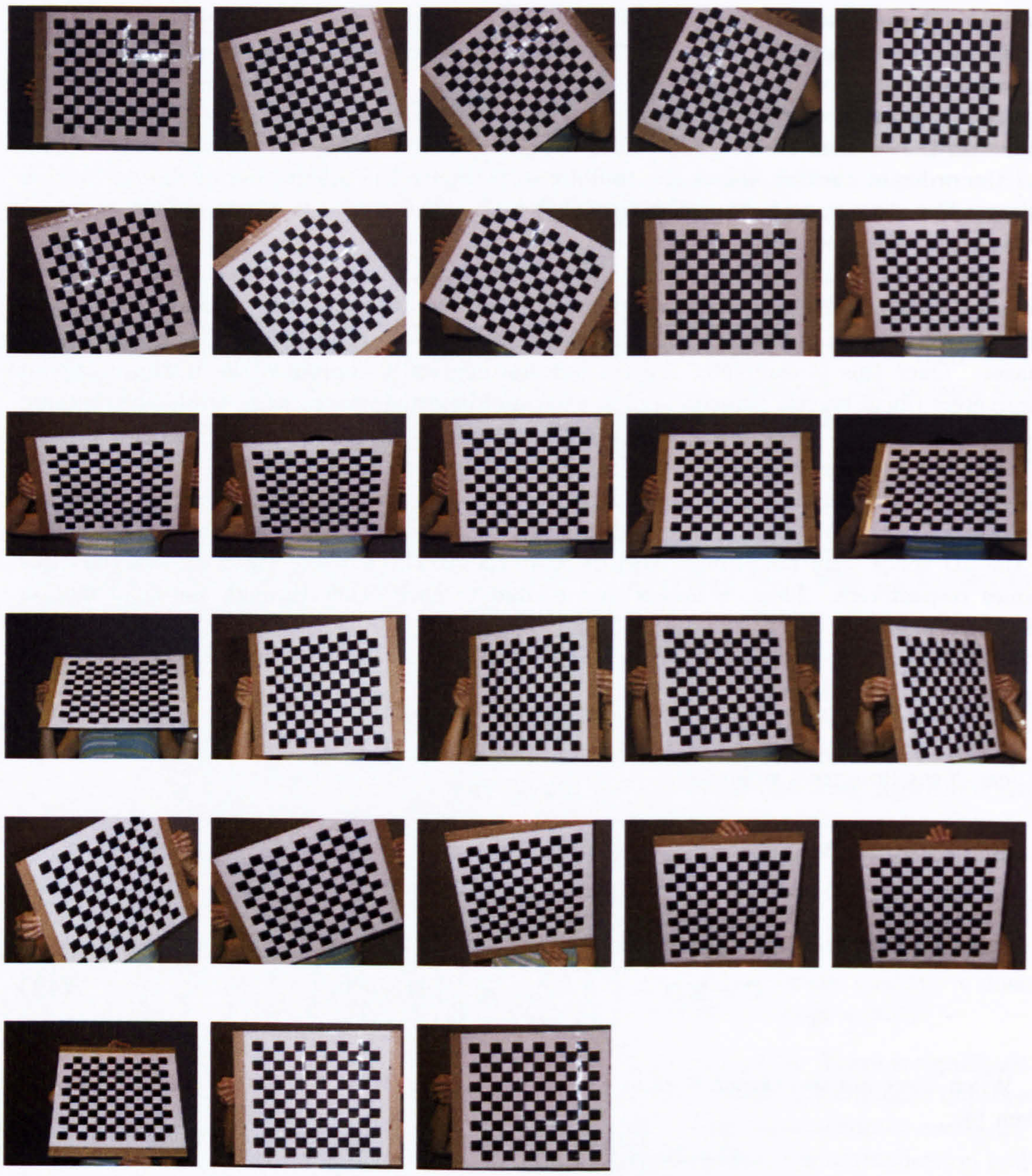


Figure 4.6: The calibration images used to calibrate the stereo cameras, using Jean-Yves Bouguet's "Camera Calibration Toolbox for Matlab".

height of a pixel, or the focal length), (u_0, v_0) are the co-ordinates of the principal point and s is the skew factor in the pixel co-ordinate system.

The camera matrices for the stereo set-up in this work are given by:

$$\mathbf{C} = \begin{bmatrix} 3811.05436 & 0 & 696.82022 & 696.82022 \\ 0 & 3829.60447 & 400.49039 & 400.49039 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.3}$$

$$\mathbf{C}' = \begin{bmatrix} 3845.55242 & -13.69196 & 1062.52133 & -786451.04710 \\ 13.722020 & 3916.78454 & 393.072111 & -11897.163528 \\ -0.051188003 & 0.023185055 & 0.99841987 & -17.3725668 \end{bmatrix} \tag{4.4}$$

The corresponding matrices $\mathbf{K_1}, \mathbf{K_2}, \mathbf{R}$ and $\mathbf{T}$ are given by:

$$\mathbf{K_1} = \begin{bmatrix} 3811.05436 & 0 & 696.82022 \\ 0 & 3829.60447 & 400.49039 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.5}$$

$$\mathbf{K_2} = \begin{bmatrix} 3895.06403 & 0 & 863.80941 \\ 0 & 3906.79300 & 482.55947 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.6}$$

$$\mathbf{R} = \begin{bmatrix} 0.9986 & -0.0087 & 0.0514 \\ 0.0098 & 0.9997 & -0.0227 \\ -0.0512 & 0.0232 & 0.9984 \end{bmatrix} \tag{4.7}$$

$$\mathbf{T} = \begin{bmatrix} -198.0569 & -0.8994 & -17.3726 \end{bmatrix}^\top \tag{4.8}$$

Note that the rotation matrix $\mathbf{R}$ is generated from

$$\mathbf{r} = \begin{bmatrix} 0.0230 & 0.0513 & 0.0093 \end{bmatrix}^\top$$

using the Rodrigues formula. The Rodrigues formula provides a convenient way of writing a rotation matrix as a rotation vector. The direction of this vector gives the axis of rotation, and its norm (or length) is the angle (or the amount) of rotation (Bouguet 1998). A rotation vector $\Omega = [\omega_x \quad \omega_y \quad \omega_z]^\top$ can be expressed as a 3×3 rotation matrix $\mathbf{R}$ by taking the exponent of

$$\tilde{\Omega} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}$$

So,

$$\mathbf{R} = e^{\tilde{\Omega}}$$

(Bouguet 1998). Note that Ω is a skew-symmetric matrix which makes it possible to compute $\mathbf{R}$ in closed-form using Rodrigues' formula (Faugeras 1993):

$$\mathbf{R} = I_{3\times3}\cos(\theta) + \tilde{\Omega}\frac{\sin(\theta)}{\theta} + \tilde{\Omega}\tilde{\Omega}^\top\frac{1-\cos(\theta)}{\theta^2}$$

where $\theta = \| \Omega \| = \sqrt{\omega_x^2 + \omega_y^2 + \omega_z^2}$, $I_{3\times3}$ is the 3×3 identity matrix and $\tilde{\Omega}\tilde{\Omega}^\top = \begin{bmatrix} \omega_x^2 & \omega_x\omega_y & \omega_x\omega_z \\ \omega_y\omega_x & \omega_y^2 & \omega_y\omega_z \\ \omega_z\omega_x & \omega_z\omega_y & \omega_z^2 \end{bmatrix}$

(Bouguet 1998).

The numerical errors corresponding to the these variables are approximately three times the standard deviations. They are computed automatically by the camera calibration toolbox, and are given in Table 4.1

Parameter	Error (%)	Mean Error (%)
$(\alpha_{u1}, \alpha_{v1})$	$\pm(2.5651, 2.5775)$	2.57
$(\alpha_{u2}, \alpha_{v2})$	$\pm(9.4139, 22.0799)$	2.45
(u_{01}, v_{01})	$\pm(9.4139, 22.0799)$	15.75
(u_{02}, v_{01})	$\pm(19.34400, 15.71865)$	17.53
T	$\pm(0.0173, 2.4382, 2.3057)$	1.59
r	$\pm(120.8696, 89.8635, 17.2043)$	75.98

Table 4.1: Camera Parameters and their associated errors

The highest errors are observed in the computation of the principal points and the rotation vector. Bouguet explains in the documentation for the toolbox that the large errors are the result of these two parameters being very difficult to estimate accurately.

Averaging the errors over the six parameters gives a mean error of 19.31%, most of which is due to the exceptionally high error value for **r**. If this is discounted, then the overall error is 8.0%.

4.5 Summary

This chapter describes the Sheffield Dataset in detail and how it was collected. The camera calibration process is described and the camera matrices are presented along with their estimated error values.

The Sheffield Dataset contains 692 images of 22 individuals (11 males and 11 females). 6 of these are photographed with and without glasses and 1 female is photographed with and without a head-scarf as well. A certain degree of control is exercised in the imaging process. For example, images with rotation about the y-axis are taken by instructing the subjects to face reference points located strategically around the room. However, since a balance between controlled and uncontrolled imaging environments was required, the subjects were given a certain degree of choice with some of the images (such as those requiring different expressions).

The main advantage of this dataset is that the same set of images can be used to test face recognition in 2D, $2\frac{1}{2}$D, 3D and the composite spaces. The camera calibration matrices and images are also available so that, in future, if required, then the 3D images can be generated again with more accurate camera matrices. In this respect, the dataset not only serves the face recognition community, but also the image matching community. The main drawbacks of the dataset are that it is not as extensive as some of the other databases (e.g. the FERET database) and the ground truth data is not available for image matching.

Cameras are calibrated using Jean-Yves Bouguet's "Camera Calibration Toolbox for Matlab". The camera matrices and the associated error values are detailed. The mean error across all the parameters is approximately 19.31%, with the highest errors being in the rotation vector and the principal points of the two cameras.

CHAPTER 5

The Correspondence Problem

5.1 Introduction

The *Correspondence Problem* refers to the search in two or more 2D images for pairs of points that are projections of the same point in the scene. Establishing a set of accurate correspondences is vital to the task of depth extraction and 3D reconstruction. Without these, the reconstruction is impossible, irrespective of whether the cameras are calibrated or uncalibrated. If the uncalibrated setup is being used, the accuracy of the correspondences becomes even more vital as all the information about the scene, the camera geometry and the epipolar geometry has to be inferred from the correspondences alone. In a calibrated setup, inaccuracies are more tolerable as some, if not all, information about the cameras is available independently of the correspondences. Assuming that the camera calibration information is accurate, the errors in the correspondences should manifest themselves as distortions in the 3D reconstruction.

Reconstructing 3D scenes from 2D images has been a major part of this work. Accurately reconstructed face models are imperative for any 3D recognition system to be successful, and the key to generating good models is a set of accurate correspondences and camera parameters. The cameras were calibrated manually and all necessary care was taken to ensure that the errors were minimal. Although it is known that camera matrices are erroneous, these errors are constant across all images and matching algorithms, and hence can be ignored. In order to make the 3D reconstruction and the recognition process as error-free as possible and to ensure that the most appropriate techniques are chosen at each stage, it is vital to evaluate these techniques quantitatively.

Quantitative evaluation techniques are rare in computer vision problems (Klette et al. 1995). Image matching is one such problem - most of the evaluation is qualitative. In (Lin & Barron 1994) Lin and Barron suggest using forward and backward image reconstruction as a means of quantitatively evaluating image matching algorithms.

Image matching is an active area of computer vision and many approaches have been put forward for solving the correspondence problem. However, these techniques are often specific to the image class they are being tested on and not generic enough to be applicable

to all classes of images. This chapter summarises some of the approaches to this problem and details the algorithms investigated in this work. It starts with some general assumptions that are adopted by most stereo vision. This is followed by a short review of the various approaches to the problem of stereo correspondence. Again, the literature in this field spans many decades and it is impossible to review all the techniques here. Evaluation techniques for stereo matching algorithms are described in Section 5.4. Section 5.5 describes two feature detection and feature matching algorithms investigated in this work. Results are presented in Section 5.6.

5.2 General Assumptions Behind Stereo Vision

In (Pan 1996*a*), Pan goes into considerable detail about the assumptions that are made, either explicitly or implicitly, by most stereo matching algorithms, and specifically by his own. These are outlined below.

1. **Scene surfaces: Static and opaque**
   The imaged scene is *static* relative to the time span of the imaging for all the stereo images. Surfaces are completely *opaque*, not transparent or semi-transparent.

2. **Lambertian Surfaces:**
   The appearance of the surfaces does not vary with the viewpoint (Scharstein et al. 2001).

3. **Illuminations: Natural or man-made, but non-specialised**
   Objects may be illuminated by natural light (e.g. sunshine) or by man-man panchromatic lamps. No specialised light sources are assumed.

4. **Optical medium: Transparent**
   The objects have opaque surfaces and are viewed through a transparent optical medium (i.e. camera lens).

5. **Camera geometry: Central perspective**
   The image sensors are central perspective cameras and the image plane is measured in a 2D Cartesian continuous co-ordinate system. If real cameras are different from this ideal model, then it is assumed that they are transformed properly into this model.

6. **Stereo configuration: Overlapping and orientation**
   Most stereo algorithms require a minimum overlap of 50% between a pair of stereo images. Individual algorithms may impose additional constraints on the minimum overlap (e.g. minimum overlap in Pan's algorithm is 60%) and orientation depending on the algorithm and the application.

7. **Images: Discrete, digital**
   The images are assumed to have digital representation and the intensity values are assumed to be discrete. The original images may be in the form of optical film but it is assumed that they are digitised appropriately.

5.3 Literature Review

There are many approaches to the solution for the correspondence problem. It is a problem that has existed for many decades without a definitive solution. The solutions are very much image-class (e.g. buildings, faces, hand gestures, etc.) and/or application dependent. A single, universally applicable or "gold-standard" technique does not exist, further highlighting the difficulty of the problem (Barron & Eagleson 1997, Read 2002). Since the problem has existed for many decades, many different techniques have been put forward for solving it. It is difficult to provide a detailed review of all the different approaches in this thesis. The reader is directed to sources such as (Barron & Eagleson 1997, Scharstein et al. 2001, Scharstein & Szeliski 2001) for detailed surveys of the literature in this field.

The existing techniques for stereo matching can be grouped into two main categories according to the matching primitives. *Feature-based* methods use sparse, high-level features such as zero-crossing points of the filtered image (Marr & Poggio 1979, Marr 1982, Pollard et al. 1985), connected edges (Ohta & Kanade 1985), segmented edges (Medioni & Nevatia 1985), and corner points (Nasrabadi & Choo 1992) and have accurate disparity values at the feature points (Kim et al. 1997). However, these methods need complicated processes such as edge-thinning and linking to avoid false matching and, post-processing such as interpolation to obtain full resolution disparity maps. In particular, they are inadequate when the texture of scenes is too dense or too sparse. Feature based methods are often used in the uncalibrated set-up. Initially, the sparse set of feature points (after the false matches have been discarded) is used to establish the camera geometry. The feature points may be manually defined points-of-interest. Once the epipolar geometry has been computed, the epipolar constraint is exploited to reduce the search area for the correspondences. The images are rectified (images are warped such that the corresponding epipolar lines in the two images are co-linear and any displacement between the images is along the x-axis only) using the epipolar geometry and an exhaustive search for the correspondences is conducted along the epipolar lines on the two images (Zhang et al. 1995). A brief review of the feature based methods may be found in (Torr & Zisserman 1999) and (Vincent & Laganière 2002) contains an empirical study of some feature matching strategies.

Direct Methods for motion and/or shape estimation refer to those methods which recover the unknown parameters directly from the measurable image quantities at each pixel in the image (Irani & Anandan 1999). These differ from the feature-based methods in that feature-based methods minimise an error measure that is based on distances between just a *few* corresponding features, while direct methods minimise an error measure that is based on direct image information collected from *all* pixels in the image (such as image brightness) (Irani & Anandan 1999). These methods are able to recover the dense 3D structure of the scene simultaneously with the epipolar geometry, and are generally used in uncalibrated stereo systems. For details on these methods, see (Irani & Anandan 1999).

Intensity-based methods fall under the category of direct methods and use dense, low-level features and intensity values themselves to calculate the disparities, without the use of feature extraction or interpolation (Levine et al. 1973, Moravec 1977, Matthies 1992). Intensity-based methods are very sensitive to noise and usually require some form of pre-processing. When stereo images have large disparities, it is often difficult to detect and treat a false match. Therefore, multiple scenes (Okutomi & Kanade 1993), hierarchical structure (Tezopoulos 1983) or neural networks with various constraints (Lee et al. 1994) are used to

solve these problems. Other algorithms for stereo matching include segmented region-based methods (Marapane & Trivedi 1989), phase-based methods (Fleet & Jepson 1990, Fleet et al. 1991, Jenkin et al. 1991), topological methods (Fleck 1991), tree matching (Cheng & In 1985), stereo matching using Gabor filters (Sanger 1988, Gennert & Malin 1992), probabilistic methods (Read 2002) and neural networks learned by constraints (Kontanzad et al. 1993). A good review of gradient based methods, frequency based methods and hierarchical methods can be found in (Magarey 1997).

Recently, some algorithms combining the intensity-based and the feature-based techniques have taken advantage of the reliable primitives of each technique (Ju & Naftel 1999). Weng et. al. (Weng et al. 1992) used some primitives simultaneously, i.e. intensity value, edgeness and cornerness, to determine the correct disparity, taking into account possible structural discontinuities and occlusions (Kim et al. 1997). Cochran and Medioni (Cochran & Medioni 1992) first employed the intensity based techniques which use the local variance of intensity pattern, and then obtained accurate disparities using edge-information as a feature-based primitive from the blurred disparity map. This method applies a set of constraints to identify and remove the low confidence matches, then performs surface interpolation to obtain full resolution disparity map.

The performance of stereo vision system based on the above methods depends mainly on extraction of the optimal features, high-level or low-level primitives which are insensitive to image translation and noise, and optimal fusion of these features (Kim et al. 1997). It is however, a complicated and a difficult process. Duplication or loss of information can occur due to each feature being extracted and processed separately. On fusion, the relative importance of each feature is determined heuristically.

Wavelets transforms have been applied to many tasks in image processing (Strela et al. 1995, Lina 1996, Kingsbury & Magarey 1997, Daubechies et al. 1999, Pastor et al. 1999, Wundrich et al. 2000), including multiresolution analysis (Mallat 1989*a*,*b*, Cohen et al. 1992, Wilson et al. 1992), image compression (Antonini et al. 1992, DeVore et al. 1992) and singularity detection (S. G & Hwang 1992, Mallat & Zhong 1992). Two-dimensional wavelets transform decomposes the image into a low-frequency and three high-frequency sub-bands. The low-frequency image is an approximation of the original, while the high-frequency components consist of the horizontal, vertical and the diagonal features (corners) (see Appendix A for details). This is a very efficient, no-loss representation of the image information. These characteristics of the transform make the feature extraction process very simple and the resulting pyramidal structure can be used for stereo matching as a coarse-to-fine strategy.

Wavelets-based methods using real-valued filters have two main drawbacks. First is that real-valued wavelets transforms are unstable with respect to translation of the input signal because these wavelets filters do not have ideal filter characteristics (Kim et al. 1997). Secondly, the distribution of energy between coefficients at different scales is very sensitive to shifts in the input data (Kingsbury & Magarey 1997), i.e. the transforms are not shift-invariant. However, many wavelets families with complex filter coefficients are now being used (Pan 1996*a*, Magarey & Kingsbury 1995, 1998*b*, Magarey 1997, Kingsbury & Magarey 1997, Kingsbury 2000*a*,*b*) to compensate for the shortcomings of the real wavelet transform. The complex wavelet families have better orientation selectivity (Magarey & Kingsbury 1995, 1998*b*, Magarey 1997) and are able to exist under the strict conditions for optimality stated by Pan in (Pan 1996*a*). These conditions are orthonormality, compact support, vanishing moments and symmetry. In addition, they produce dense disparity maps, which are preferable

for surface reconstruction.

The use of wavelet techniques to address the correspondence problem has grown steadily in the past five years or so. All the techniques take advantage of the multiresolution feature of wavelets and use it as a coarse-to-fine matching strategy with varying results. Shim (Shim 2000) proposed a technique based on multi-channel wavelet transform. The trends in wavelet coefficients are used to provide overall context throughout the framework, and the transients are used to give refined local details of the image. A locally adapted lifting scheme is used to maximise the sub-band decorrelation energy by the transients. An intra scale correlation and an inter scale backtracking technique using the Multimodality Sum of Sum of Squared Difference (MSSSD) on the transform coefficients is introduced. It is claimed that this technique provides cumulative confidence in selecting corresponding points at two contiguous analysing levels as well as within scale.

Kim et. al. (Kim et al. 1997) propose a pyramid using modified wavelet decomposition in order to achieve translation invariance in the matching process. The image transformed by the proposed method is converted into appropriate multiple features without loss of information. Since the importance of each feature is determined heuristically, it is very difficult to fuse them adequately. They propose to attach weights to each of the features depending on the similarities between the intensities in the local region of each left and right wavelet channels. The window size used for the decision of weight and disparity values influences the processed result considerably. Hence, it is chosen adaptively to ensure it is large enough to obtain a large signal to noise ratio, but not so large that it induces the effects of projective distortion. They also propose a new relaxation algorithm which can reduce false matches without blurring the disparity edge. Good performance on both synthetic and real images (e.g. collection of toys and aerial photograph of the Pentagon) has been reported.

Pan and Magarey propose a multiresolution phase-based bidirectional stereo matching algorithm in (Pan & Magarey 1999). Gabor phase is used in this scheme as a basis for dense multiresolution matching due to its stability. It is a full information matching scheme that transforms the whole image into some new domain or feature space - in this case, the Gabor-phase-based space. This matching strategy is based on the Fourier shift theorem, whereby the phase rotation in the Fourier coefficients of the signal are related to the global signal translation. Gabor phase is the ideal candidate for this strategy as its windows have optimum localisation properties in both time and frequency domains. Additionally, Gabor phase is a robust feature space that is insensitive to perturbations in illumination and affine distortion of objects. Again, a coarse-to-fine strategy is used, which effectively imposes a smoothness criterion on the final matching map without the need for explicit regularisation. The relaxation scheme used in the paper also deals with discontinuities and occlusions. A variant of this algorithm, proposed in (Pan 1996*b*,*a*) is one of the algorithms that is investigated in this work. Mathematical details are provided in Appendix C and results of its application to face images are presented in Chapter 6.

Magarey and Dick (Magarey & Dick 1998) have approached the subject of image matching in a very similar way, using Gabor phase based methods and a similarity measure to compute the disparity. However, they regularise the disparity at each level of hierarchy in order to provide a global compromise between feature similarity and disparity field continuity, resulting in feature-sensitive smoothing. They claim that the algorithm is very well suited for analysing and reconstructing facial images, though they do not report having tested it on many images.

Dick also uses complex wavelets and multiresolution and applies it to the Anandan (Anandan 1989) approach, after having reported disappointing results on his own relaxation and smoothing techniques. In his paper (Pan 1996*a*), Pan also defines some useful criteria for a robust matching algorithm. He investigates several real and complex wavelet families and suggests the use of symmetric complex wavelets as they satisfy all the conditions for optimality. The algorithm is not fully implemented yet and so the paper is somewhat sparse on results. A detailed mathematical exposition of the algorithm is covered in Appendix C.

5.4 Evaluation of Stereo-Matching Algorithms

Although there are many stereo image matching algorithms in the public domain, very few of them have been subject to quantitative evaluation (Klette et al. 1995). It has been suggested in (Barron et al. 1994) that because of difficulty in accessing accurate estimates of the 2D motion field and the ill-posed nature of the correspondence problem, only qualitative information can be extracted.

Much of the evaluation and comparison of competing algorithms is qualitative and is done by inspection, particularly if the data is real. Many techniques exist for reporting errors in disparity values for synthetic data. Some of these are presented in (Barron et al. 1994). Egnal et. al. (Egnal et al. 2002) present a technique for establishing a confidence metric for the performance of stereo algorithms using single view imagery. They search for correspondences in a pair of images from a single view, as opposed to images from two different views. Disparities significantly far from zero are erroneous as there is no motion between the two images. This yields precise quantitative performance data for real images as "ground-truth" data is readily available. This technique was investigated for the data used in this work but the results were very inaccurate and the technique was not pursued any further. According to this technique the disparity between two identical images was ¿¿10, which is obviously incorrect.

Most sources in literature adopt one of two general approaches for quantitative comparison of stereo matching algorithms. If the ground-truth data is available then comparison is done using error statistics computed with respect to this data (Scharstein & Szeliski 2001, Barron et al. 1994). Alternatively, synthetic images are obtained by warping the reference or unseen images by the computed disparity map. These images are then evaluated using appropriate metrics (usually the Root Mean Squared (RMS) error between the estimated and the actual data) (Szeliski 1999). In (Pan 1996*a*), the performance of the matching algorithms is evaluated by taking the difference between the computed disparity values and the mean disparity value in the 3×3 neighbourhood of the pixel. A threshold is set for the minimum acceptable difference. If the difference is below this threshold value then the disparity value at the pixel is taken to be correct, otherwise not. The main disadvantage of this approach is that it relies on the assumptions that the disparity values in a given neighbourhood vary smoothly and that these disparity values are correct. These assumptions are often hard to justify unless the matching technique is already known to produce good results.

In (Lin & Barron 1994), Lin and Barron use forward and backward image reconstruction using the optical flow fields to evaluate various image matching techniques. Forward and backward image reconstruction are used to generate the next image in the sequence of images for which the correct optical flows are known. The RMS differences between the actual images and their reconstructed versions are used as a metric for comparing different image matching

techniques. Lin and Barron found it to be a good indicator of optical flow errors between different methods of computation. Backward reconstruction and RMS are used in this work as a means of comparing matching accuracy between two image matching techniques when the correct disparity values are not known. The following section explains the backward image reconstruction process as described in (Lin & Barron 1994).

5.4.1 Backward Image Reconstruction

Without loss of generality, let the left image be the first in the sequence of images and the right image the subsequent image in the sequence. Let the intensity at image location (x, y) in the left image be denoted by $I(x, y)$. Let the optical flow or the disparity between the left and the right image be denoted by $\mathbf{d}(u, v)$. The new image is created by using $I(x-u, y-v)$ as the intensity at location (x, y). For example, if $\mathbf{d} = (1.43, 2.31)$ at location $(50, 50)$, then the intensity at $(48.57, 47.69)$ in the left image is used as the intensity at $(50, 50)$ in the right image. The RMS error is computed using

$$\text{RMS error} = \sqrt{\frac{\sum_x \sum_y (I(x,y) - I'(x',y'))^2}{M \times N}} \tag{5.1}$$

where $I(x, y)$ and $I'(x', y')$ are the actual and reconstructed images of size $M \times N$ (Lin & Barron 1994).

The main drawback of this method is that it assumes that illumination across both the images is uniform. In this work, this assumption is satisfied, though no special effort is made to ensure this. In situations where this may not be true (e.g. when the images are taken with cameras that are far apart, or when the images are not captured simultaneously), backward image reconstruction may not be an appropriate evaluation tool.

5.5 Algorithm Choices for Matching Face Images

Face images belong to an especially difficult class of images. The image of a face contains large areas of low-frequency regions (e.g. cheeks and forehead) and pockets of very high-frequency regions (e.g. eyes, nostrils and lips). Most algorithms are very good at identifying features in the high-frequency regions. Low-frequency areas pose a problem as there are no sharp edges, corners or significant changes in the intensity. Yet features this region do need to be located and matched correctly. Although cheeks and forehead generally lack features of distinction, they are distinct for each individual (e.g. presence or absence of prominent cheek-bones and the shape and the size of the forehead) and contain useful classification information that may be essential for the recognition in the 3D space. Consequently, an algorithm which yields dense, accurate correspondences in both high and the low frequency regions is essential.

Although it can intuitively be seen that feature based methods would struggle to identify features in low frequency areas of the face, two feature detection and feature matching methods were investigated to illustrate this point. Harris corner and edge detector (Harris & Stephens 1988) and Smallest Univalue Segment Assimilating Nucleus (SUSAN) (Smith 1995) were used for feature detection. The features identified by these were then matched using Pilu's (Pilu 1997) Singular Value Decomposition (SVD) method and Torr's (Torr 2002) correlation based matching. These methods are simple and clearly underscore the above point.

They have not been applied to the class of face images before. Both the feature detection methods claim that they are well suited to feature detection tasks in high frequency and low frequency, textured regions.

A set of dense matches is required to reconstruct 3D face images. If the correspondences are sparse but accurate, "filled-in" flow fields can be used. In (Lin & Barron 1994) Lin and Barron claim that in general, "filled-in" flow fields are more accurate than unthresholded dense flows. Linear interpolation between pairs of non-adjacent horizontal and vertical image disparities is used to fill-in the missing disparity values. However, this relies heavily on the fact that the chosen algorithm is able to correctly identify enough matches in very high-frequency regions such as the eyes, *and* it is able to absorb the salient information present in the gentle contours of the cheeks, say. This is in general difficult to accomplish even for the most sophisticated feature-based methods. This is especially true for face images as the low-frequency areas are not uniformly textured or contoured as in synthetic objects.

The chosen feature-based methods were tested against two wavelets based methods as techniques that yield dense disparity maps. Pan's uniform information matching scheme (Pan 1996*b*,*a*) and Magarey and Dick's motion estimation algorithm (Magarey 1997, Kingsbury & Magarey 1997, Magarey & Dick 1998) are tested. Both the algorithms are described in Appendices C and D respectively. Coarse-to-fine multiresolution matching and the use of complex wavelets is common to both techniques. Advantages of a wavelets-based scheme, particularly one that uses complex valued filters have been outlined in Section 5.3. In addition, Magarey and Dick's algorithm has been successfully applied to face images (Magarey & Dick 1998, Magarey et al. 1999). Pan's algorithm however, has not been applied to this class of images before. It has only been applied to aerial photographs and has reported good results.

Smallest Univalue Segment Assimilating Nucleus (SUSAN)

SUSAN, proposed by Smith and Brady, uses brightness comparisons within a circular mask (Smith 1995) centred on a pixel to identify corners. It assumes that within a relatively small circular region, pixels belonging to a given object will have relatively uniform brightness.

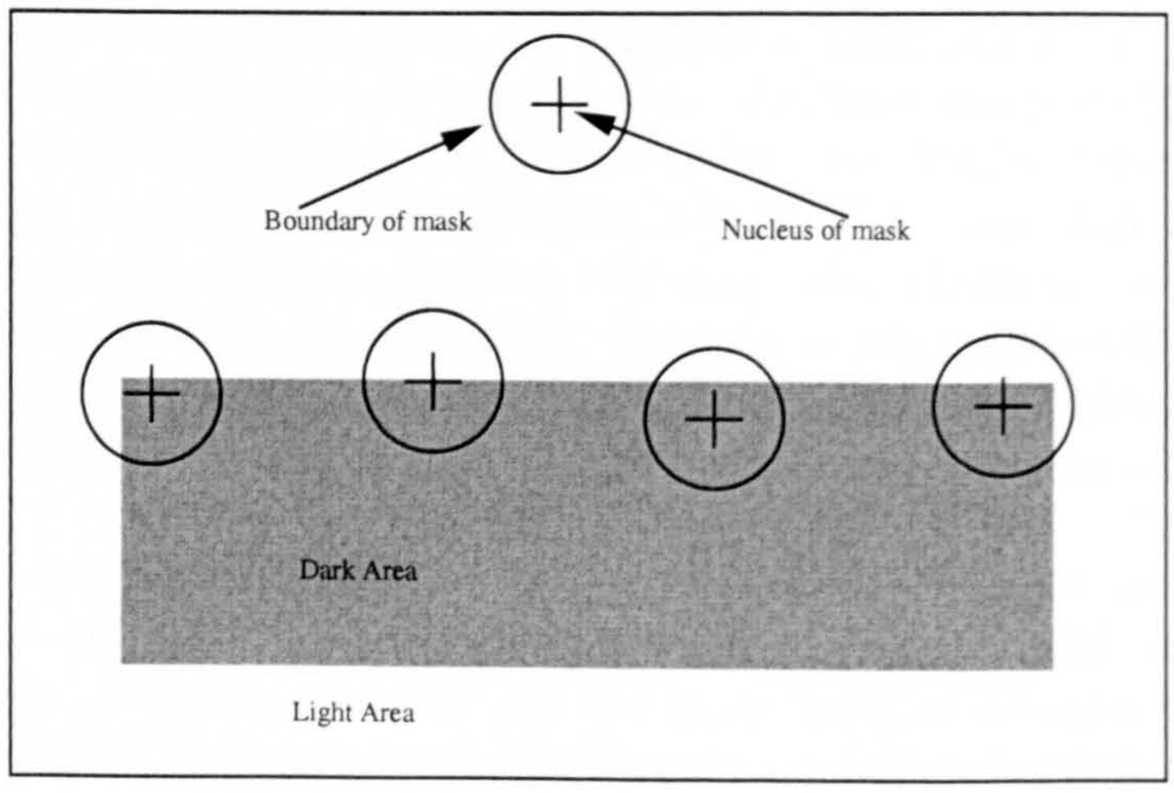


Figure 5.1: Four circular masks at different places on a simple image

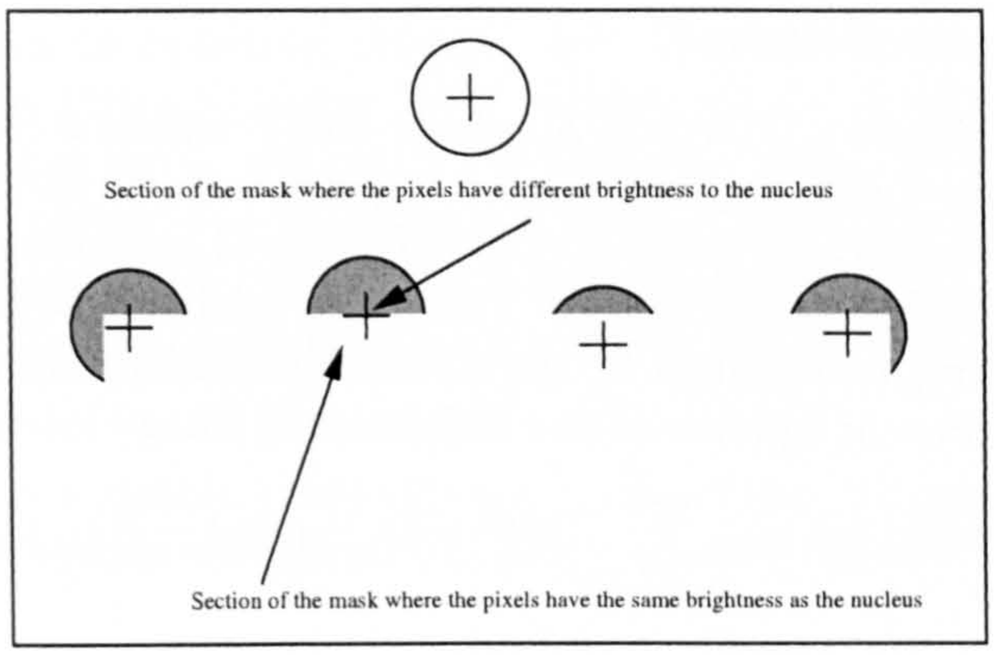


Figure 5.2: Four circular masks with similarity colouring; USANs are shown as the white parts of the masks

The intensity of the central pixel or the Nucleus, is compared with the intensity of every other pixel in the mask. This allows the segmenting of the region in the mask where the intensity is similar to the intensity of the nucleus. This region is known as the USAN, and is implemented using

$$c(r, r_0) = 100e^{-\left(\frac{I(r)-I(r_0)}{t}\right)^6}, \tag{5.2}$$

where r_0 is the position of the nucleus, r is the position of any other pixel, t is the intensity difference threshold, $I(r)$ is the intensity at position r and $c(r, r_0)$ is the output of the comparison. For each pixel in the mask, the number of pixels which have similar brightness to the nucleus are counted using

$$n(r, r_0) = \sum_r c(r, r_0). \tag{5.3}$$

Then, n is compared with a geometric threshold g, which is set at $\frac{n_{\max}}{2}$. $n_{\max}$ is the maximum value that n can take (number of pixels in the mask $\times$ the maximum value of c).

Note that the intensity difference threshold t selects the minimum contrast of corners which will be detected and the geometric threshold g controls the strength of the corners. False features are removed by imposing the condition that all the pixels between the centre of gravity of the USAN and the centre of the mask must belong to the USAN (Ghita et al. 2001).

The main advantage of this method is that it does not rely on any image derivatives and so requires no noise reduction. Every point in the input image is used as the nucleus of a small circular mask and its USAN is determined. The area of the USAN is maximised when the nucleus lies in a flat region, it gradually decreases and is halved as an edge is approached and is reduced further as the nucleus approaches a corner. The local minima of the USAN's are used to identify corners in the image. The algorithm is implemented using the publicly available code from Stephen Smith's SUSAN pages[1].

[1] www.fmrib.ox.ac.uk/ steve/susan/susanl.c

Harris corner and edge detector

Harris corner and edge detector (Harris & Stephens 1988) combines corner and edge detection to cater for image regions containing both high-frequency features and low-frequency textures. The corner detector calculates an interest operator defined according to an auto-correlation of Gaussian smoothed images (Torr 2002). There is a trade-off between the localisation of corners and noise-filtering, determined by the size of the convolution mask. Auto-correlation may be defined as the sum of squares of the difference of image intensities (Torr 2002)

$$\delta I(\delta x, \delta y) = \sum_{i,j \in \text{patch}} (I_1(i+\delta x, j+\delta y) - I_1(i,j))^2 \tag{5.4}$$

whose analytic Taylor expansion is

$$\delta I(\delta x, \delta y) = (\delta x, \delta y)\mathbf{N}\begin{pmatrix}\delta x \\ \delta y\end{pmatrix} \tag{5.5}$$

where

$$\mathbf{N}(x,y) = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{5.6}$$

The two eigenvalues of $\mathbf{N}$ are proportional to the principal curvatures of the local auto-correlation function, and form a rotationally invariant description of $\mathbf{N}$ (Harris & Stephens 1988). Large values of trace of the matrix correspond to an edge, while large values of the determinant correspond to an edge or a corner. Corner strength signal is given by

$$\Phi(x,y) = \det \mathbf{N}(x,y) - \kappa * \text{Trace}^2\left(\mathbf{N}(x,y)\right), \tag{5.7}$$

where κ is a weighting factor derived empirically.

Feature Matching Techniques

In order to establish correspondences between the identified features, two feature matching techniques are used: Torr's cross-correlation method (Torr 2002) and Pilu's SVD method (Pilu 1997).

Torr uses the difference between the image intensity over two $N \times N$ areas centred on each feature:

$$C = \sum_{i,j \in \text{patch}} (I_2(i,j) - I_1(i,j))^2 \tag{5.8}$$

where $I_n(i,j)$ is the image intensity at co-ordinate (i,j) in the n^{th} image. Every feature in one image is cross-correlated with every feature in the other image. The match with the maximum strength is stored for each corner from the first to the second image. The same process is then applied in reverse from the second to the first image. Matches are only accepted if the difference between their intensities is minimum in both comparisons. This has the effect of removing corners which are ambiguous in that they have multiple candidate matches. The main drawback of this method is that its execution time is directly proportional to the number of features identified in both the images. The Harris corner and edge detector and Torr's cross-correlation method for feature matching were implemented using Philip Torr's "Structure and Motion Toolkit in Matlab"[2].

[2]http://cms.brookes.ac.uk/staff/PhilipTorr

The second approach to matching features first identified either by SUSAN or by the Harris corner detector is Pilu's Singular Value Decomposition (SVD) approach (Pilu 1997). This method builds on the landmark paper by Scott and Longuet-Higgins (Scott & Higgins 1991), in which they exploited the properties of SVD to associate features of two arbitrary patterns. Pilu's algorithm, similarly to Scott and Longuet-Higgins's, has three stages.

Let I and J be two images containing m features $I_i (i = 1, \ldots, m)$ and n features $J_j (j = 1, \ldots, n)$ respectively.

1. Build a *correlation-weighted proximity* matrix $\mathbf{G}$ of the two sets of features where each element $G_{i,j}$ is a Gaussian weighted distance between two features I_i and J_j:

$$G_{ij} = \frac{C_{ij} + 1}{2} e^{-r_{i,j}^2 / 2\sigma^2} \tag{5.9}$$

where $r_{ij} = \|I_i - J_j\|$ and C_{ij} represents the correlation weights (the normalised cross-correlation) in the proximity matrix.

$$C_{ij} = \frac{\sum_{u=1}^{W} \sum_{v=1}^{W} (A_{uv} - \overline{A})(B_{uv} - \overline{B})}{W^2 \cdot \sigma(\mathbf{A}) \cdot \sigma(\mathbf{B})} \tag{5.10}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two $W \times W$ arrays of pixel intensities centred on features I_i and J_j. $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ are the average of $\mathbf{A}$ and $\mathbf{B}$, and $\sigma(\mathbf{A})$ and $\sigma(\mathbf{B})$ are the standard deviations of all the elements of $\mathbf{A}$ and $\mathbf{B}$ respectively. The values of C_{ij} vary from -1 for completely uncorrelated patches to 1 for identical patches.

$\mathbf{G}$ is positive-definite and the its values decrease monotonically from 1 to 0 with similarity. The parameter σ controls the degree of interaction between the two sets of features: a small value of σ enforces local interactions, while a larger value permits more global interactions.

2. Perform the *singular value decomposition* (SVD) of $G \in M_{m,n}$:

$$\mathbf{G} = \mathbf{TDU}^{\top}$$

where $\mathbf{T} \in M_m$ and $\mathbf{U} \in M_n$ are orthogonal matrices and the diagonal matrix $\mathbf{D} \in M_{m,n}$ contains the (positive) singular values along its diagonal elements D_{ii} in non-increasing numerical order. If $m < n$, only the first m columns of $\mathbf{U}$ have any significance (Scott & Higgins 1991).

3. Convert $\mathbf{D}$ to a new matrix $\mathbf{E}$ obtained by replacing its diagonal elements D_{ii} with 1 and then compute the product

$$\mathbf{P} = \mathbf{TEU}^{\top}$$

The new matrix $\mathbf{P} \in M_{m,n}$ has the same shape as the proximity matrix $\mathbf{G}$. If P_{ij} has the maximum value in both the row and the column *and* its value is above the threshold τ, then the features I_i and J_j are in $1:1$ correspondence with each other (Scott & Higgins 1991). The squares of elements in each row of $\mathbf{P}$ add up to 1, meaning that a given feature I_i cannot be strongly associated with more than one feature J_j (Scott & Higgins 1991).

The main disadvantage of this method is that a large number of matches need to be identified in both images prior to matching.

5.6 Feature Detection and Feature Matching Algorithms: Results

The feature detection and feature matching algorithms described in Section 5.5 were applied to the set of face images described in Chapter 4. As stated in Section 5.5, in order to sufficiently reconstruct the entire surface of the face, a dense set of matches is required since depth values can only be computed for the points that are visible and have been matched in both images. It was certain that feature detection algorithms would in general not be able to identify many matches in the low frequency areas such as the cheeks and the forehead. However, if a sufficient number of correct matches can be identified in the high frequency regions such as the eyes, nose and the lips, then this small number of correct matches may be better suited for reconstruction and matching purposes than a dense disparity field that is riddled with mismatches. Filling-in the flow fields as suggested by Lin and Barron in (Lin & Barron 1994) was also considered as a possibility if a few good matches in the low frequency regions could be found.

5.6.1 SUSAN Feature Detection

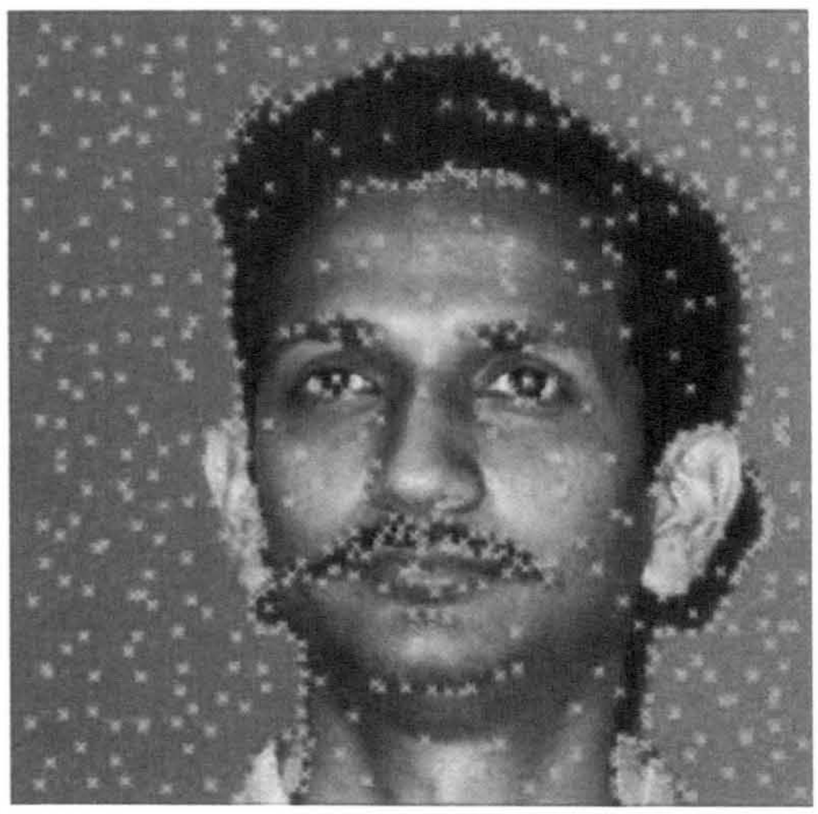
For the feature detection task in SUSAN, a 37 pixel circular mask is used, as recommended in (Smith 1995). The value of t determines the maximum difference in intensity between two pixels which allows them to be considered part of the same "region". A high value of t causes fewer corners to be detected, while a small value of t increases the number of corners that are detected. Three values of t were tested. The average number of corners detected for 300 images (150 left and 150 right images) along with associated t values is given in the Table 5.1.

t	Average number of corners detected
10	554
20	372
30	180

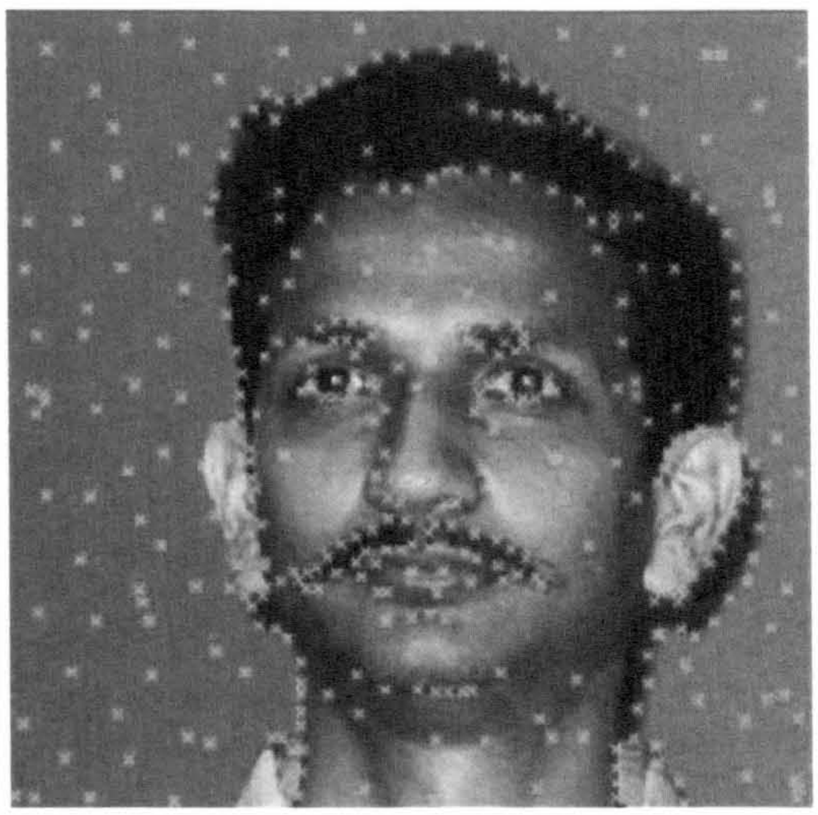
Table 5.1: Intensity difference threshold t and the average number of corners detected for 150 left and right images (total of 300 images).

Figures 5.3(a), 5.3(b) and 5.3(c) illustrate a typical set of corners detected in one of the images in the dataset.

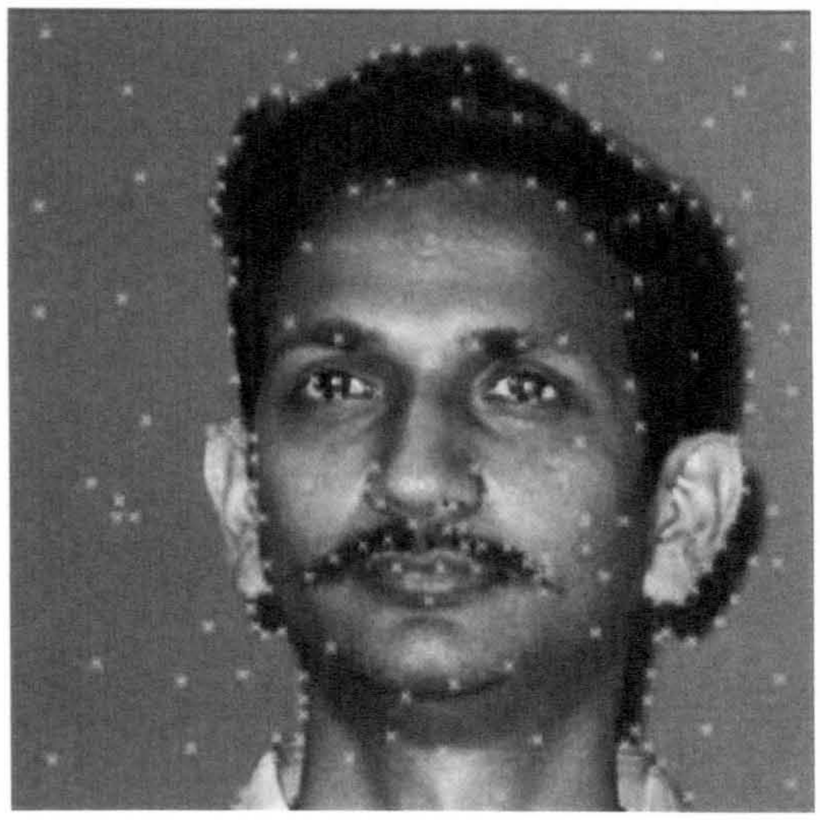
It is clear that although a low value of t does detect more corners, these are not necessarily located in regions that are useful for the reconstruction or the recognition tasks. In fact most of the corners detected are in the background and are of little use. On the other hand, a high value of t detects relatively fewer matches. It is interesting to note that a large number of matches are still detected in the background. This was not expected since it was to avoid this very situation that a matte, monotonic background was chosen for the image capture. However, it does illustrate the point made in Section 3.3 that minute changes in illumination that are not necessarily visible by eye are most certainly recorded in the digital images in the form of small changes in the intensity values.



(a) Corners detected with $t = 10$



(b) Corners detected with $t = 20$



(c) Corners detected with $t = 30$

Figure 5.3: Typical set of corners detected in the face images using SUSAN. Intensity difference threshold (t) values of 10, 20 and 30 are tested.

Matches obtained with the threshold value of $t = 20$ were thought to give a suitable balance between the number of matches and their location. These features were then matched using the two feature matching algorithms. On average, between 300 and 400 matches were fed into the matching algorithms.

5.6.2 Harris Corner Detector

The Harris corner detector requires two parameters to be chosen: the size s of the Gaussian convolution mask and σ_g, its standard deviation.

The size of the convolution mask represents a tradeoff between identifying high frequency features and eliminating noise. A large value of s results in a large number of features being found in the background rather than on the actual face, while a small value restricts the vast majority of the features to lie on the face. It should be noted at this stage that the number of corners identified was restricted to 500. Since ground truth data for the face images is not available, the detected features have to be verified manually. This is a very tedious and time-consuming task and an upper bound has to be placed on the number of matches to be identified by the algorithm.

Standard deviation, σ_g, of the mask controls the degree of smoothing applied to the images. A large value of σ_g increases the blurring effect and much of the high frequency information is lost. Figures 5.4(a) to 5.4(c) depict the effects of changes in the values of this parameter.

It is obvious that values of $\sigma_g > 1$ cause a drastic deterioration in the quality of the matches detected. This makes it easy to assign it a value of 1 since it yields matches in the best locations in the image, i.e. on the face rather than the background. The choice of a value for the size of the convolution mask is not so easy. Its effects manifest gradually and it is much harder to identify a cut-off point beyond which the quality of the matches is too poor to consider. Figures 5.4(a) and 5.4(d) show the quality of matches for the chosen value of s ($s = 5$) and an extremely large value, 50.

5.6.3 Feature Matching

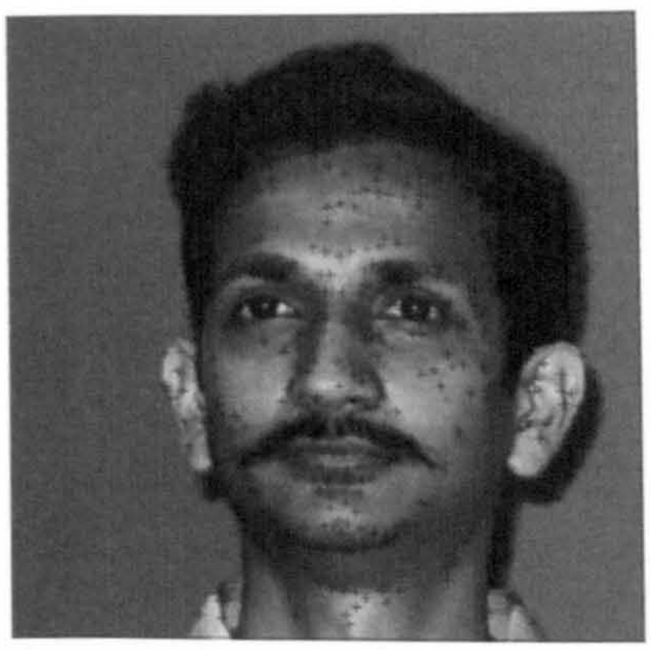
In general, both the matching algorithms, after the removal of incorrect or low confidence matches, were able to match approximately half the corners originally identified. Pilu's SVD based method requires three parameters to be chosen by the user:

- σ, the degree of interaction between the two sets of features.
- W, the size of the $W \times W$ patch centred on the features to be matched.
- τ, the minimum correlation threshold indicating a match between the features.

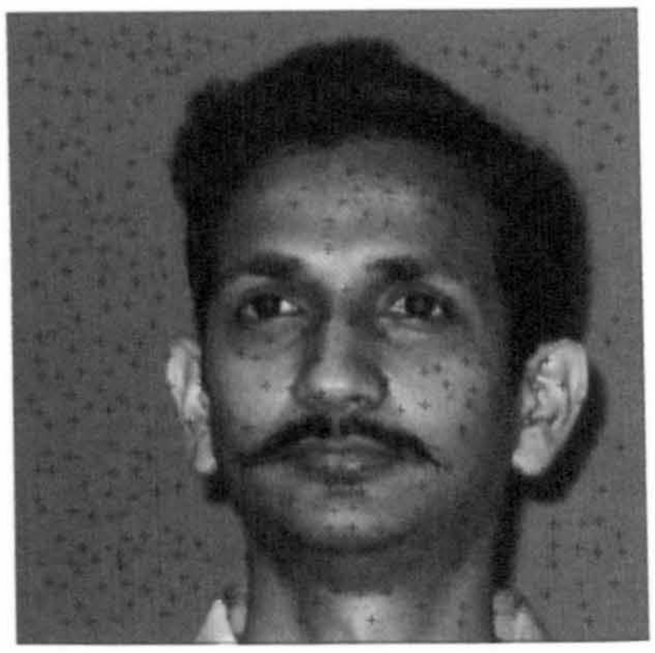
Two parameters need to be chosen for Torr's cross correlation based technique:

- d, the maximum expected disparity between the images.
- N, the size of the correlation window.

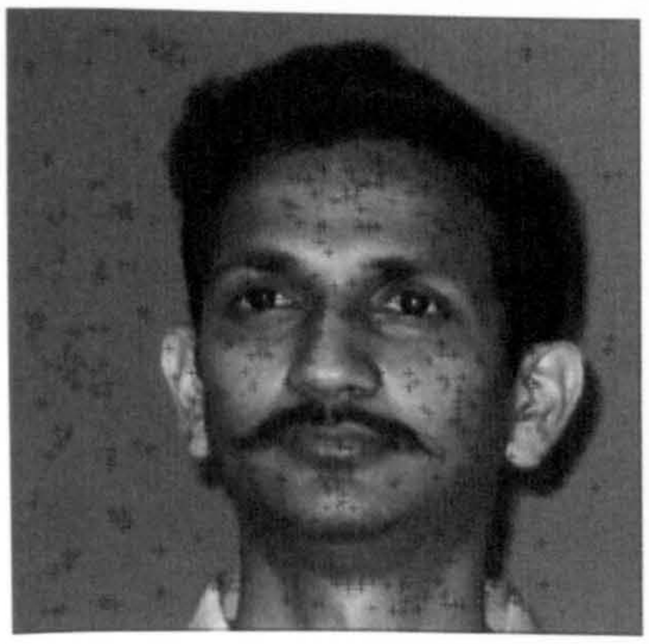
Since ground truth data for these images is not available, the validity of the matches produced by both these images were checked manually. This is an extremely time-consuming



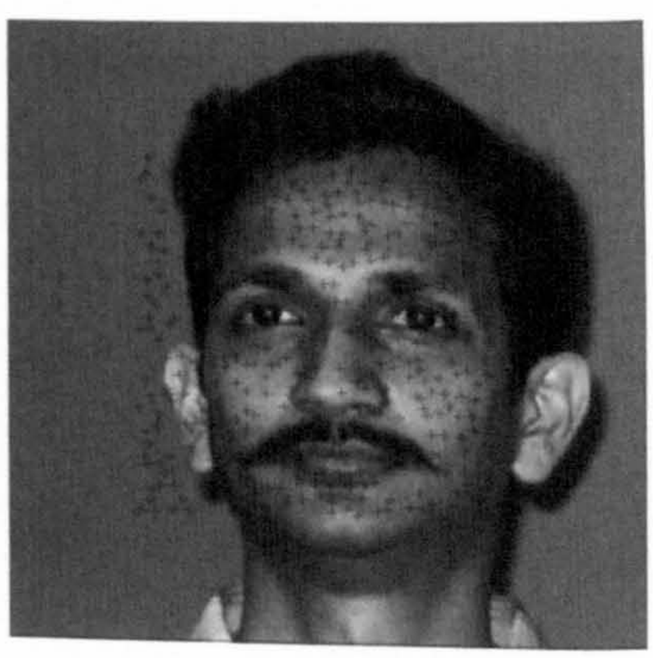
(a) Gaussian convolution mask $s = 5$, standard deviation $\sigma_g = 1$



(b) Gaussian convolution mask $s = 5$, standard deviation $\sigma_g = 2$



(c) Gaussian convolution mask $s = 5$, standard deviation $\sigma_g = 10$



(d) Gaussian convolution mask $s = 50$, standard deviation $\sigma_g = 1$

Figure 5.4: 500 features detected using the Harris corner detector with varying values for the Gaussian convolution mask and the standard deviation.

task and often the issue of whether or not two features match is subjective. As a result, only a fraction of values were investigated for these parameters. The values that resulted in the most number of correct matches are given in Table 5.2.

SVD Parameters		Cross-Correlation Parameters	
σ	10	d	15
W	5	N	7
τ	0.25		

Table 5.2: Parameter values that result in the greatest number of correctly matched features for the two feature matching algorithms.

Once the optimal (or as near optimal as possible) parameters had been determined, the features identified using SUSAN and the Harris detector were matched using each of the two feature matching algorithms, resulting in the four combinations: SUSAN and Pilu's SVD method (SP), SUSAN and Torr's cross-correlation method (ST), Harris' corner detector and Pilu's SVD method (HP) and Harris' corner detector and Torr's cross-correlation method (HT). Figure 5.5 illustrates the average *match rate* (percentage of correctly matched points) for 150 images using a combination of the two feature detection and feature matching algorithms.

It is easy to see that the use of SUSAN results in the fewest correct matches. This comes as no surprise since the quality of the features detected is very poor, compared to the Harris corner detector. In the feature matching algorithms, the SVD method performed poorly compared to the cross-correlation based algorithm. It is worth noting that the performance of the SVD method is poorer when combined with SUSAN than when combined with the Harris corner detector. This is indicative of the importance of the quality of the initial features detected when using the SVD method.

Although the Harris-Torr combination algorithm performs reasonably well on this difficult class of input images, it was felt that the match rate was not high enough for an application such as face recognition. After the elimination of incorrect matches, from the original set of 500 features, on average, only about 190 were correct. This amounts to a mere 0.0029% of the total number of image points (65,536 points for the 256×256 input images used here). Even if it can be assumed that 75% of the image pixels do not contain useful information, there are still an insufficient number of correct matches to justify filling-in the flow field as suggested by Lin and Barron (Lin & Barron 1994).

Consequently, algorithms that are designed specifically to produce dense disparity matches were explored. Two wavelet algorithms were chosen and their description and the results of their application can be found in the subsequent chapters.

5.7 Summary

This chapter discusses the Correspondence Problem in detail. Briefly, it is concerned with searching and the matching of features across two images. It is an extremely difficult problem to solve due to its ill-posed nature. Although, many researchers have over the years tried

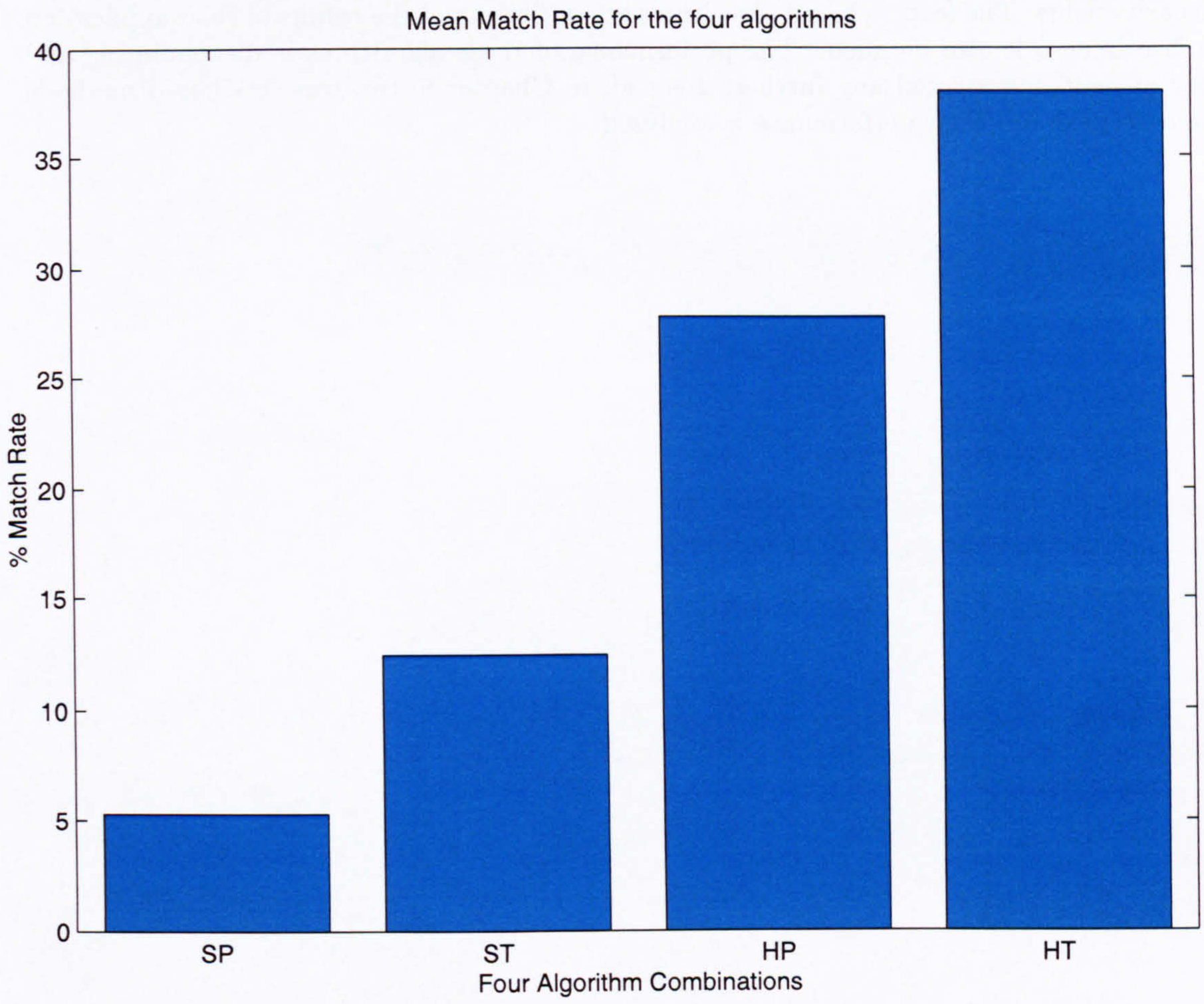


Figure 5.5: Mean percentage of correctly matched features using combinations of feature detection and feature matching algorithms (SP = SUSAN + Pilu, ST = SUSAN + Torr, HP = Harris + Pilu, HT = Harris + Torr).

to address the problem, the solution still remains image-class specific. Face images form a particularly challenging class of images due to the lack of sharp corners and edges; and also because they contain vital information in both the high and the low frequency areas. An overview of some of the existing solutions is provided along with evaluation techniques for stereo matching algorithms.

A set of two feature detectors and two feature matchers are chosen to perform feature based matching on the face images, and two wavelets based methods are chosen to yield dense disparity fields. The feature based algorithms are outlined and the results of their application to face images is also detailed. The performance of these algorithms is disappointing and they are not investigated any further. Instead, in Chapter 6, two wavelets based methods are described and their performance is analysed.

CHAPTER 6

Image Matching: Results

6.1 Introduction

To reconstruct information-rich surfaces with very limited sharp features, dense disparity maps are required. Faces are an example of such surfaces. Subtle features such as cheek bones and the curvature of the forehead which may be useful for automatic face recognition are generally difficult to reconstruct using feature based methods, unless these methods are particularly sophisticated.

In this work two wavelets based methods are investigated for matching face images: Pan's full information image matching algorithm (Pan 1996*b*,*a*) and Magarey's motion estimation algorithm using complex wavelets (Magarey 1997). An overview of both these algorithms is presented in this chapter, with detailed descriptions in Appendices C and D respectively. The two algorithms are applied to a selection of images from the Sheffield Dataset. Qualitative and quantitative results are compared in Sections 6.5.1 and 6.5.2. This is followed by the conclusions and a summary of the chapter.

6.2 Image Matching Algorithms

Wavelets based matching algorithms tend to exploit the inherent facility of multi-resolution in the form of a coarse-to-fine matching strategy. This can be expected to result in more accurate matches, as at each stage in the multi-resolution hierarchy, false matches are eliminated and only the correct ones are propagated to the next level. Pan's algorithm has been tested on aerial photogrammetry images and promising results have been reported. This work investigates its application to face images. Magarey's algorithm has been tested on face images and has yielded very good results when the disparity field is smoothed and regularised at each level of decomposition (Dick 1997). For detailed mathematical exposition, derivations and test results for non-face images, the reader is directed to the original sources.

Both the algorithms have a hierarchical structure, shown in Figure 6.1 (Magarey 1997). The nature of the motion estimator and the interpolator varies for each algorithm. Details are presented in the subsequent sections.

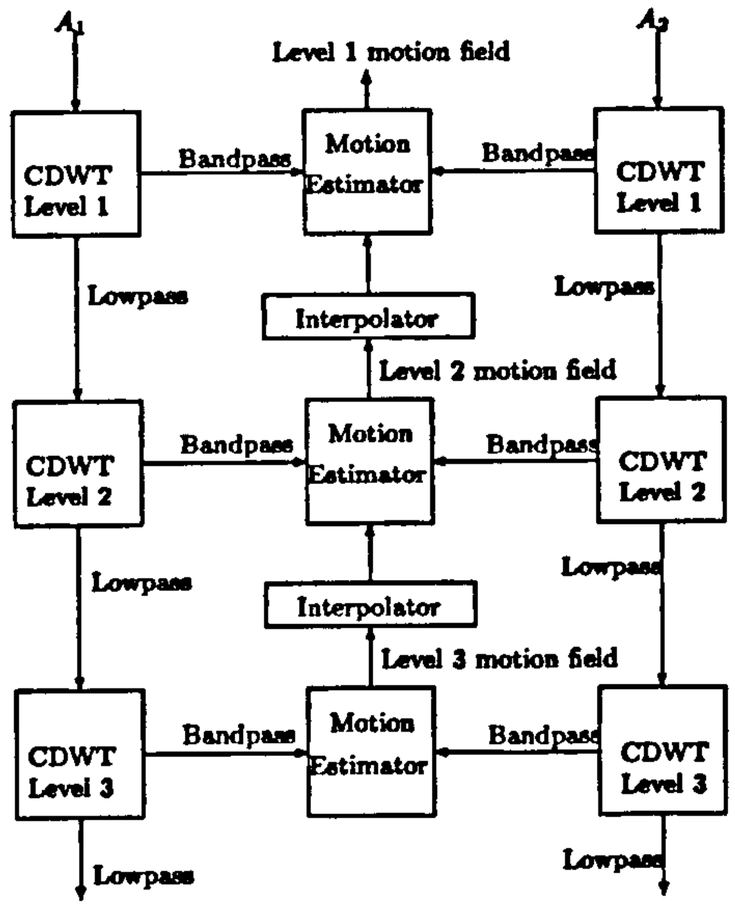


Figure 6.1: Hierarchical structure of CDWT-based motion estimation algorithm. Transforms proceed from top to bottom, estimation from bottom to top. The motion estimation stops at level m_{min}, and the corresponding motion field is interpolated m_{min} times to attain full resolution.

6.3 Pan's complex wavelets

Pan presents a top-down complex wavelets based matching algorithm. Complex wavelets are used for phase-based matching. In addition to the assumptions stated in Section 5.2, this algorithm assumes that the minimum overlap between the two images to be matched is 60% - the industry standard in aerial photogrammetry. The two image planes are not required to be parallel, however, the vergence angle (the angle between the two cameras' lines of "sight") formed by the two image planes is required to be less than $\pi/2$. The scales of the two images may be different, as long as the above conditions are met.

A uniform and full information representation is one in which the constructs are related to the salient information in the original signal $f(x, y)$ (Pan 1996*b*). A mathematical definition of full information representation can be found in Section C.1. Two examples of uniform and full-information representations are Fourier analysis and wavelets analysis. Fourier analysis has the disadvantage that it is extremely poor in spatial localisation. Wavelets analysis on the other hand has not only good localisation in space and frequency domains, but it also has a number of other desirable properties. Details of these and the complex wavelets recommended by Pan for use with his algorithm can be found in Section C.8.

The matching algorithm proceeds with the definition of *implicit feature vectors*, $B_m(x, y)$ (equation C.4). Implicit feature vectors are constructed, for every pixel, from the coefficients of the Approximation and Detail images. These correspond to the outputs of the low-pass and the high-pass channels respectively. Estimation of disparity starts at the highest level of wavelet decomposition m_{max}, where the images are the coarsest (See (f) in Figures 6.3, 6.4 and 6.5). If the assumption of 60% or more overlap is satisfied, then the pixels in the centre

of the reference image (say, left) are guaranteed to have a match in the search image (say, right).

The search for potential matches for the central area starts with an exhaustive search at the coarsest level. The similarity distance defined by equation C.14 yields an approximate disparity matrix at level m_{max} $\mathbf{M}_{m_{max}}(0,0)$ for the central area. The disparity vector for each integer-indexed position of the central area is initialised as

$$\mathbf{M}_m(k,l) \approx \mathbf{M}_m(0,0), \quad \{(k,l) \in \mathcal{N} - (0,0)\} \tag{6.1}$$
$$\mathcal{N} = \{(0,0),(-0.5,-0.5),(-0.5,0.5),(0.5,-0.5),(0.5,0.5)\}$$

and fine-tuned using the *similarity distance measure* S_m (equation C.14). A pixel at (x,y) in the left image corresponds to the pixel (x',y') in the right image that minimises S_m. Search takes place on both integer-indexed positions and half-integer-indexed positions to give a sub-pixel accuracy.

The disparity field is assumed to be continuous in the local neighbourhood. So, based on the correspondences for the pixels in the central area, the remainder of the field at the coarsest level is initialised. Known disparity vectors are propagated outwards, ring by ring, from the central area (spiral propagation). Once the field is initialised, it is fine tuned and refined. Gross errors of the resultant disparity field can be detected and corrected automatically using the local continuity constraint

$$|\mathbf{M}_m(k,l) - \overline{\mathbf{M}}_m(k,l)| \leq T_M,$$

where $\overline{\mathbf{M}}_m(k,l)$ denotes the mean (or median for robustness) of the disparity field vector on the smallest (e.g. 4-connected) neighbourhood centred on the position (k,l) of level j. T_M denotes the maximal allowed disparity difference usually $1 \leq T_M \leq 2$. If the disparity at a pixel value differs significantly ($> T_M$) from the disparity in its neighbourhood, it is adjusted appropriately (see the pseudocode in Figure 6.2).

After image matching on a higher $(m+1)^{\text{th}}$ level, the disparity field is then propagated to the next lower (finer) m^{th} level (hierarchical propagation). The initial disparity field at level m can be obtained by interpolating the disparity field at the $(m+1)^{\text{th}}$ level. The inverse of the similarity distance measure S_m (equation C.14) for each position on the higher level may be taken as the weighting factor for linear or nonlinear interpolation. The matched disparity field $\mathbf{M}_m$ on each m^{th} level yields pairs of matched image points. These, along with the relative imaging geometry (obtained through camera calibration), are used to reconstruct the object surfaces via triangulation.

The algorithm is presented in pseudo-code form in Figure 6.2.

Pan recommends three complex wavelets: Symmetric Complex Daubechies Wavelets of length 6 (SCD-6), Symmetric Complex Daubechies Wavelets of length 4 (SCD-4) and Magarey and Kingsbury's Complex Wavelets of length 4 (MKC-4). These are described in more detail in Section C.8. Figures 6.3, 6.4 and 6.5 show the Approximation and Detail images for wavelet decomposition levels 1-6 using SCD-6, SCD-4 and MKC-4 (Note that images from only the complex channel are shown for MKC-4 since the images from the complex conjugate channel are mirror images of these.). At higher levels of decomposition, the features are gross, and it is easier to conduct exhaustive searches for potential matches. As the images get finer, it is harder to search exhaustively and to be certain that the correct match has

Algorithm 1: Image Matching using Pan's Algorithm

Input: Images X_1, X_2 ($N \times N$), complex wavelet (SCD-4, SCD-6, MKC-4) and levels of decomposition m_{max}
Output: Disparity field M_m ($N \times N$)

Perform Complex Discrete Wavelet Transform (CDWT) on X_1 and X_2 using complex valued low-pass and high-pass filters (see Section C.8)

Output: Approximation & Detail images of size $N/2^m$ for each image X_1 and X_2 at levels $m = 1 : m_{max}$. Single channel filter-banks (SCD-4, SCD-6) produce 1 Approximation & 3 Detail images. Two channel filter-banks produce twice as many.

for $m = m_{max} : 1$ // *m_{max} is the coarsest level of decomposition*
// *Compute disparity field at each level using:*
if $m == m_{max}$
Exhaustively search for matches in central area
- Find co-ordinates of central pixel
- Compute implicit feature vector (IFV) for this point & its neighbours in left & right images
- Compute similarity distances S_m
- Establish match for central pixel based on $min(S_m)$

Repeat above process for pixels neighbouring central pixel
Spiral Propagation
Set the disparity values of pixels surrounding the central area the same as their adjacent pixels
end if

if $m \neq m_{max}$
// *If this is not the coarsest level then an initial disparity field is already available.*
// *Refine this field using:*
Spiral Search
for every pixel, its potential match & the neighbours of this match
- Compute IFV's
- Compute similarity distances S_m
- Of the potential match & its neighbours, determine the correct match using $min(S_m)$

end for
end if

Refine disparity field (starting from centre & going out) by:
for every pixel location (k,l) in central area
Compare disparity value $\mathrm{M}_m(k,l)$ at (k,l) with the mean disparity value $\overline{\mathrm{M}}_m(k,l)$ in its 3×3 neighbourhood
if error value $e = |\mathrm{M}_m(k,l) - \overline{\mathrm{M}}_m(k,l)| \leq T_M$
// *T_M: threshold value for the maximum allowable difference in the disparity*
// *values of neighbouring pixels*
Do nothing
else
Correct disparity value
using error-weighted sum of computed disparity $\mathrm{M}_m(k,l)$ at (k,l) & mean disparity $\overline{\mathrm{M}}_m(k,l)$ of the neighbouring pixels
$(e \times \overline{\mathrm{M}}_m(k,l)) + (1 - e) \times \mathrm{M}_m(k,l))$
end if

Hierarchical Propagation using linear interpolation
Interpolate the disparity field to obtain the correct resolution at the next finer level $m - 1$.
The interpolated field forms the initial estimate of disparity values at level $m - 1$.
end for

Figure 6.2: Pseudo code for Pan's complex wavelets based image matching algorithm

been isolated. As a result, the hierarchical structure of wavelets based algorithms provides a means of narrowing down the search area for potential matches in larger images. The horizontal and vertical disparity maps corresponding to each of these sets of decomposition images are presented in Figures 6.7, 6.8 and 6.9.

6.4 Magarey's complex wavelets

Magarey's complex wavelets based motion estimation algorithm (Magarey & Kingsbury 1995, 1996, Magarey 1997, Magarey & Kingsbury 1998*b*, Castellano 1999) is summarised in this section. Further mathematical details are provided in Appendix D.

A pair of images X_1 and X_2, belonging to the reference and the current frames respectively, are decomposed using the Complex-valued Discrete Wavelet Transform (CDWT). A pair of complex conjugate wavelet filters in a two channel filter-bank (one channel for the complex filters, one for the complex conjugate filters) are used to implement the CDWT. The 1D filter pair $\{h_0, h_1\}$ are rational-valued complex kernels (equations 6.2 and 6.3) and were designed by Magarey and Kingsbury (Magarey & Kingsbury 1995, 1996, Kingsbury & Magarey 1997, Magarey 1997, Magarey & Kingsbury 1998*b*). They are described in more detail in Appendix B.

$$h_0 = \Big[1-j \quad 4-j \quad 4+j \quad 1+j\Big]/10 \tag{6.2}$$

$$h_1 = \Big[-1-2j \quad 5+2j \quad -5+2j \quad 1-2j\Big]/14 \tag{6.3}$$

The wavelet functions generated using these filters were also used in Pan's matching algorithm and are referred to as MKC-4.

Image matching starts at the coarsest level m_{max}. Since MKC-4 wavelets have 2 channels, at each level of decomposition, there are twice as many coefficients (2 Approximation and 6 Detail images). The 6 Detail images from each image frame (X_1 and X_2) are used to compute the disparity value at each sub-pixel and a confidence measure for these values.

The matching criterion used in this algorithm takes the form of a *sub-band squared difference* (SSD) surface at pixel location $\mathbf{n}$ over the real valued offset vector $\mathbf{f}$ (see Section D.2 for further details on this (equation D.1)):

$$SD^{(n,m)}(\mathbf{n},\mathbf{f}) = \left|D_1^{(n,m)}(\mathbf{n},\mathbf{f}) - D_2^{(n,m)}(\mathbf{n})\right|^2 \tag{6.4}$$

$D_1^{(n,m)}(\mathbf{n},\mathbf{f})$ is an interpolated sub-band (Detail) coefficient from image frame 1 at offset $\mathbf{f}$ from $\mathbf{n}$, and allows for disparities at sub-pixel locations to be estimated. It is assumed that the input spectrum has no sharp peaks in the support region of the associated Gabor-like wavelet filter (Magarey & Kingsbury 1996). This means that the phase-behaviour of the coefficient follows that of the filter and it has a constant magnitude, given by equation D.9. As a result, the matching criterion $SD^{(n,m)}(\mathbf{n},\mathbf{f})$ can be approximated as an elliptical surface which is quadratic around its minimum.

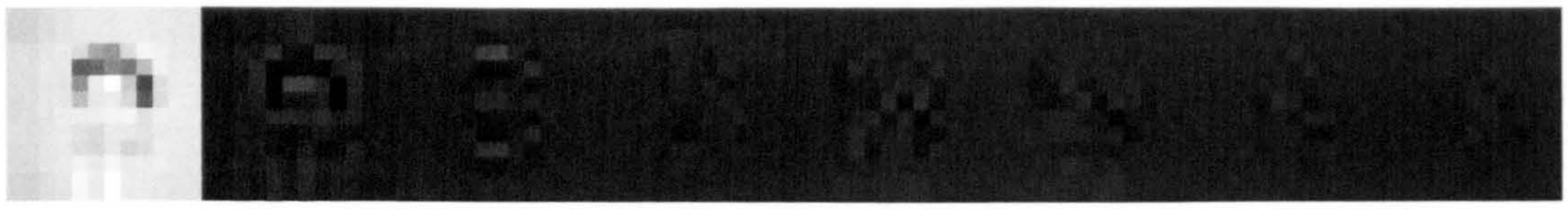
$SD^{(n,m)}(\mathbf{n},\mathbf{f})$ is computed for each pixel across all the six surfaces and summed to give a single quantity (equation D.4) $SD^{(m)}$ for each level of decomposition. This averages the disparity over the six oriented sub-bands rather than over a region of pixels in any particular sub-band. If $SD^{(m)}$ is plotted, it also produces an elliptical bowl shaped surface, which can

(a) Real and Imaginary components of the four subimages at level 1 of decomposition using SCD-6

(b) Real and Imaginary components of the four subimages at level 2 of decomposition using SCD-6

(c) Real and Imaginary components of the four subimages at level 3 of decomposition using SCD-6

(d) Real and Imaginary components of the four subimages at level 4 of decomposition using SCD-6



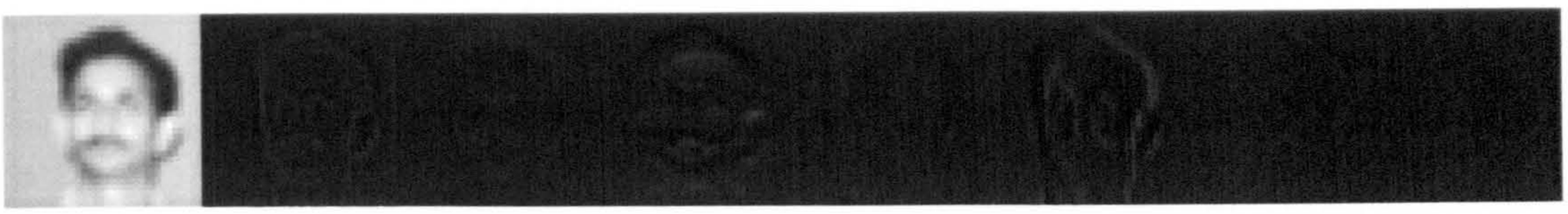
(e) Real and Imaginary components of the four subimages at level 5 of decomposition using SCD-6

(f) Real and Imaginary components of the four subimages at level 6 of decomposition using SCD-6

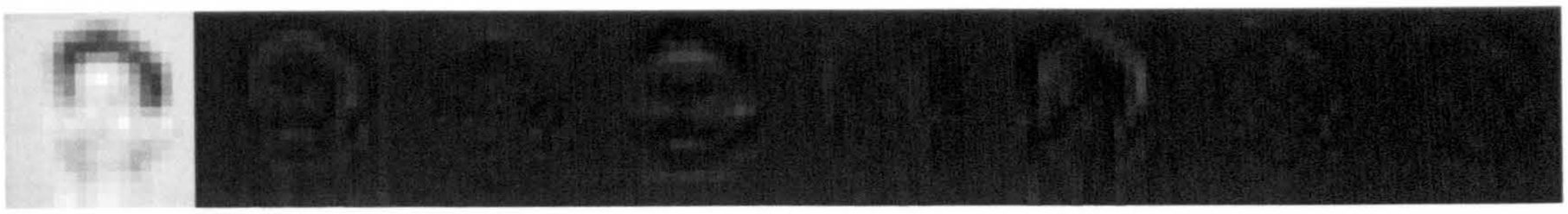
Figure 6.3: Real and imaginary components of the Approximation and Detail images (horizontal, vertical and diagonal) at various levels of decomposition using Symmetric Complex Daubechies wavelet of length 6 (SCD-6).

(a) Real and Imaginary components of the four subimages at level 1 of decomposition using SCD-4

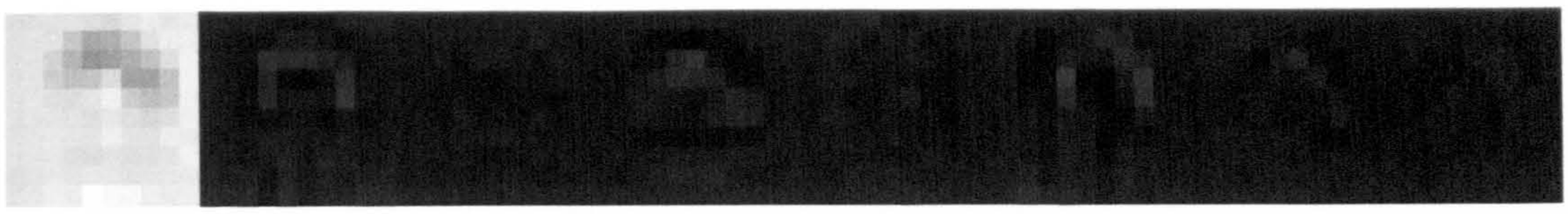
(b) Real and Imaginary components of the four subimages at level 2 of decomposition using SCD-4



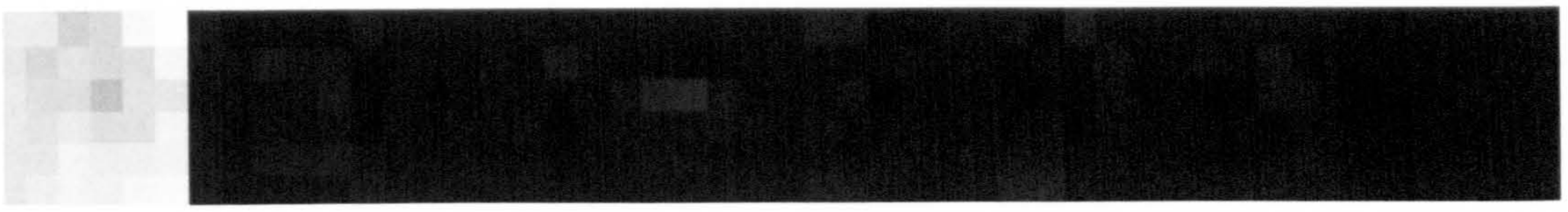
(c) Real and Imaginary components of the four subimages at level 3 of decomposition using SCD-4



(d) Real and Imaginary components of the four subimages at level 4 of decomposition using SCD-4



(e) Real and Imaginary components of the four subimages at level 5 of decomposition using SCD-4



(f) Real and Imaginary components of the four subimages at level 6 of decomposition using SCD-4

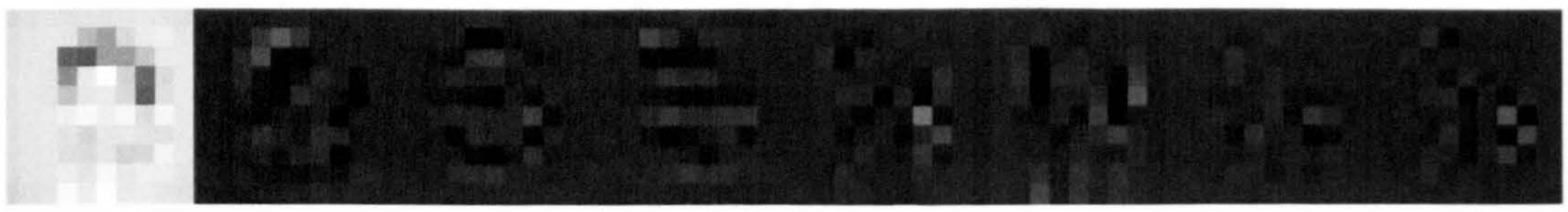
Figure 6.4: Real and imaginary components of the Approximation and Detail images (horizontal, vertical and diagonal) at various levels of decomposition using Symmetric Complex Daubechies wavelet of length 4 (SCD-4).

(a) Real and Imaginary components of the four subimages at level 1 of decomposition using MKC-4

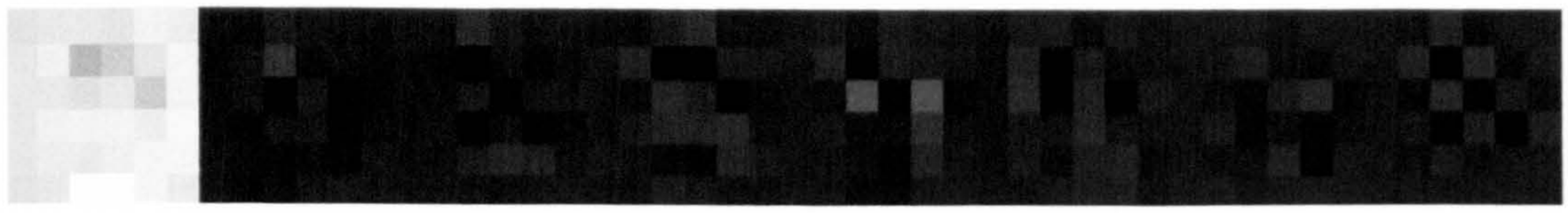
(b) Real and Imaginary components of the four subimages at level 2 of decomposition using MKC-4

(c) Real and Imaginary components of the four subimages at level 3 of decomposition using MKC-4

(d) Real and Imaginary components of the four subimages at level 4 of decomposition using MKC-4



(e) Real and Imaginary components of the four subimages at level 5 of decomposition using MKC-4



(f) Real and Imaginary components of the four subimages at level 6 of decomposition using MKC-4

Figure 6.5: Real and imaginary components of the Approximation and Detail images (horizontal, vertical and diagonal) at various levels of decomposition using Magarey and Kingsbury's complex wavelets of length 4 (MKC-4).

be fully described from the 12 coefficients $D_1^{(n,m)}$ and $D_2^{(n,m)}$ (Dick 1997). This in turn allows $SD^{(m)}$ to be approximated as a quadratic surface (equation D.21):

$$SD^{(m)}(\mathbf{n},\mathbf{f}) \approx Af_1^2 + BF_2^2 + Cf_1f_2 + Df_1 + Ef_2 + G$$

The coefficients $\{A, B, C, D, E, F, G\}$ are defined in terms of the coefficients of the band-pass images $D^{(n,m)}$ and the centre frequencies $\mathbf{\Omega}^{(n,m)}$. Closed form expressions for these coefficients are given in equations D.22-D.28.

By completing the square, $SD^{(m)}$ can be transformed from the representation in (equation D.21) to (equation D.29):

$$SD^{(m)}(\mathbf{n},\mathbf{f}) \approx \alpha(f_1 - f_{10})^2 + \beta(f_2 - f_{20})^2 + \gamma(f_1 - f_{10})(f_2 - f_{20}) + \delta$$

where $\mathbf{f}_0 = [f_{10} \quad f_{20}]^\top$ are the co-ordinates of the surface minimum. $\mathbf{f}_0$ is taken to be the level m disparity value at sub-pixel $\mathbf{n}$ of frame 2. The curvature parameters $\alpha, \beta, \gamma = A, B, C$ together define the curvature matrix of the surface at its minimum point (equation D.31) and give a measure of confidence in the estimated disparity value. These are referred to as "ellipses of confidence" in (Magarey & Kingsbury 1996). The steeper the surface (large curvature parameters), the more precise the disparity value and the higher the confidence in the disparity estimates.

Once the disparity field has been computed at the coarsest level it is refined and propagated to the next finer level. The SSD is refined (see Section D.5.5) using the information from the ellipses of confidence. To account for the images at the next finer level being twice the size, the SSD surface is scaled and interpolated using the bilinear kernel (see (Magarey 1997) for details). The interpolated field of level m surfaces is denoted by $SD'^{(m)}(\mathbf{n},\mathbf{f})$, with parameters $\{A', B', C', D', E', G'\}$ or $\{\mathbf{f}_0', \alpha', \beta', \gamma', \delta'\}$.

$\mathbf{f}_0'$ acts as an initial estimate of disparity at the next finer level $m-1$. The disparity estimates, or the SSD's at level $m-1$, $SD^{(m-1)}$ are established using the procedure described above. The interpolated field $SD'^{(m)}$ is then added to $SD^{(m-1)}$ to form the *cumulative squared difference* (CSD). This incorporates the information from the previous level into the disparity estimates at this level. The CSD is then interpolated and used as the initial estimate for the next finer level. This procedure is repeated until the required level of detail is obtained.

The algorithm is presented in pseudo-code form in Figure 6.6.

One of the main advantages of phase based technique such as this one is that it provides a certain robustness to local changes of illuminance. This is very beneficial for this particular application as it adds a degree of versatility to the system and retains some of the benefits of face recognition.

However, the algorithm also has some drawbacks and it may not be applicable in all situations. In particular, the model breaks down when the motion is "out of range" for a given search set, i.e. if one or both components of motion exceed 0.5×2^m pixels. Also, if the images contain motion discontinuities of some kind (eg. objects moving over one another or over a stationary, textured background) or if there is significant rotational, dilational or shear component, then the algorithm would result in noisy or inaccurate matches.

Algorithm 2: Image Matching using Magarey's Algorithm

Input: Images X_1, X_2 $(N \times N)$ and levels of decomposition m_{max}
Output: Disparity field SD $(N \times N)$

Perform Complex Discrete Wavelet Transform (CDWT) on X_1 and X_2 using complex valued low-pass and high-pass filters:
$h_0 = [1-j, 4-j, 4+j, 1+j]/10 \quad h_1 = [1-2j, 5+2j, -5+2j, 1-2j]/14$

Output: Six Detail images $D^{(n,m)}$ $\{n = 1, \ldots, 6\}$ and two Approximation images $A^{(1,m)}$ and $A^{(2,m)}$ of size $N/2^m$ for each image X_1 and X_2 at levels $m = 1 : m_{max}$.

for $m = m_{max} : 1$ *// m_{max} is the coarsest level of decomposition*
// Compute the sub-band squared difference (SSD) for each of the 6 sub-bands at each level using:

for sub-bands 1:6
// Compute disparity between the pixels in each of the sub-bands of X_1 and the corresponding sub-bands of X_2.
for every pixel location $\mathbf{n}_1$ in X_1
// Find a match in the image X_2 using:

- **Sub-band squared difference (SSD):** $SD^{(n,m)}(\mathbf{n}, \mathbf{f}) = \left| D_1^{(n,m)}(\mathbf{n}, \mathbf{f}) - D_2^{(n,m)}(\mathbf{n}) \right|^2$
- **Compute the energy weighting** $\sigma P^{(n,m)}$ and divide the SSD (equation D.1)
- **Sum over the 6 sub-bands:** $SD^{(m)}(\mathbf{n}, \mathbf{f}) = \sum_{n=1}^{6} SD^{(n,m)}(\mathbf{n}, \mathbf{f})$ (equation D.4)

end for

- **Compute the coefficients of the quadratic surface** $\{A, B, C, D, E, G\}$ and $\{\mathrm{f}_0, \alpha, \beta, \gamma, \delta\}$ *// See Section D.4*
- **Locate surface minimum:**
  The correspondent for the left pixel $\mathbf{n}_1$ is the indicated by the minimum $\mathbf{f}_0$ of the quadratic surface
- **Curvature Correction:**
  Subtract circular bowl shaped surface with the same minimum as the elliptical surface (equation D.41) to increase accuracy of the estimates (equations D.41 & D.42).
- **Confidence Filtering:**
  - Compute scalar confidence measure C^m (equation D.38).
  - Eliminate matches for which $C^m < 0.95$

  // Low confidence matches eliminated. Motion estimates at subsequent levels not affected
- **Smooth & Regularise:**
  - Compute principal axes & curvatures using the curvature matrix (equation D.31)
  - Compute directional confidence measures using equations D.43 & D.44

  if $c_{max}c_{min} < t$ *// overall poor reliability of match*
  Eliminate match
  end if
  - Solve equation D.45 using Gauss-Seidel iterations

end for

if $m \neq m_{max}$ *// If this is not the coarsest level*

- **Combine with previous level estimates** (Section D.5.2)
  $CSD^{(m)}(\mathbf{n}, \mathbf{f}) = CSD'^{(m+1)}(\mathbf{n}, \mathbf{f}) + SD^{(m)}(\mathbf{n}, \mathbf{f}) \quad m_{min} \leq m < m_{max}$

end if

// Propagate motion estimates up the image pyramid. Scale & interpolate disparity field to account for
// increased no. of pixels & decreased spacing between adjacent pixels.

- **Interpolate using bilinear kernel:**
  - Upsample & then apply [1 3 3 1]/4 first to columns, then to rows.
- **Scale disparity field:** See equations D.32-D.36

end for

Figure 6.6: Pseudo code for Magarey's complex wavelets based image matching algorithm

6.5 Image Matching Results

Although many image matching algorithms report excellent results on synthetic images such as the "Diverging/Translating Tree"[1] and the "Yosemite Sequence" [2], they are not always suitable for matching face images. This is mainly due to lack of sharp features such as edges and corners in face images. Since face recognition and even face reconstruction are rarely performed using stereo, very few image matching algorithms are tested on face images. Furthermore, ground truth data for face images is extremely hard to establish. This makes the task of choosing appropriate algorithms for this application particularly difficult.

Magarey's complex wavelets based image matching algorithm has been tested on face images with excellent results (Magarey & Dick 1998). Pan's algorithm has been tested on aerial photogrammetry and terrain images (Pan 1996*a*), and promising results have been reported. Initially, only qualitative comparisons were made, as is common in most image matching literature. In due course, quantitative assessment of the algorithms and wavelets used was carried out to support the findings of the qualitative assessment.

6.5.1 Qualitative Analysis

Figures 6.7, 6.8 and 6.9 show the horizontal and the vertical disparity maps generated for the decomposition images shown in the previous section. Pan's algorithm has been used with the three complex wavelets SCD-6, SCD-4 and MKC-4. Compare these with the disparity maps using MKC-4 wavelet with Magarey's algorithm in Figure 6.10. The coarsest level maps from Pan's algorithm show how the disparity values generated for the central area are propagated outwards as an initial estimate. This leads to the entire central region of the image having the same disparity values, indicated by a block of the same colour. By contrast, the coarsest level map from Magarey's algorithm (Figure 6.10, (f)) computes the disparity values for each pixel individually and does not rely on the initial estimates for the central region being correct.

The effect of the assumption of disparity values in a neighbourhood being similar also ripple through to the finer level maps. Pan's algorithm uses this assumption at each level of decomposition to refine the matches. If this assumption is not satisfied, the disparity values are "corrected" to satisfy it. This manifests in the form of "pockets" of erroneous disparity estimates with the entire neighbourhood having incorrect values. The maps resulting from Magarey's algorithm are much smoother (noise-free) at finer levels. They capture the subtle variations in the disparity values as the algorithm is more sophisticated and the refinement procedure adopted is more robust to errors in the matching process.

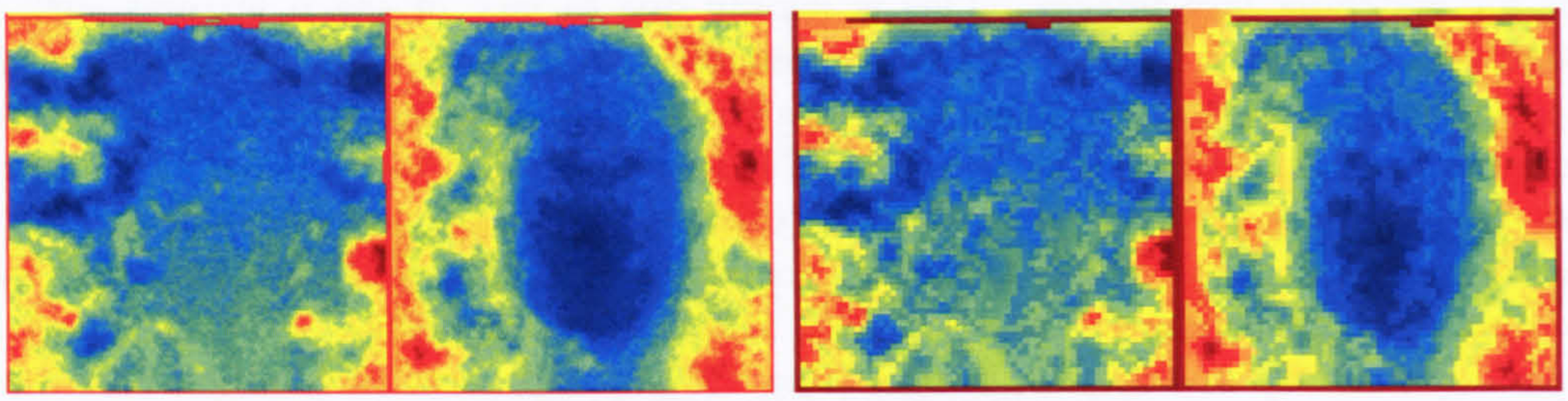
The disparity maps give some indication about the quality of the two algorithms. But since ground truth data is not available, it is not possible to verify the quality of the algorithms with certainty. In such situations, it is common for the matching algorithms to be compared on the basis of their 3D reconstructions. It is in general easier to identify errors in the reconstruction than it is in the disparity maps since the errors and the noise manifest as distortions in the surface.

In this work, the two algorithms were initially compared qualitatively only, using the 3D reconstructions. The reconstructions clearly showed that Magarey's algorithm was by far the

---

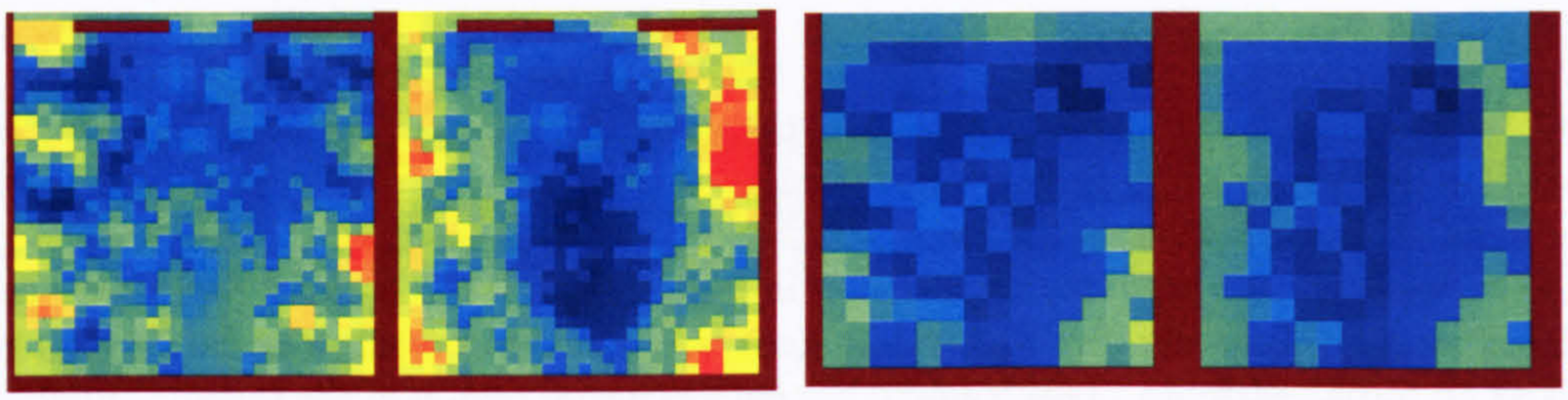
[1]Diverging/Translating Tree sequence was created by David Fleet and

[2]Yosemite sequence was created by Lynn Quam. Both these test sequences are available at ftp://ftp.csd.uwo.ca/pub/vision



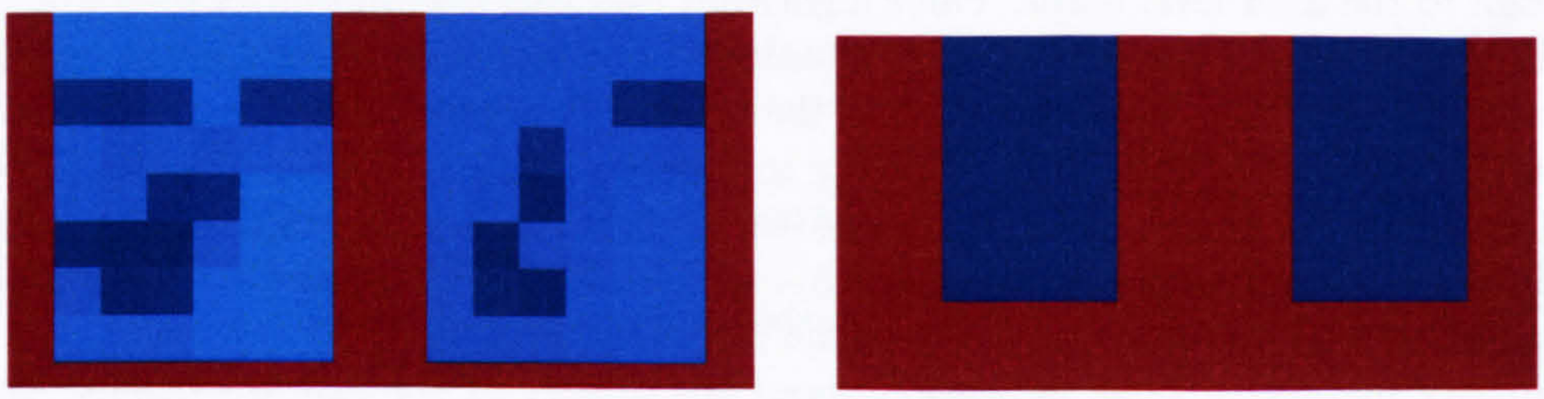
(a) Vertical and horizontal disparity maps at level 1 of decomposition

(b) Vertical and horizontal disparity maps at level 2 of decomposition



(c) Vertical and horizontal disparity maps at level 3 of decomposition

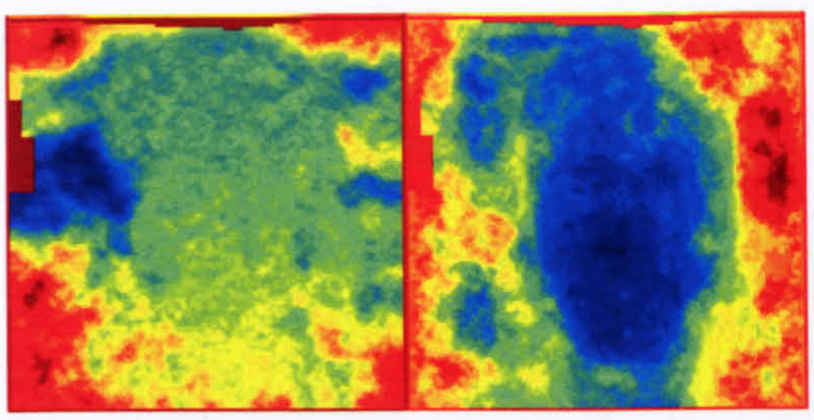
(d) Vertical and horizontal disparity maps at level 4 of decomposition



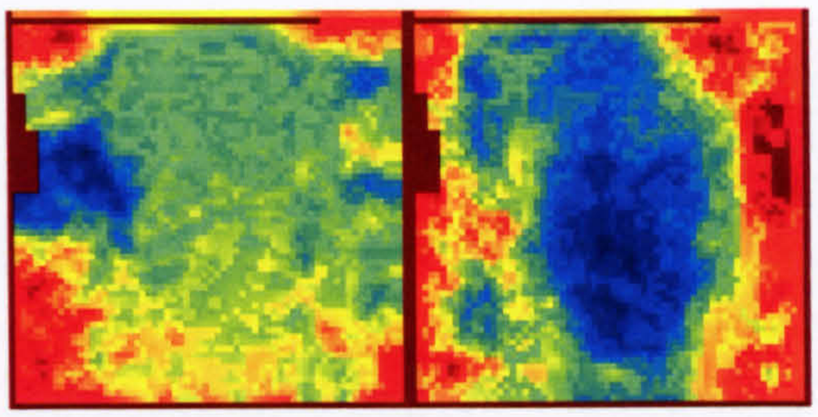
(e) Vertical and horizontal disparity maps at level 5 of decomposition

(f) Vertical and horizontal disparity maps at level 6 of decomposition

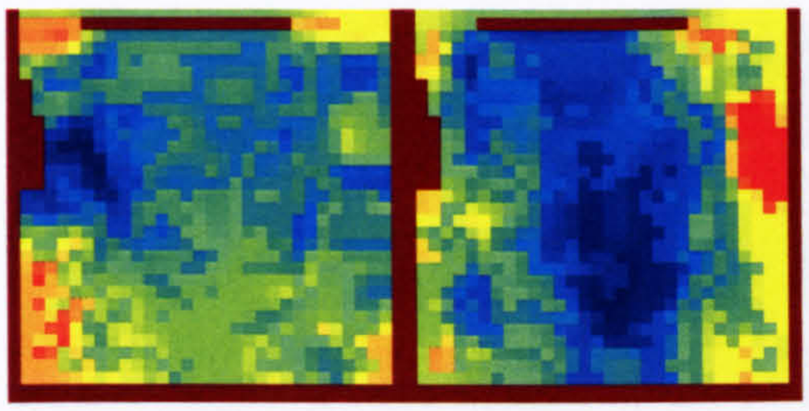
Figure 6.7: Vertical and horizontal disparity maps obtained at each of the six levels of decomposition using Pan's algorithm with Symmetric Complex Daubechies wavelet of length 6 (SCD-6).



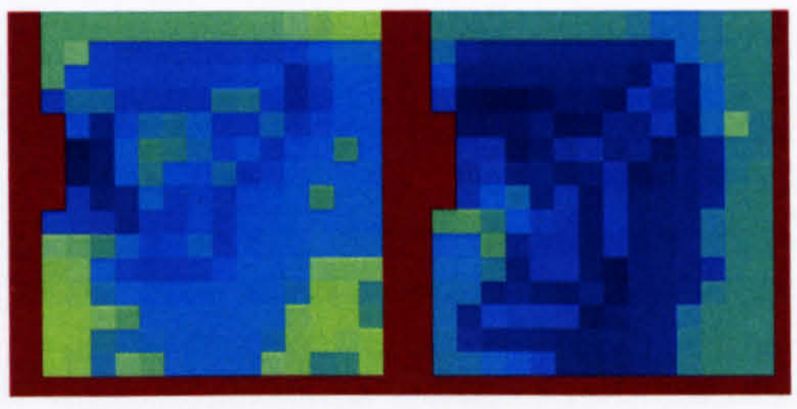
(a) Vertical and horizontal disparity maps at level 1 of decomposition



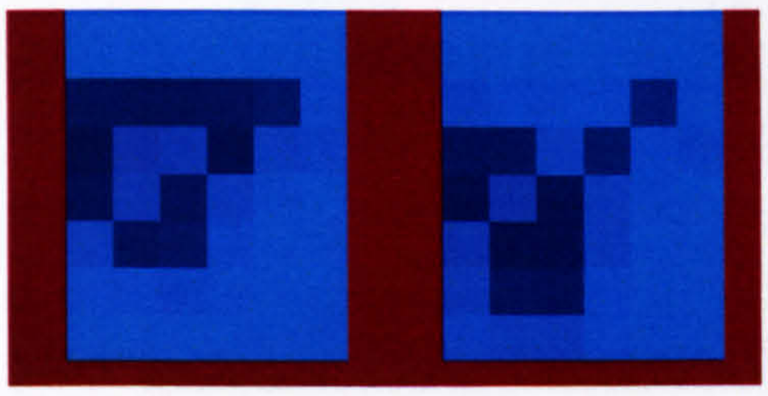
(b) Vertical and horizontal disparity maps at level 2 of decomposition



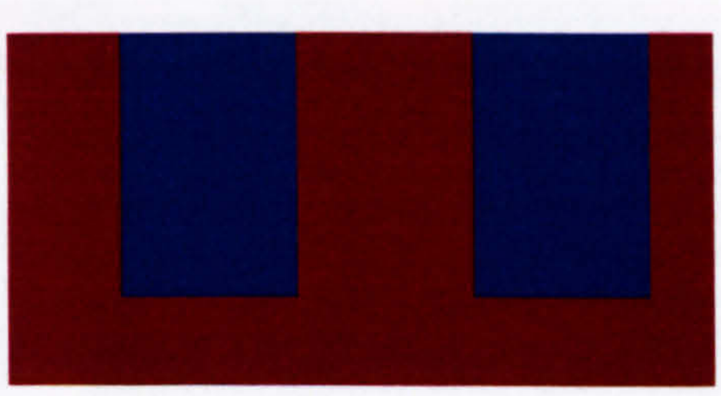
(c) Vertical and horizontal disparity maps at level 3 of decomposition



(d) Vertical and horizontal disparity maps at level 4 of decomposition

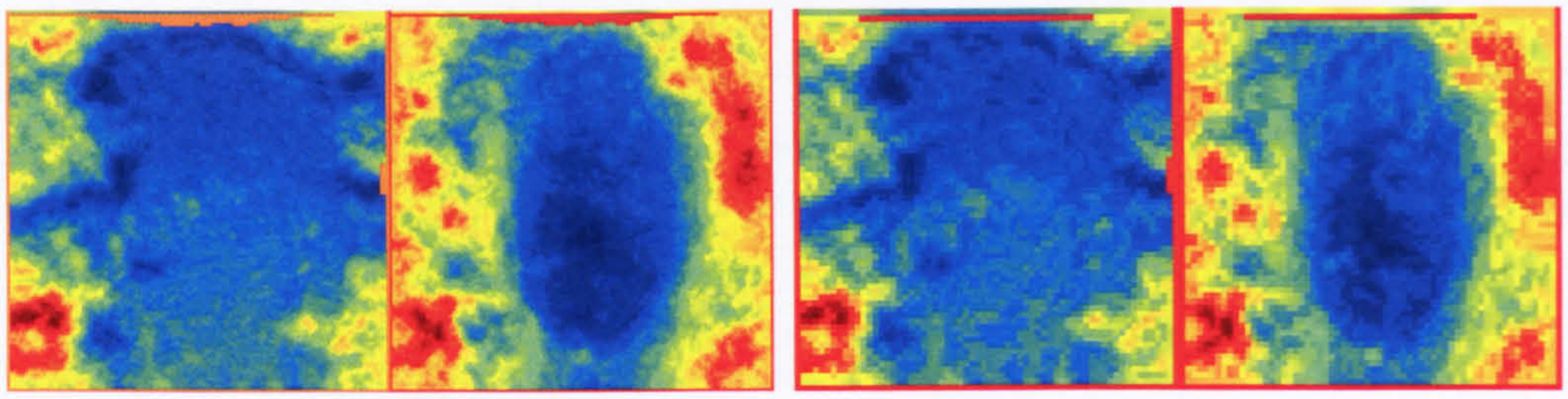


(e) Vertical and horizontal disparity maps at level 5 of decomposition



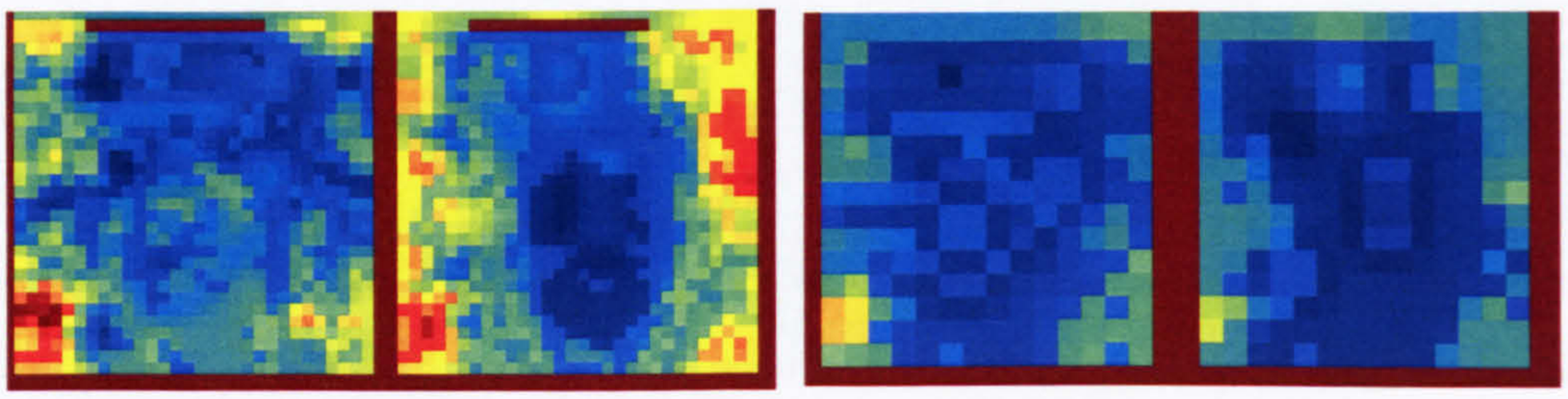
(f) Vertical and horizontal disparity maps at level 6 of decomposition

Figure 6.8: Vertical and horizontal disparity maps obtained at each of the six levels of decomposition using Pan's algorithm with Symmetric Complex Daubechies wavelet of length 4 (SCD-4).



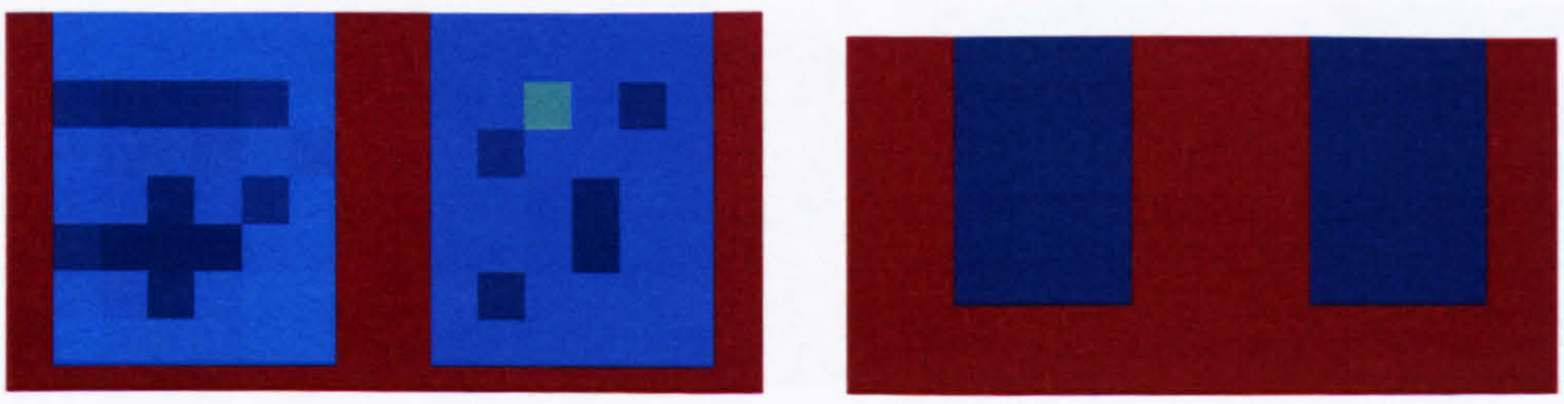
(a) Vertical and horizontal disparity maps at level 1 of decomposition

(b) Vertical and horizontal disparity maps at level 2 of decomposition



(c) Vertical and horizontal disparity maps at level 3 of decomposition

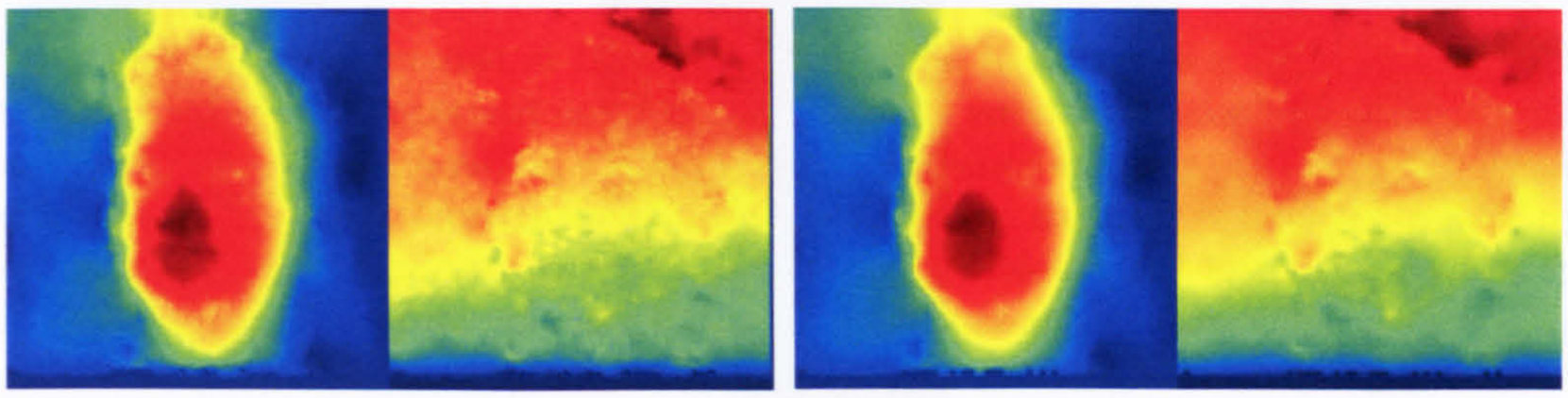
(d) Vertical and horizontal disparity maps at level 4 of decomposition



(e) Vertical and horizontal disparity maps at level 5 of decomposition

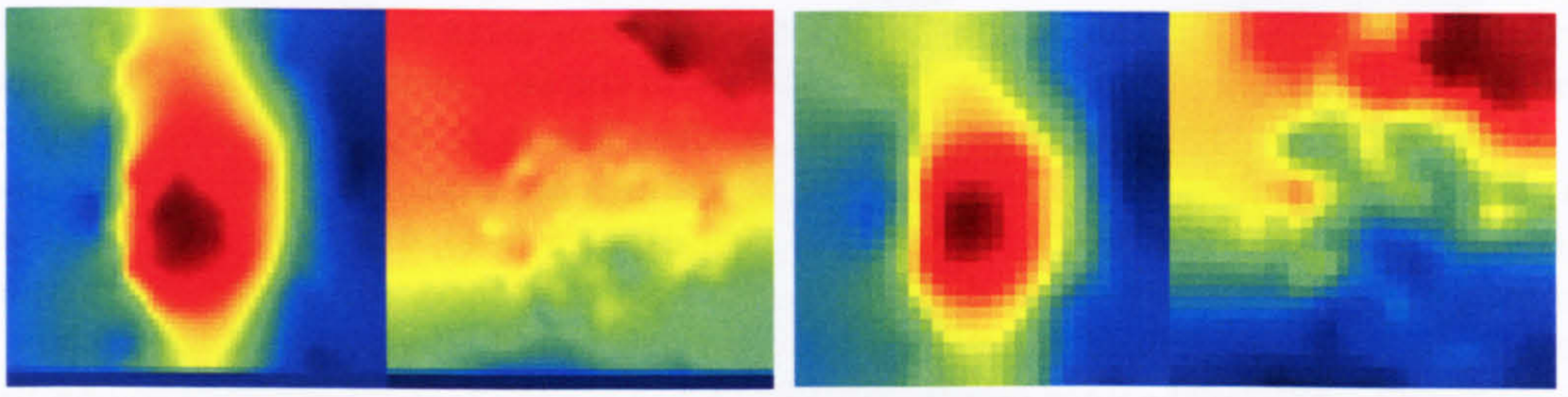
(f) Vertical and horizontal disparity maps at level 6 of decomposition

Figure 6.9: Vertical and horizontal disparity maps obtained at each of the six levels of decomposition using Pan's algorithm with Magarey and Kingsbury's complex wavelets of length 4 (MKC-4).



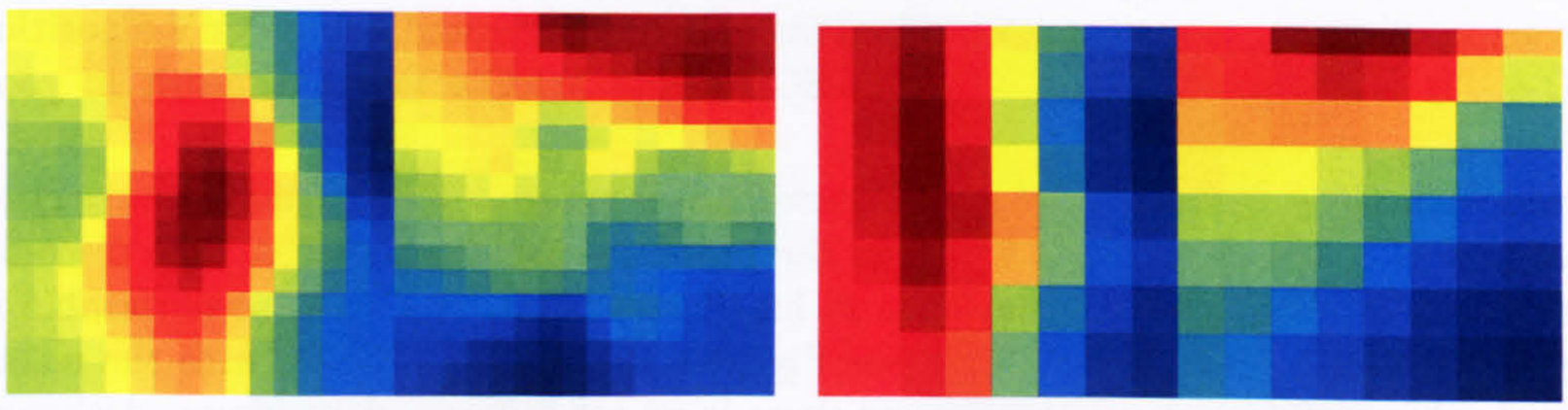
(a) Horizontal and vertical disparity maps at level 1 of decomposition

(b) Horizontal and vertical disparity maps at level 2 of decomposition



(c) Horizontal and vertical disparity maps at level 3 of decomposition

(d) Horizontal and vertical disparity maps at level 4 of decomposition



(e) Horizontal and vertical disparity maps at level 5 of decomposition

(f) Horizontal and vertical disparity maps at level 6 of decomposition

Figure 6.10: Horizontal and Vertical disparity maps obtained at each of the six levels of decomposition using Magarey & Kingsbury's complex wavelets and motion estimation algorithm.

more superior. However, it was in general difficult to draw any conclusions about the how the three wavelets used with Pan's algorithm compared.

Comparison using 3D Reconstructions

The reconstructions for a selection of images are presented in this section. Their quality is poor compared to the reconstructions generally seen in the literature. They are produced from images captured using simple cameras, while those seen in literature are usually obtained through specialist equipment such as laser scanners or structured light projectors. This highlights the difficulty of the problem but does not hinder the evaluation of the algorithms since all the reconstructions use the same camera parameters and are equally affected by the noise.

The reconstructions presented in this section are generated using the full frontal images shown in Figure 4.2. The 3D reconstructions are then rotated about the y-axis to show different perspectives.

It is obvious looking at the reconstructions that Magarey's algorithm is better at identifying correct matches between the images. The reconstructions using Pan's algorithm preserve only the gross relief information such as the concavity of the face relative to the background. Depth of smaller features such as nose and chin is often lost (e.g. Figure 6.13 (a-c), Figure 6.15 (a) and (e-f), 6.16 (f-g) and 6.18 (a)). If these features are captured, then they are usually extremely noisy (e.g. 6.11 (c-f), 6.12 (a-f), 6.13 (e-f), 6.16 (a-e), 6.18 (b-f)). Without the texture maps, it would be extremely difficult to establish that these surface reconstructions represent faces. Identification of an individual from these would be virtually impossible.

A further indication of the poor quality of the matches are the texture maps corresponding to each of the reconstructions. Bilinear interpolation is used to compute the intensity values at each 3D position, using the intensity values at the matching points in the two images. If the matches are incorrect, the corresponding intensity values will also be incorrect. This effect is most visible when using SCD-4 wavelet (see Figures 6.12 (a-f), 6.15 (a-b), 6.18 (a)). The most erroneous texture maps are obtained for matches generated using Pan's algorithm with SCD-4. It is harder to gauge which of SCD-6 and MKC-4 wavelets result in good matches.

The reconstruction of subject 13's face (Figure 6.15) is extremely poor and is a very good example of how the algorithm breaks down in the presence of discontinuities. The discontinuity in this case is in the form of reflection in his glasses. This effect is also noticeable in other images with reflection in the glasses (see Figure 6.19). Compare these with the reconstructions of subject 11g (Figure 6.14), where there is no reflection in the glasses. Although the reconstruction is unsatisfactory, it does not have any distortions due to reflection.

Figure 6.20 shows the reconstructions produced by applying Magarey's algorithm to some of the non-frontal images from Figure 4.3. Images with extreme rotation (a,b,g) result in poor models since only the visible portions of a scene can be reconstructed accurately. This is further exemplified by image (f) which is reconstructed using a profile image. The effects of harsh lighting are also evident. The part of the face that is well lit is well-reconstructed (k), while there is considerable loss of depth information in the poorly lit regions (l). This is particularly noticeable in the region near the eyes, which in this case is completely flat. The subtlety of the information captured by Magarey's algorithm is highlighted in surfaces with varying expressions. Minor deformations in the region near the eyes in (h) and (i) are

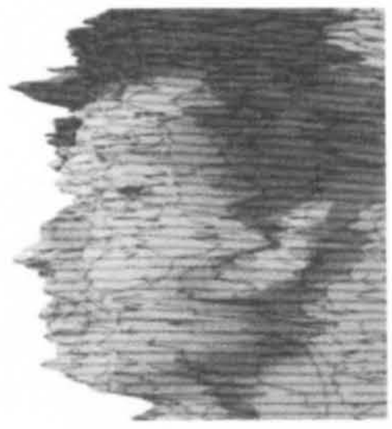
(a) Pan & SCD-4 (b) Pan & SCD-4 (c) Pan & SCD-6

(d) Pan & SCD-6 (e) Pan & MKC-4 (f) Pan & MKC-4

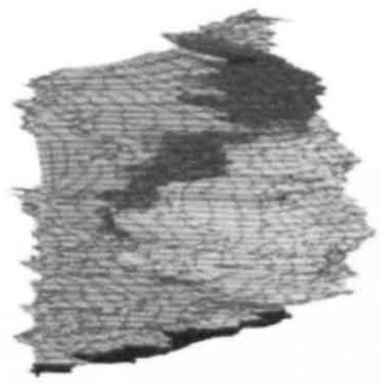
(g) Magarey & MKC-4 (h) Magarey & MKC-4

Figure 6.11: 3D reconstructions of subject 4.

(a) Pan & SCD-4



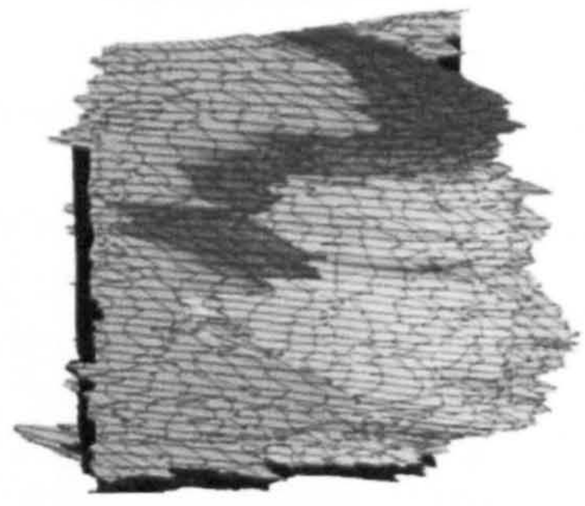
(b) Pan & SCD-4



(c) Pan & SCD-4

(d) Pan & SCD-6

(e) Pan & SCD-6

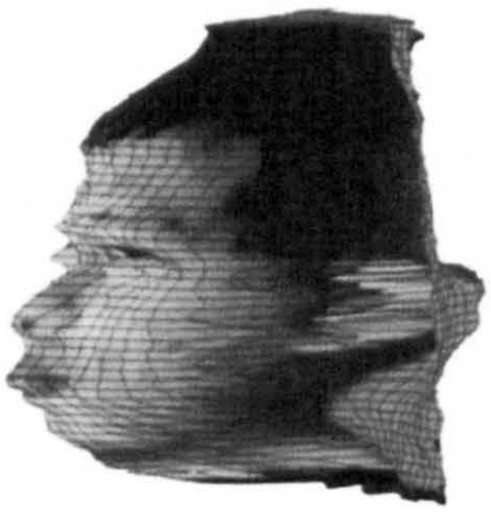


(f) Pan & SCD-6

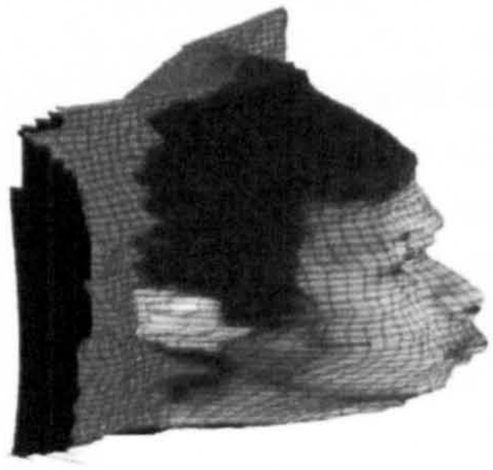
(g) Pan & MKC-4

(h) Pan & MKC-4

(i) Magarey & MKC-4

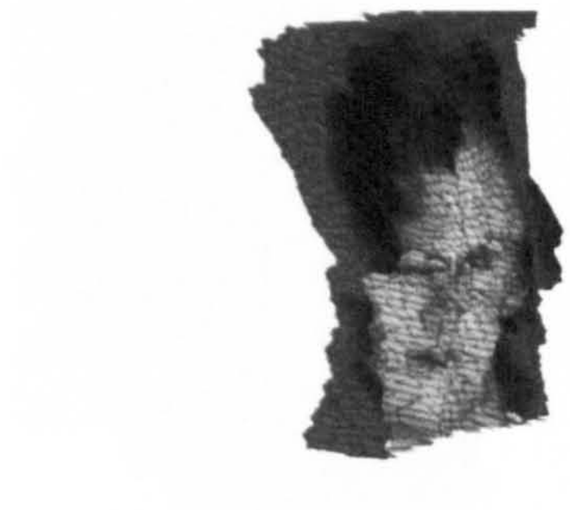


(j) Magarey & MKC-4

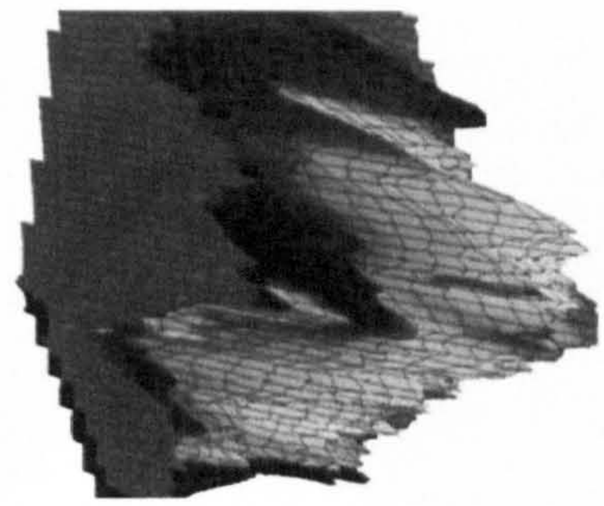


(k) Magarey & MKC-4

Figure 6.12: 3D reconstructions of subject 7.



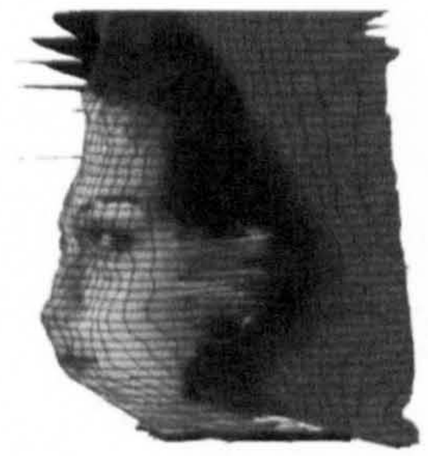
(a) Pan & SCD-4



(b) Pan & SCD-6

(c) Pan & SCD-6

(d) Pan & MKC-4



(e) Magarey & MKC-4

(f) Magarey & MKC-4

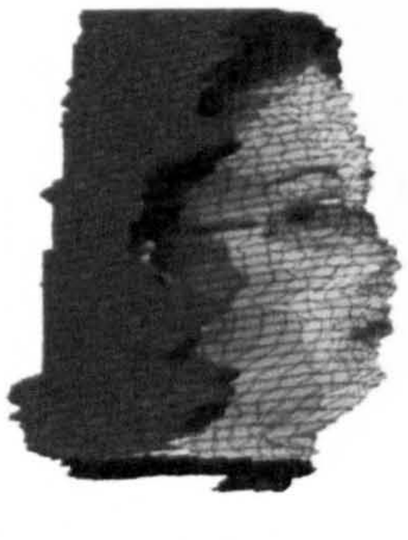
(g) Magarey & MKC-4

Figure 6.13: 3D reconstructions of subject 11.

(a) Pan & SCD-4

(b) Pan & SCD-6

(c) Pan & SCD-6



(d) Pan & MKC-4

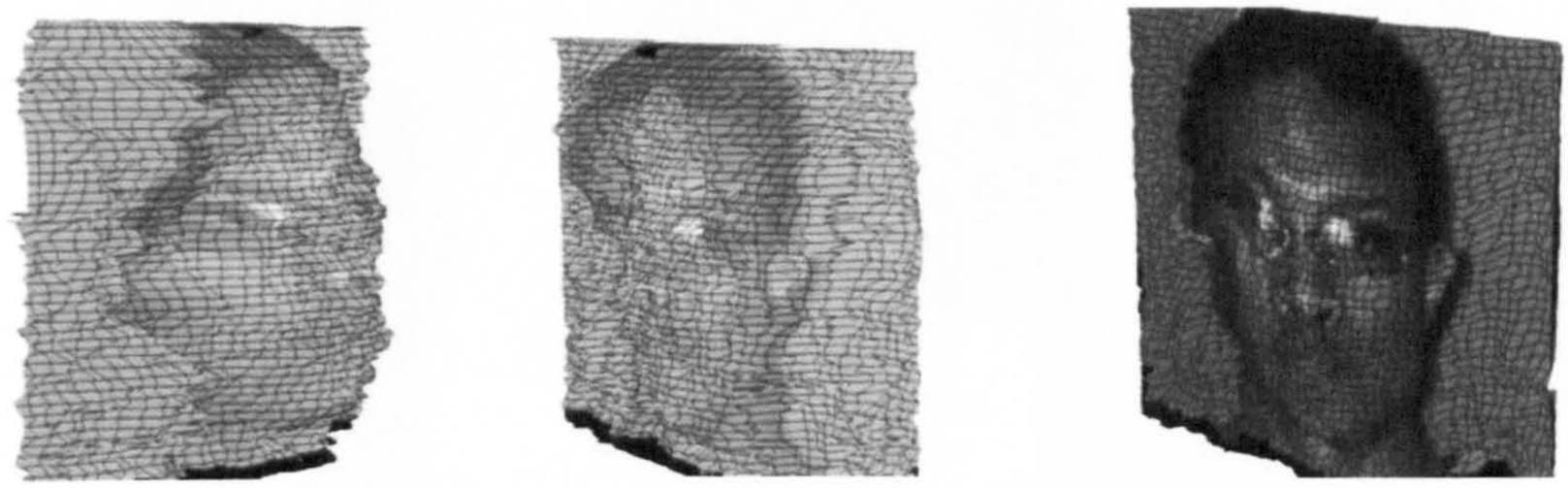
(e) Pan & MKC-4

(f) Pan & MKC-4

(g) Magarey & MKC-4

(h) Magarey & MKC-4

Figure 6.14: 3D reconstructions of subject 11 with glasses.

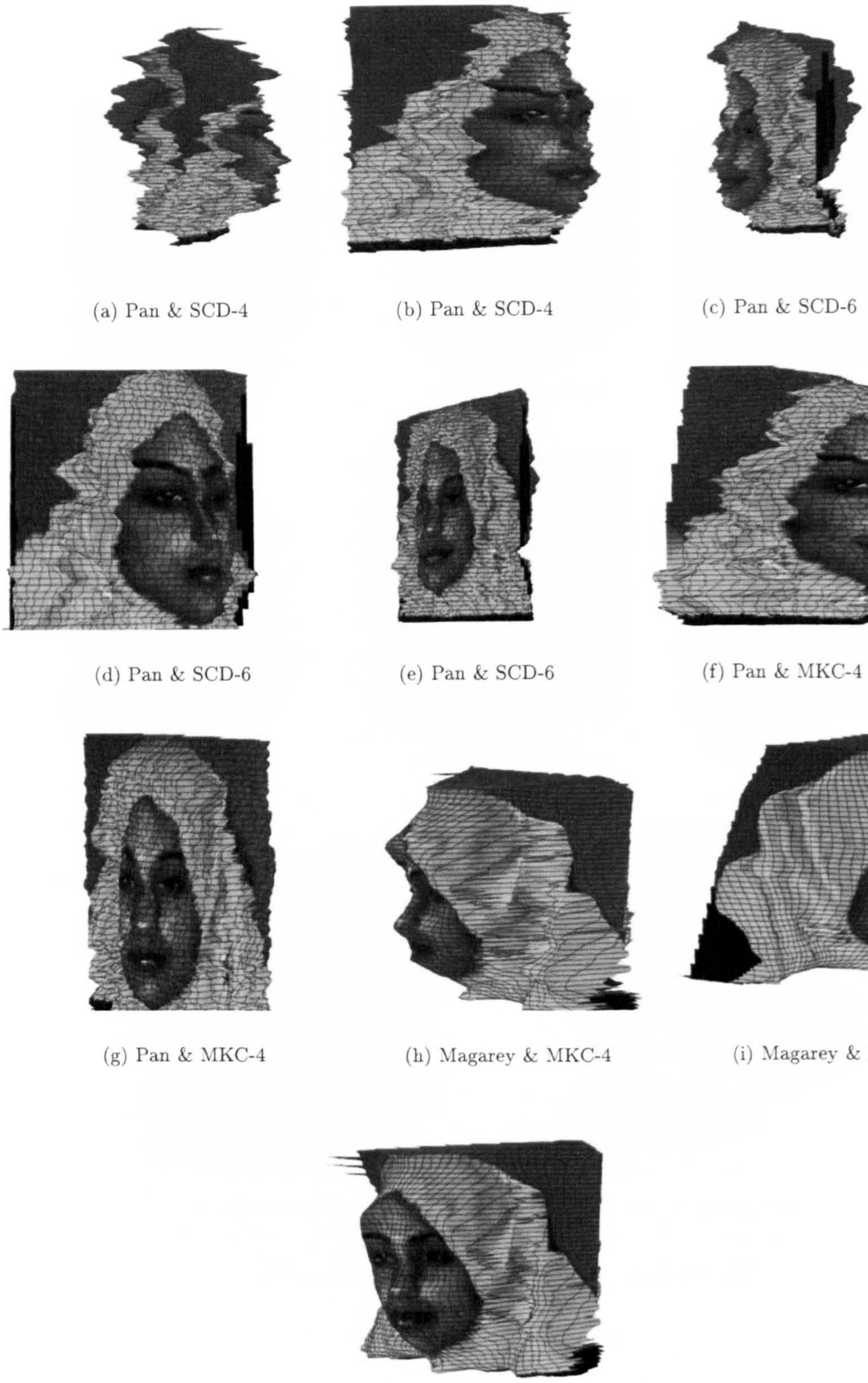


(a) Pan & SCD-4 (b) Pan & SCD-4 (c) Pan & SCD-6

(d) Pan & SCD-6 (e) Pan & MKC-4 (f) Pan & MKC-4

(g) Magarey & MKC-4 (h) Magarey & MKC-4

Figure 6.15: 3D reconstructions of subject 13.



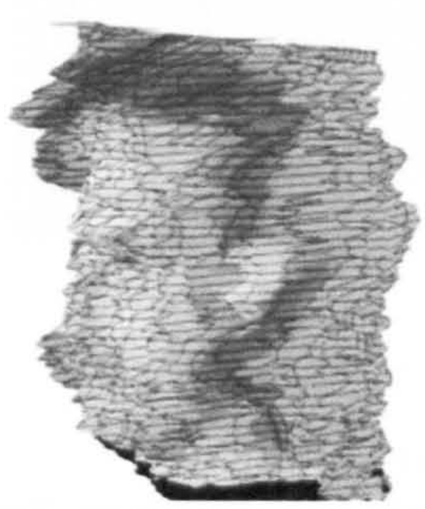
(a) Pan & SCD-4 (b) Pan & SCD-4 (c) Pan & SCD-6

(d) Pan & SCD-6 (e) Pan & SCD-6 (f) Pan & MKC-4

(g) Pan & MKC-4 (h) Magarey & MKC-4 (i) Magarey & MKC-4

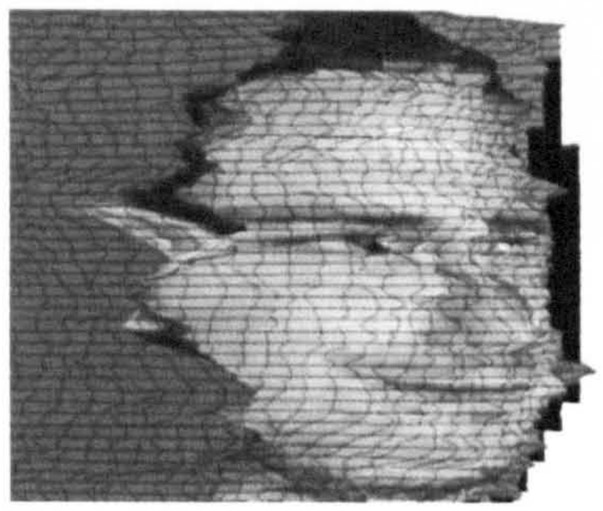
(j) Magarey & MKC-4

Figure 6.16: 3D reconstructions of subject 17 with scarf.



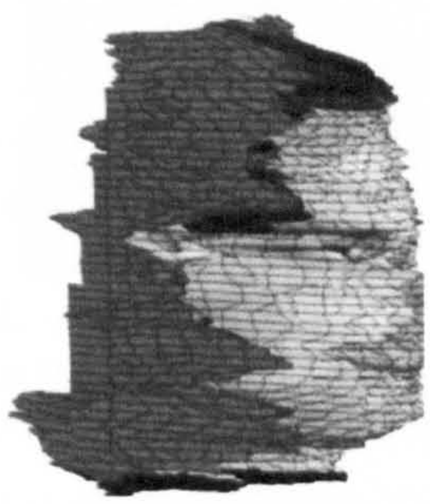
(a) Pan & SCD-4

(b) Pan & SCD-6



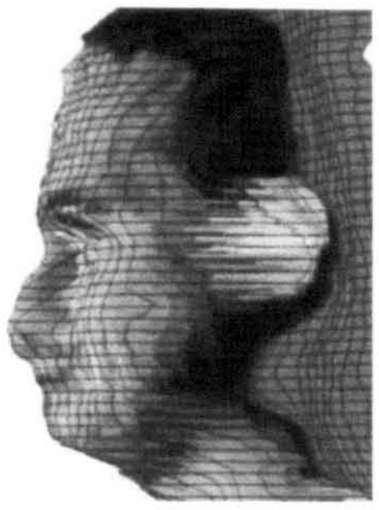
(c) Pan & SCD-6

(d) Pan & MKC-4

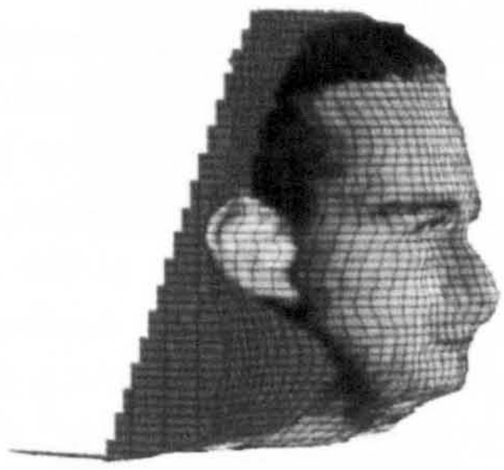


(e) Pan & MKC-4

(f) Pan & MKC-4



(g) Magarey & MKC-4



(h) Magarey & MKC-4

(i) Magarey & MKC-4

Figure 6.17: Pan + SCD6

Figure 6.18: 3D reconstructions of subject 18.

(a) (b) (c)

Figure 6.19: Effects of Reflection in the glasses (Magarey & MKC-4)

captured along with major deformations in the mouth and chin regions in (j).

6.5.2 Quantitative Analysis

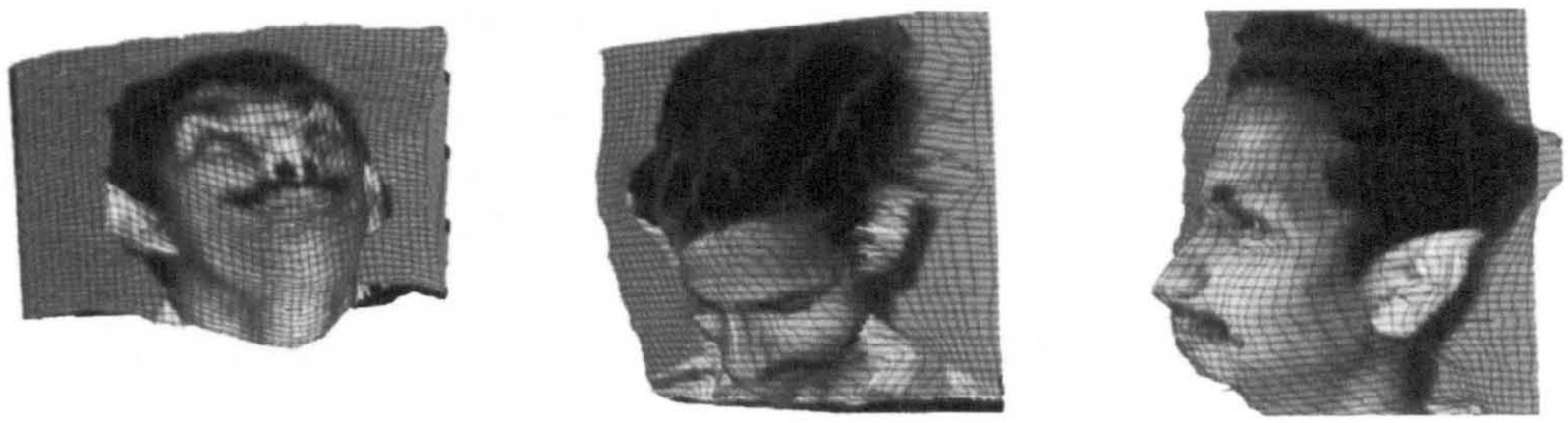
Backward image reconstruction (Lin & Barron 1994) is used to quantitatively evaluate the two matching algorithms used in this work. The concept is explained in Section 5.4.1 along with a simple numerical example.

The accuracy of the backward image reconstruction depends on the precision of the sub-pixel intensities. This in turn depends on the quality of interpolation and on the satisfaction of the implicit assumption that local intensity varies smoothly (Lin & Barron 1994).

Sub-pixel intensities are computed using bilinear interpolation (see Figure 6.21). The intensity value of the new pixel is computed by taking the weighted average of the four pixels in the 4-connected neighbourhood of the original pixel.

It is known that the assumption of smooth variation in local intensity, inherent to this technique, is not satisfied at occlusion boundaries and in image regions containing detailed texture (Lin & Barron 1994). However, for evaluating matching algorithms in this work, this assumption is taken to be satisfied. This is justified since large areas of the face are low frequency, for which the intensity values *do* vary smoothly. Also, the algorithms are compared using the same set of images, so the performance of each of the algorithms is affected equally due to the violation of this assumption. Furthermore, backward image reconstruction is one of the few reliable methods of quantitatively comparing the performance of matching algorithms in the absence of ground-truth data. Together, all these factors justify the use of this technique despite the fact that the assumption is not always satisfied (eg. in the presence of reflection)

The results of backward image reconstruction using Pan's algorithm in conjunction with the two Symmetric Complex Daubechies Wavelets , SCD-4 and SCD-6, and the Magarey and Kingsbury wavelet MKC-4 are presented in Table 6.1. For Magarey's matching algorithm, the backward image reconstruction is performed twice. First, matches for the pixels in the left image are sought in the right image (LR) and then vice-versa (RL). This is not done for Pan's algorithm since the error values for the first application of the reconstruction process are so high. This, combined with poor surface reconstructions contributed to the algorithm not being investigated any further. The actual RMS values are presented in the table along



(a) Look Up (b) Look Down (c) $\frac{3}{4}$ View

(d) $\frac{3}{4}$ View (e) $\frac{3}{4}$ View (f) Profile

(g) XY Look Down (h) Expression 1 (i) Expression 1

(j) Expression 2 (k) Harsh Lighting (l) Harsh Lighting

Figure 6.20: Bala models

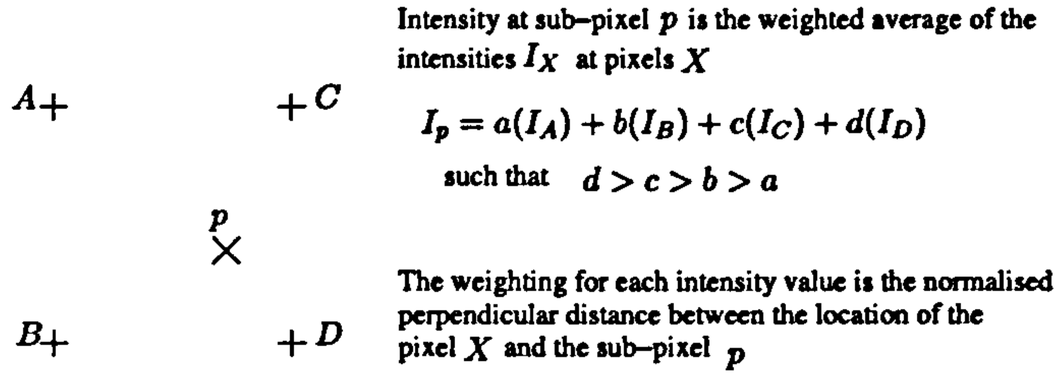


Figure 6.21: Bilinear interpolation to compute the intensity of sub-pixel p using the intensities at integer-indexed pixels $\{A, B, C, D\}$.

with the error values as a percentage of 256, the maximum possible disparity between the pixels.

Algorithm + Wavelet	Actual RMS	% Error
Pan + SCD-4	115.40	45.08
Pan + SCD-6	114.95	44.90
Pan + MKC-4	115.24	45.01
Magarey + MKC-4 (LR)	41.27	16.12
Magarey + MKC-4 (RL)	28.89	11.28

Table 6.1: Evaluation of Image matching algorithms using backward image reconstruction. The table shows the actual RMS values and the RMS values as a percentage of the maximum disparity between the pixels (256) for combinations of algorithm and complex wavelets.

As expected, the RMS error values for Magarey's algorithm are significantly lower than those for Pan's. It is surprising that the difference between the performance of the different wavelets in Pan's algorithm is so insignificant. The SCD-4 performed the worst, followed by MKC-4 and SCD-6. The large difference in the RMS values of LR and RL images was also unexpected since there are no obvious reasons for this asymmetry. In order to investigate this phenomenon, the algorithm needs to be tested on more images. It is possible that this is peculiar to the particular dataset used in this work. This investigation is not carried out in this work since it was felt that this would shift the focus from original problem of face recognition. It was in fact preferable that this distortion not be corrected prior to the face recognition stage so that its effects on the classifier accuracy can be determined.

The poor performance of Pan's algorithm can be attributed to its reliance on the matches at the coarsest level being correct. If the matches at this level are incorrect, the errors are propagated up to the finer levels since the refinement strategy is rather simplistic. However, given that this is the case, the results are very impressive.

6.6 Summary

This chapter described the two wavelets-based image matching algorithms investigated for matching the face images. Pan's algorithm is simple and easy to implement and has reported good results on aerial photographs. However, it has not been applied to face images. Magarey's algorithm on the other hand is complex to understand and its implementation is non-trivial. However, it has been tested on many images with ground truth data (Magarey 1997), and more importantly, it has been tested on face images (Magarey & Dick 1998) with impressive results.

As expected, Magarey's algorithm is more robust and this is indicated by lower RMS values determined using Backward Image Reconstruction process. This is not surprising since Magarey's algorithm is more involved and adopts robust regularisation and refinement strategies. By contrast, Pan uses a rather simple error correction technique. However, Pan's algorithm, despite its simplicity does produce some acceptable reconstructions. Magarey's algorithm fails in the presence of image discontinuities such as reflection (from subjects' glasses, for example). There is also a significant difference in the error rates of LR and RL images when using Magarey's algorithm. The reason for this was not investigated here since it was deemed beyond the scope of this work. It was instead decided to investigate if this difference in the LR and RL disparity maps has any bearing on the recognition rates. Pan's algorithm is not investigated any further in this work. The LR disparity maps are used to reconstruct 3D surfaces, which form the inputs for the 3D face recognition algorithm.

CHAPTER 7

Surface Matching

7.1 Introduction

Three-dimensional models generated using triangulation form the inputs for the recognition task in the 3D space. These models are noisy and very unlike the smooth laser scanned models generally seen in literature. The main sources of noise are errors in the image matching process, camera calibration, inaccuracies in triangulation and finally, in the process of going from 3D points to 3D surface meshes.

Johnson's spin-image representation and recognition algorithm (Johnson 1997, Johnson & Hebert 1997, Johnson et al. 1998, Carmichael et al. 1999, Johnson & Hebert 1999, Ruiz-Correa et al. 2001) is used for 3D face recognition. This technique was pioneered by Johnson at the Carnegie Mellon University's Robotics Institute in 1997 and has been successfully applied to 3D object recognition (Johnson 1997, Johnson & Hebert 1997, Johnson et al. 1998, Carmichael et al. 1999, Johnson & Hebert 1999, Ruiz-Correa et al. 2001), 3D object retrieval (Assfalg et al. 2004) and surface registration (Brusco et al. 2005), among others.

An overview of the representation and the recognition algorithm is presented in this Chapter. Further details can be found in Appendix E. The main advantages and disadvantages of the technique with reference to face recognition are discussed in Section 7.3. A description of the spin-image parameters used in this work and details of mesh pre-processing are presented in Section 7.4. This is followed by the results and conclusions.

7.2 3D Object Recognition using Spin Images

In the spin-image representation, 3D objects are represented by a polygonal surface mesh, and for every *oriented point* (3D point with surface normal) on this mesh, a 2D image - *spin image*, is generated. Informally, the spin-image concept can be thought of as follows: Imagine standing on the surface mesh at one of its vertices and doing a 360° turn on the spot, observing the view. Spin-images capture this 360° view of the object from each of the points on the mesh. It can intuitively be seen that the view captured in each of these images is entirely independent of the orientation of the object. This is an inherent advantage of

object-centred representation systems. The scene that is viewed does however depend very much on the "orientation" of the viewer - whether the viewer is standing on the outside or the inside of the surface and on other constraints. So in terms of 3D object representation, the spin-image contents vary based on whether the surface normal is oriented towards the inside or the outside of the object surface and it is constrained by the spin-image generation parameters.

7.2.1 Spin-Image Generation

More formally, an oriented point $\mathcal{O}$, at a surface mesh vertex is defined using the 3D position of the vertex, $\mathbf{p}$, and the surface normal, $\mathbf{n}$, at the vertex. Surface normal at a vertex is computed by fitting a plane to the points connected to the vertex by edges in the surface mesh (Johnson et al. 1998). A crucial requirement for the spin-image representation is that the oriented points are oriented outside the object surface. If the surface mesh is created from a sensor with a single viewing direction, then the normal direction can be chosen as the one pointing towards the sensor. Otherwise, each of the normals have to be oriented using the heuristic outlined in Appendix E (Section E.1).

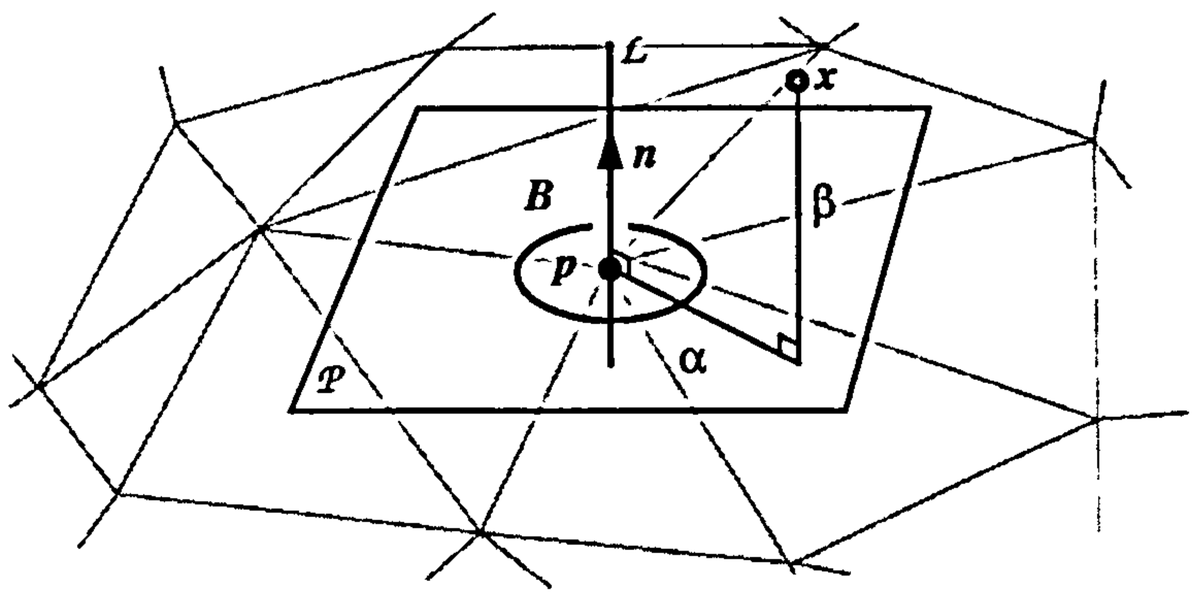


Figure 7.1: An oriented point basis created at a vertex in a surface mesh. The oriented point is defined by the 3D position $\mathbf{p}$ of the vertex and the direction of the surface normal $\mathbf{n}$ at the vertex. α is the radial distance to the surface normal line $\mathcal{L}$ and β is the axial distance above the tangent plane $\mathcal{P}$. Note that α can only take positive values, β however, can take both positive and negative values. $\mathbf{x}$ is another 3D point on the surface.

An oriented point, such as that shown in Figure 7.1, defines a basis or a local co-ordinate system with 5 degrees of freedom (DOF). The two co-ordinates of the basis are α, the perpendicular distance to the surface normal line $\mathcal{L}$, and β, the signed perpendicular distance to the tangent plane $\mathcal{P}$. An oriented point basis is a cylindrical co-ordinate system that is missing the polar angle co-ordinate, as this cannot be determined using just surface position and normal.

For every oriented point $\mathcal{O}$, a 2D accumulator indexed by α and β is also created. This encodes the density of points in each of the spin-images. Then for each vertex $\mathbf{x}$ on the surface of the object, the spin-map co-ordinates with respect to $\mathcal{O}$ are computed.

$$\mathcal{S}_{\mathcal{O}}(\mathbf{x}) \rightarrow (\alpha, \beta) = \sqrt{\|\mathbf{x}-\mathbf{p}\|^2 - (\mathbf{n}\cdot(\mathbf{x}-\mathbf{p}))^2}, \mathbf{n}\cdot(\mathbf{x}-\mathbf{p})) \tag{7.1}$$

If $\mathbf{x}$ meets some criteria based on distance from $\mathcal{O}$ (*support distance*) and angle between $\mathcal{O}$ and the surface normal of $\mathbf{x}$ (*support angle*), then the bin corresponding to its spin-map co-ordinates is determined. The 2D array is updated by incrementing the surrounding bins in the array - *not* the bin to which the point is spin-mapped. This takes into account noise in the data by spreading the position of the point in the 2D array. The point is bilinearly interpolated to the four surrounding bins in the 2D array. Thus by spreading the contribution of the point in the 2D array, it is made less sensitive to its position. Once all of the points on the surface of the object have been accumulated, a 2D array representation of the spin-image is generated. This process is depicted in Figure 7.2.

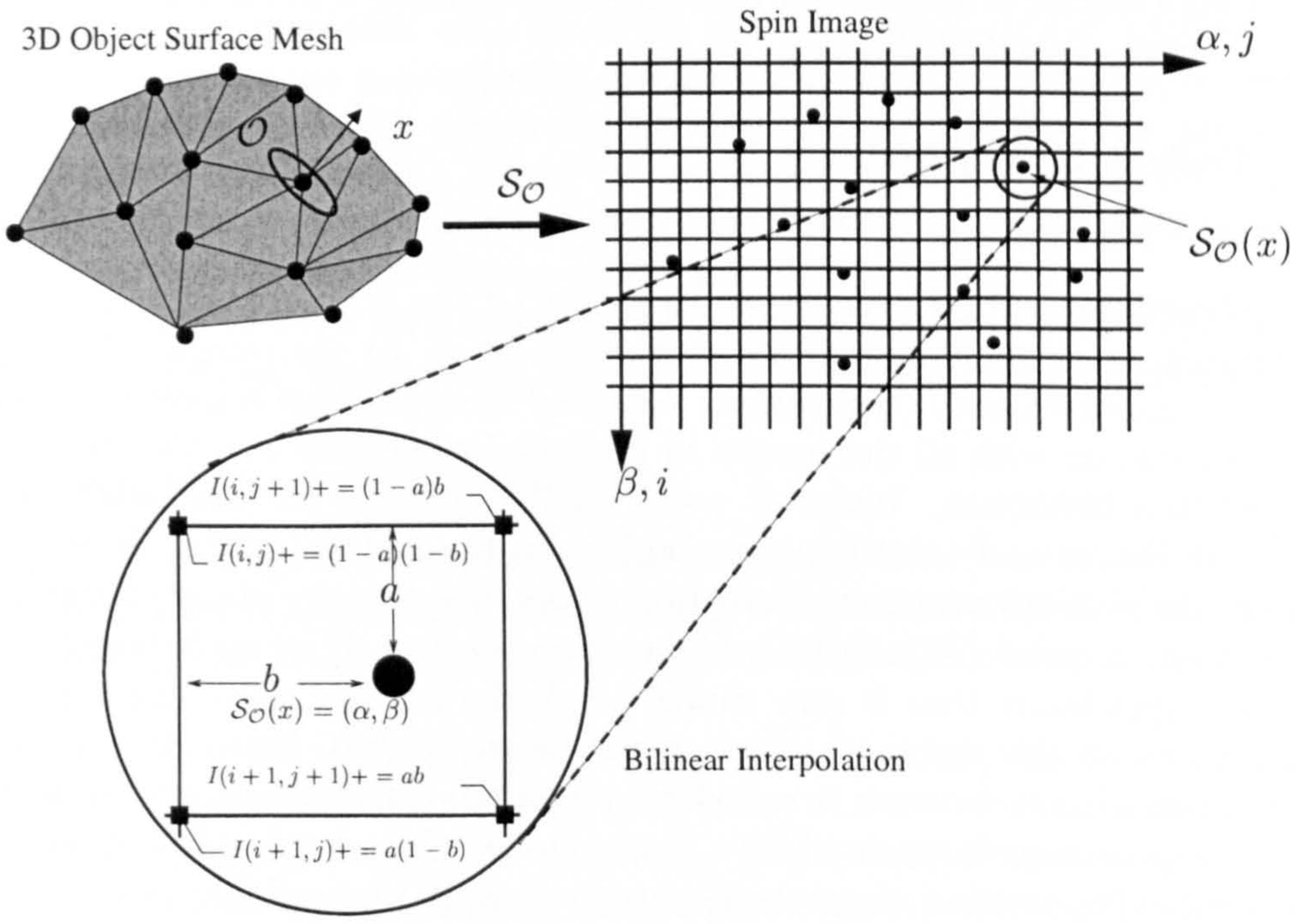


Figure 7.2: The spin-image generation process. A *spin-map* $\mathcal{S}_\mathcal{O}$ is a function that projects 3D points $\mathbf{x}$ to the 2D co-ordinates of a particular basis $(\mathbf{p}, \mathbf{n})$ corresponding to oriented point $\mathcal{O}$. Each oriented point $\mathcal{O}$ on the surface of an object has a unique spin-map $\mathcal{S}_\mathcal{O}$ associated with it, and when this is applied to all of the vertices of a surface mesh $\mathcal{M}$, a set of 2D points is created.

The descriptiveness of the spin-images is controlled by three spin-image generation parameters. *Bin Size* is the storage size of the bins in the spin-images and regulates the effect of individual point positions. *Image Width* sets the size of the spin-image and controls how "localised" or "globalised" the spin-images are. For a given bin size, the larger the image width, the more globalised the spin-images. *Support Angle* limits the angle between the normal of the oriented point basis and the normal of points contributing to the spin-image. A large support angle may map points from outside the object surface to the spin-images. The parameters are explained in more detail in Appendix E.

7.2.2 Spin-Image Matching

Two spin-images P and Q are compared using the *similarity measure*, $C(P,Q)$ (equation E.6). It combines the normalised linear correlation coefficient with a confidence measure in the value of the correlation coefficient. Linear correlation coefficient is a standard way of comparing linearly related images. The confidence measure in the correlation coefficient is obtained by considering the amount of overlap between two surfaces to be matched. Combining both these measures in the similarity measure means that the spin-image matching process takes into account only those regions of the surfaces where there is an overlap. This is a distinct advantage when data from real scenes is compared, since real data is often riddled with noise in the form of clutter and occlusion. A large similarity measure between two spin images indicates high degree of correlation between them. Hence, a point correspondence between the two surfaces is established.

7.2.3 Object Recognition

Spin-Image Matching

For the object recognition task, spin-images for all the surface models in the training set are generated a-priori and stored in a *spin-image stack*. At the recognition stage, a point is selected at random from the test surface mesh and its spin-image is generated. The test spin-image is correlated with all the images in the spin-image stack and the similarity measures are stored in a histogram. Incorrect point correspondences are eliminated by ranking the similarity measures and identifying the outliers in the histogram (See Section E.5). Upper outliers in the similarity measure histogram correspond to pairs of test/training spin-images with similarity measures significantly higher than the rest. If no outliers exist, then the test point has a spin-image that is very similar to all the training spin-images and definite correspondences with this test mesh point cannot be established. From these outliers, plausible point correspondences between oriented points on the test/training surface mesh are established. This procedure is repeated for a fixed number of randomly selected points on the test surface mesh. The number of points selected is directly proportional to the amount of clutter in the test model. Johnson (Johnson 1997) recommends setting this number somewhere between 0.5 and 0.05.

This results in a set of oriented point correspondences between test/training surface meshes, for *each* mesh in the training set. These sets of point correspondences are then filtered and grouped together using *geometric consistency* to compute the transformation from training set model to the test set model. The surface matches are verified and refined using a modified iterative closest point (ICP) algorithm. A match between the surfaces is established if the number of correspondences between the two surfaces is greater than a threshold value, usually set as a percentage of the total number of points in meshes.

Correspondence Filtering

Although some of the incorrect matches are eliminated using the histogram of similarity measures, many still remain. These may be due to symmetry in the data, which causes the spin-images of two points to be similar. It may also occur because proximal points on the surface mesh can have similar spin images or because there are multiple occurrences of the model in the test scene.

Incorrect matches are eliminated by filtering the point correspondences on the basis of geometric consistency (equation E.7). Geometric consistency is a measure of likelihood that two correspondences can be grouped together to calculate a transformation of the test model to the training model (or vice-versa). If a correspondence is not geometrically consistent with other correspondences then it cannot be grouped with other correspondences to calculate a transformation and should be eliminated.

Correspondences are filtered by setting a threshold value T_{gc} for the geometric consistency (equation E.7). Two point correspondences $\mathbf{C}_1$ and $\mathbf{C}_2$ are said to be geometrically consistent if their geometric consistency $D_{gc}(\mathbf{C}_1, \mathbf{C}_2) < T_{gc}$. Johnson (Johnson 1997) suggests setting T_{gc} equal to 0.25 to enforce a strong geometric consistency between the correspondences.

First, all the correspondences $\mathbf{C}_i$ are put in a list L. Next, for each "test correspondence" $\mathbf{C}_i$ in L, $D_{gc}(\mathbf{C}_i, \mathbf{C}_j)$ is computed with all of the other $\mathbf{C}_j$ in L. If the number of correspondences in the list that are geometrically consistent with the "test correspondence" (i.e. the number of correspondences $\mathbf{C}_j$ where $D_{gc}(\mathbf{C}_1, \mathbf{C}_2) < T_{gc}$) is at least one quarter of the number of correspondences in L, $\mathbf{C}_i$ passes the geometric consistency test. Otherwise it is not geometrically consistent with enough correspondences in L and is removed from the list L. This process is repeated for the correspondences $\mathbf{C}_i$ in L.

This results in a list of point correspondences $L = \{\mathbf{C}_1, \ldots, \mathbf{C}_n\}$ that are most likely to be correct and can be grouped together into sets that can be used to compute transformations.

Grouping geometrically consistent matches

The *grouping criterion* W_{gc}, is the geometric consistency distance (equation E.8) augmented by a weight that promotes grouping of correspondences that are far apart. Correspondences that are too close together result in erroneous transformations. This is a consequence of noise in the point position.

Given a list of most likely correspondences $L = \{\mathbf{C}_1, \ldots, \mathbf{C}_n\}$, the grouping procedure is applied for each correspondence in the list as follows. First, a seed correspondence $\mathbf{C}_i$ in L is selected and a group $G_i = \{\mathbf{C}_i\}$ is initialised. Then, the correspondence $\mathbf{C}_j$ in L, for which $W_{gc}(\mathbf{C}_j, G_i)$ is a minimum (correspondences geometrically consistent and far apart) is established. $\mathbf{C}_j$ is added to G_i if $W_{gc}(\mathbf{C}_j, G_i) < T_{gc}$. The geometric consistency threshold T_{gc} is set between 0 and 1 (usually around 0.25). This process of adding the correspondence with minimum grouping criterion continues until no more correspondences can be added to G_i. This procedure is repeated for each correspondence in L, and the end result is n groups, one for each correspondence in L.

This grouping algorithm allows a correspondence to appear in multiple groups. This is essential for handling model symmetry, as correspondences along the plane of symmetry contribute to two distinct transformations. A set of potential rigid body transformations are calculated from each of the groups of correspondences. The transformations and the associated correspondence groups are then verified to eliminate any further mismatches and establish the final test/training model match.

Verifying Matches

This stage of the matching process aims to find the best match(es) between the training set models and the test set model by eliminating inconsistent data. During the verification process, point correspondences are spread over the surfaces of the test model and the training

set models from the initial correspondences established by the matching process. If many correspondences are established through spreading the matches, an association (or a match) between the training model and the test model is validated. In addition to verifying possible matches, this method also improves the transformation between the two models.

The verification algorithm is a modified version of the Iterative Closest Point Algorithm (ICP) of Besl and McKay (Besl & McKay 1992) and Zhang (Zhang 1994). Johnson (Johnson 1997) modifies the generic ICP by limiting the closest point distance measurement only to those areas in the two sets that overlap. This is accomplished by growing the closest point correspondences from initial correspondences established by the matching process thus far.

The verification process involves first computing the transformation from the training model to the test model using the correspondences established so far. This transformation is then applied to the training model. Each of the initial correspondences is propagated outwards. If the distance between the test model and the closest training set model is less than a threshold D_v then the test model points directly connected to it by edges are turned into correspondences. Johnson recommends setting D_v at two times the mesh resolution in order to allow for noise but not for establishment of correspondences in areas of no overlap. This correspondence propagation process is repeated until no more correspondences can be created.

If the initial transformation is correct, then a large number of points will be brought into correspondence. If the transformation is poor, the number of correspondences will remain close to the original number. Hence, the measure of validity of the match is *number of correspondences after verification.* This is related to the total aligned (or overlapped) surface area between the test and the training models. If the number of correspondences is greater than 1/10 the total number of points on the model, then the transformation is considered valid, otherwise not. The additional matches generated from this process are used to refine the transformation, after it has been verified and accepted.

The matching process described here is repeated for all the models in the training set and a match is established with the model that gives the highest number of correspondences between the points on the test and the training surface mesh.

In (Johnson 1997), Johnson also presents three variations on the spin-image representation and the recognition algorithm. These are described in Section E.9 of Appendix E.

7.3 Spin-Images: Advantages and Disadvantages

7.3.1 Advantages

Spin-image representation of 3D objects has many advantages, the main one being that it is an object-centred representation rather than a viewer-centred one. Object-centred systems can be more compact than viewer-centred systems because a single surface representation describes all views of the surface (Johnson & Hebert 1999). Object-centred systems are also view-independent. This means that unlike 2D recognition systems, only one model of the object is required for training the classifier. This has the obvious advantages of speeding up the processing time and decreasing storage requirements. Pose invariance is also inherent to these systems. From the point of view of face recognition, this is a very important consideration since most 2D face recognition systems perform poorly in the face of pose-variation.

Spin-image generation process and the recognition algorithm both avoid error-prone feature extraction and segmentation. This is particularly advantageous since in the approach adopted in this work, the input data is already so erroneous. This may not be a major concern for applications that use accurate data acquisition methods such as laser scans. However, when the 3D data comes from noisy stereoscopy, it becomes very important to avoid any kind of additional processing that may introduce more errors.

Surface matching algorithms often require some knowledge of the transformation between the models in the training set and the test model. While this may be possible for many applications, it was not an option for a face recognition system. A system that is implemented in a very controlled environment can enforce such conditions by limiting the subjects' freedom of movement during the capture of test data. However, this work imposes no such constraints on the input data. Spin-image representation is designed to work in the absence of any information regarding the transformation between the training and the test images. If the transformation is known, then it can increase the speed and accuracy of matching, but it is not a requisite.

The other major consideration behind the choice of spin-images for face recognition was that they have been reported to give highly accurate results when the data is noisy and there is only a small overlap between the training and the test models. As explained before, input data from stereo matching is extremely noisy. Often there is only a small region on the entire surface that has reconstructed well. Most surface registration algorithms would fail under these circumstances. However, because of the way spin-image matching works, even a small overlap or a small noise-free area on the surface can, in theory, produce an accurate match.

It is also a very flexible representation and imposes few constraints on the training and test data. For example, the mesh resolution of the training and the test images can be different, as long as the mesh generation parameters are adjusted accordingly. Since determining the mesh statistics is a simple process (which can be automated), the mesh generation parameters can also be adjusted appropriately without any user interference. This makes the recognition algorithm not only immune to variations in pose, but also to variations in scale. However, scale invariance is not dealt with in this work.

Finally, a major concern in this work has been to develop a system that can eventually function in uncontrolled environments. In applications such as crowd surveillance, the system needs to perform to accurately even when the input image contains more than the object to be recognised. And equivalently, when the object to be recognised is not entirely visible. The robustness of spin-images has been tested in both these situations and good performance has been reported in the presence of both, clutter and occlusion.

7.3.2 Disadvantages

Although the authors of the technique report very good recognition accuracy, these claims should be considered with caution. The main reason being that it has only been tested on two libraries (toy and plumbing) (Johnson 1997). The 3D objects used in these libraries have very distinct shapes. Faces on the other hand, are instances of the same 3D object, and most of the time the differences between these instances is subtle rather than gross. Spin-images from different faces are not as distinct as those for the toy or the plumbing libraries, and this is a major drawback of this technique. This may also detract from the benefits of the approach being object-centred and therefore requiring only one training model per class.

More than one models may be required per class to train the system adequately. Since the processing time increases linearly with the number of models in the library (Johnson 1997), this can make the system, however accurate, useless for practical applications.

The other problem is that Johnson's library contains fewer models than the Sheffield Dataset. His libraries contain at most 20 models, whereas the Sheffield Dataset has 692. In addition, his models are complete, closed models, much like the 3D objects themselves. The models in the Sheffield Database are surfaces rather than closed models, since the back of the head is not imaged during data acquisition. Further, since Johnson's objects are small, the corresponding meshes are also small (100's of vertices). However, the face meshes are extremely large in comparison (1000's of vertices). It is envisaged that the combination of large models and a large library will make the recognition process very sluggish. Compressing the model or the library spin-images is not an option at this stage.

Hence, although the spin-image representation has many desirable features, its performance may be hindered by certain factors. On balance though, it was felt that the advantages may outweigh the disadvantages and that it was worth investigating the technique's application to the face recognition task.

7.4 Spin-Image Parameters for face images

Choice of appropriate parameters is crucial for the generation of appropriate spin-images and for accurate object recognition. First, the mesh pre-processing parameters are chosen.

Depth information obtained from the image matching and the triangulation processes has to be represented as a polygonal mesh for spin-image generation. Triangular meshes were generated using Delaunnay triangulation. The meshes then have to be processed to eliminate errors and ensure that spin-images containing useful information are generated (as opposed to spin-images containing misleading, erroneous information).

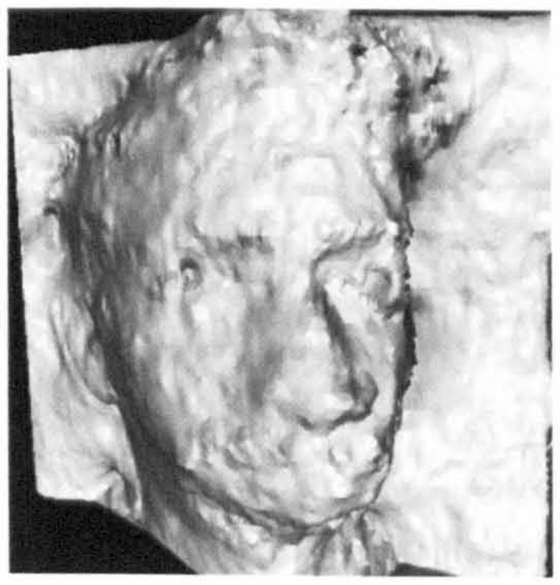
Mesh pre-processing is summarised below:

- The normals are then oriented outwards. Without this crucial step the algorithms are likely to fail, as one of the core assumptions of the spin-images theory is violated.

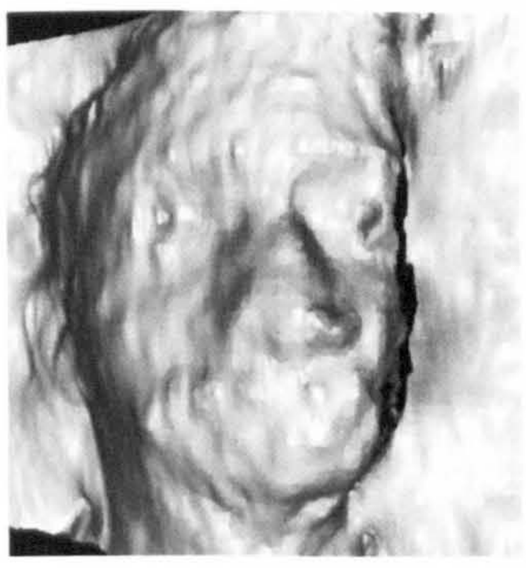
- The meshes are cleaned in order to remove overly-long edges. Overly long edges are defined as edges that are longer than 2 times the mesh resolution. Mesh resolution is defined as the median edge length. Similarity in the edge length is one of the implicit assumptions in this representation. Overly long edges correspond to outlier points and hence, are removed. Random patches (< 100 vertices) that are not connected to anything are also removed. They are seen as belonging to the background noise, and would not contain any useful classification information.

- Meshes are then smoothed using 25 iterations of "smoothing without shrinking filter" of Taubin (Taubin 1995). The number of iterations are chosen after a qualitative analysis of the meshes. Too few iterations mean that meshes are still very noisy, while too many mean that valuable information is lost.

- The mesh is simplified, so that it consists of user-specified number of polygons. The number of polygons results in a trade-off between the level of detail and the processing time. It may sometimes be difficult to achieve the desired number of polygons. This

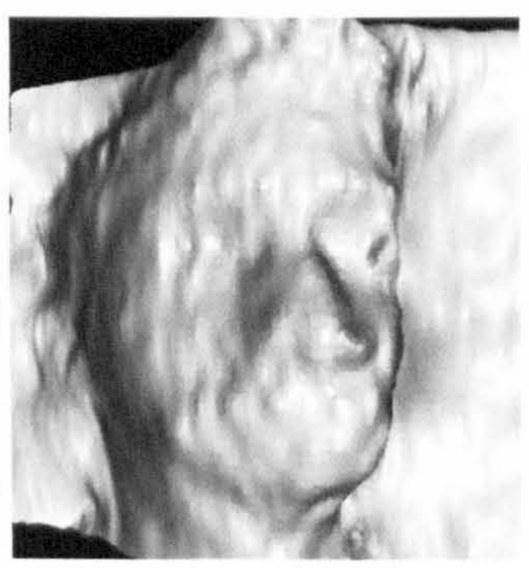


(a) Unprocessed mesh



(b) Smoothed, 10 iterations

(c) Smoothed, 25 iterations



(d) Smoothed, 50 iterations

Figure 7.3: An unprocessed mesh, along with meshes smoothed using 10, 25 and 50 iterations of the smoothing filter. Excessive smoothing in the last mesh means features such as the eyes are less discernible. This effect is less noticeable when the mesh is smoothed using 25 iterations.

is a consequence of the mesh generation and the simplification process. In that case, reduce the mesh size to the minimum possible, going up in steps of 250 or 500 polygons. This speeds up the mesh pre-processing and results in fewer errors. Controlling the approximate size of all the meshes in the library also allows the user to control the mesh resolution and the processing time to an extent.

(a) Unprocessed mesh

(b) Simplified, 500 polygons

(c) Simplified, 1000 polygons

(d) Simplified, 5000 polygons

Figure 7.4: The original mesh with approximately 32,000 polygons, simplified to 500, 1000 and 5000 polygons. At 5,000 polygons, it is just about possible to correspond the simplified mesh with the original.

The algorithm requires setting the threshold values for some of the variables. The threshold values recommended by Johnson (Johnson 1997) are used in this work and are listed below.

- Geometric consistency threshold, T_{gc} is set to 0.25 (Section E.6).
- Geometric consistency weight γ is set to 4 times the mesh resolution to encourage grouping between correspondences which are at least four times the mesh resolution distance away from each other (Section E.7).

- ν, the factor that weighs the surface normal information against position information is set at 2 times the mesh resolution. The normals of vertices, as a result, have more of an effect on the distance metric than the positions of the vertices (see Section E.8).
- Verification threshold, D_v, is set to 2 times the mesh resolution. This dictates, at the verification stage, which points on the training model may have matches on the test model (see Section E.8).

For the spin-image generation parameters, Johnson recommends setting the bin size between 1 and 2 times the mesh resolution as this blurs the position of the points sufficiently, without compromising on the shape description. For face images, it was set to 1.5 to strike a balance between too much averaging so that there is loss of information, and not enough so that the images are riddled with noise. The image width and the support angle were chosen by conducting a leave-one-out cross validation experiment on a small dataset. The mesh size was set to 1200 polygons. This is an extremely coarse mesh, so considerable information is lost, however, resulting computations are extremely fast. Dataset D_2 containing 165 images is used. Figure 7.5 is an example of the 2D images in this dataset. The 3D models corresponding to these images are used in this study.

Figure 7.5: A sample image class from the dataset used in the preliminary experiments to determine optimum values for the spin-image parameters.

Spin-images are generated and matched using the "Mesh Toolbox" developed at the Vision and the Autonomous Systems Centre within the Robotics Institute at Carnegie Mellon University, where the technique was pioneered [1]. The results of this preliminary experiments are depicted in Figure 7.6. It is easy to see that the best recognition rates are obtained when the image width is set to 15 and the support angle to 90°. The recognition rates achieved are extremely low, even though all the points on the training and the test models were used for matching.

Consequently, the mesh resolution was altered to obtain better recognition rates for the experiments using the entire Sheffield Dataset. The original meshes generated using Delaunnay triangulation consist in the region of 50,000 polygons. Ideally, matching process should involve as many of these points as possible. But this would slow the system down to the point of collapse. In order to strike a reasonable balance between processing time and recognition accuracy, the mesh size for the next set of experiments was set at 5,500 polygons, and 80% of the points on the training and test models were used for matching.

The final mesh size is a mere $\frac{1}{10}$ of the original mesh, and this loss of information manifests in poor recognition rates. However, owing to the complexity of the algorithm, this simplification of the meshes was necessary.

[1] www.cs.cmu.edu/~vmr/software/meshtoolbox/introduction.html

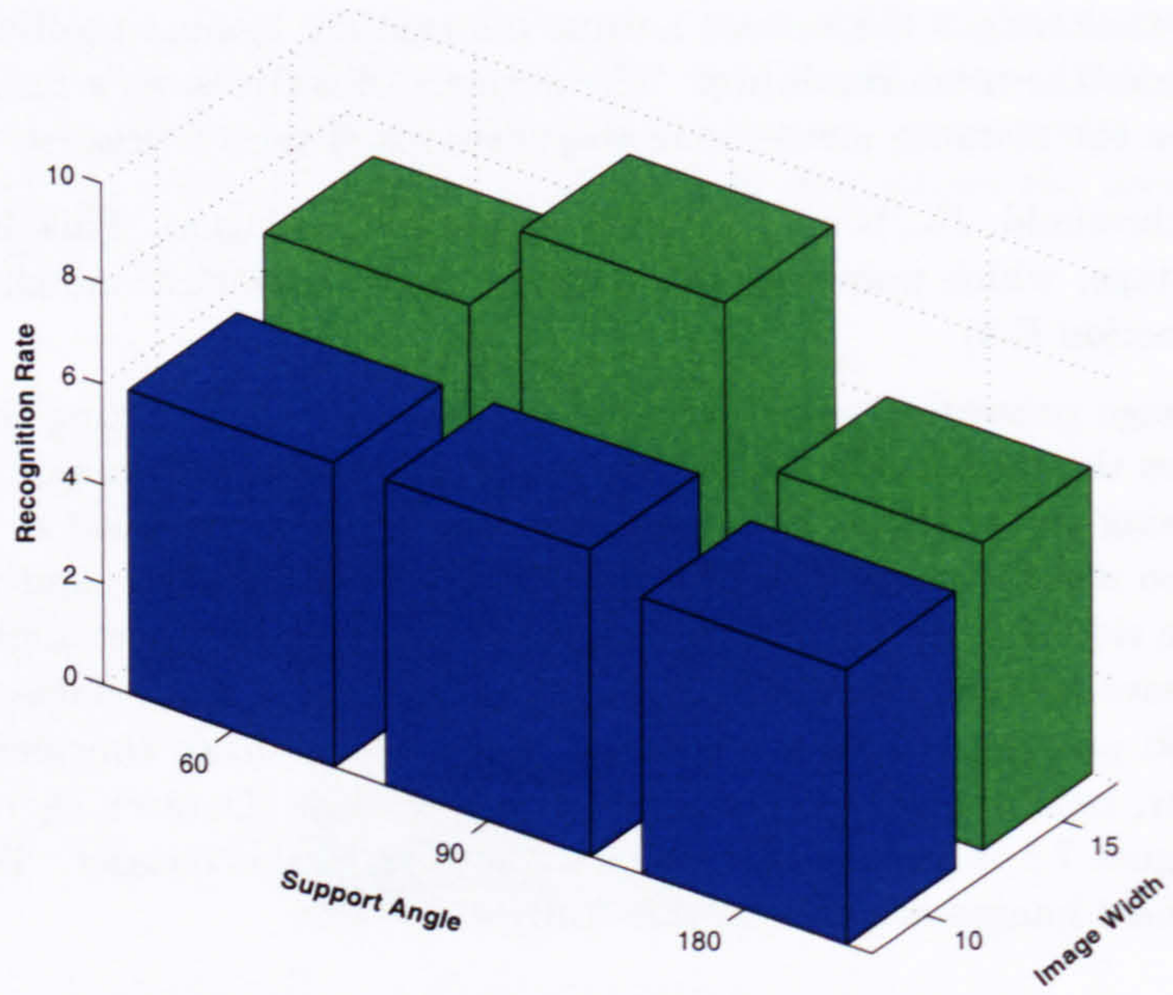


Figure 7.6: A sample image class from the dataset used in the preliminary experiments to determine optimum values for the spin-image parameters.

7.5 Face Recognition Results using Spin-Images

Spin-image recognition algorithm was tested using Dataset D_1 in two experiments: Leave-One-Out (LOO) cross validation and recognition using a reduced training set.

Leave-One-Out (LOO) cross validation partitions the dataset of size D into a training set of size $D-1$ and a test set of size 1. The classifier is trained D times, leaving out 1 image each time (for testing). All the images in the dataset are used for training the classifier, which can potentially maximise its accuracy. In addition, all data can be used to test the classifier. This has the advantage of not only being able to test the classifier more thoroughly, but it may also highlight some important features about the dataset. Furthermore, since it does not involve any random sampling, it always returns the same results, given that the same classifier and associated parameters are used. The main drawback of this method is that the algorithm has to be run D times. For large datasets, or complex algorithms, this can take a long time.

The second experiment uses only one image per class to train the classifier. The 3D models corresponding to the 2D images in Figure 4.2 form the training set. The main aim of this experiment was to test the claim that object-centred systems only require one training image due to their pose-invariance. If this claim can be verified, then it may negate some of the disadvantages of spin-image representation, and may still justify their use in real-time applications. It also tests how well the classifier can cope with input images that deviate significantly from the training set.

Note that during the experiments, images of an individual with glasses and without glasses are treated as belonging to two different classes. It was felt that this would give a better idea

of how well the algorithm is able to distinguish between images of the same individual with and without glasses. The spin-images of an individual with glasses are very different to the spin-images of the same individual without glasses in the region near the eyes. Spin-images from other regions of the face are identical. The idea was to see if the recognition algorithm could distinguish between the two image classes of the same individual and classify them correctly. The recognition rate in both the experiments is defined as the number of correctly identified individuals divided by the total number of images in the dataset. Recognition rates for the first nearest-neighbour (denoted by 1-NN) and the recognition rate when the correct match is identified among the top 5 matches (denoted by Top5-NN) are presented in Table 7.1.

The results of the two recognition experiments are provided in table 7.1.

Experiment	1-NN (%)	Top5-NN (%)
Leave-One-Out	18.42	32.75
Reduced Training Set	17.90	25.33

Table 7.1: Results of the preliminary experiment using Dataset D_2, to determine the optimum values for the spin-image generation parameters.

Two things are immediately obvious from Table 7.1. First, that the recognition rates are extremely low - much lower than those reported in the literature for even the most basic algorithms. Second, that the difference between the 1-NN recognition rates of the two experiments is extremely small - 0.52%.

The poor recognition rates are extremely disappointing. The algorithm was not expected to yield the kind of recognition rates that are generally reported in face recognition literature, primarily, because of the nature of the input data. Both, training and test models are extremely noisy. They are not the smooth laser-scanned or structured-light scanned models seen in literature. The noise levels in the data would have a significant impact on the recognition rates. Johnson (Johnson 1997) tests for the effects of noise in the input data, but the noise he introduces into his models is not as extensive as that seen in the face models. Further, it is assumed in this work that optimal spin-image parameters determined using the coarses meshes are also optimal for the more detailed meshes. This may not necessrily be the case, however, time constraints did not permit further investigation in this direction. This combined with the similar nature of the models, and hence the spin-images, is thought to be responsible for the algorithm's poor performance.

This problem can be dealt with in a number of ways. First, the models can be smoothed extensively at the pre-processing stage. However, this causes valuable information to be lost. Smoothing may eradicate much of the high-frequency data, which is mistaken for noise. This is a problem since much of the useful information on the face is also in high-frequency areas. The low-frequency areas such as the forehead and the cheeks contain salient information that aids classifier accuracy, but on their own, they are not sufficient to classify faces correctly, especially using this technique.

The classifier performance may be improved by increasing the mesh size to at least half the original size. This reduces the errors that result from the mesh pre-processing tasks. This has the advantage of the meshes being more information-rich, but is also has the disadvantage of the meshes retaining much of the noise. The best alternative perhaps is to use noise-filtering

at each stage prior to recognition, from image matching to triangulation. This would result in the input meshes being more accurate.

The marginal difference between the 1-NN recognition rates of the LOO cross-validation and the reduced training set experiments largely verifies the claim that object-centred systems are indeed very elegant and do not require extensive training data to perform accurately. It also underscores the point that it is the accuracy of the input data rather than the amount of training data available that determines accuracy of this classifier. This fact is also echoed by the results of the experiments conducted to determine the value of parameters. The meshes used in those experiments were much smaller, containing 1200 polygons on average, whereas the meshes used in these experiments contained on average 5500 polygons.

The recognition algorithm used by the spin-image representation is essentially a nearest-neighbours algorithm. One of the main drawbacks of nearest-neighbours is that the test image needs to be compared against every image in the training set. For an algorithm as complex as this one, classifying each image becomes an extremely time-consuming task. However, the results of the reduced training set experiment have shown that because of the way the spin-image representation was designed, the need for more than one training image per class has been eliminated,

The difference in the Top5-NN recognition rates between the two experiments is much larger - circa 7.4%. This difference is a consequence of the LOO cross validation training set be considerably larger. In a nearest-neighbours algorithm, if the training set is large, then there is an increased chance of the test image being matched with another image of the same person.

7.6 Conclusions

The results from the previous section are extremely disappointing. They are also very inconclusive with regards to whether or not the depth information has any added benefits for face recognition. What is certain is that although the spin-image representation is powerful, it is not suitable for this particular approach to face recognition. Depth computation using stereo images is very error prone. When both training and test data are so error-prone, most classification algorithms are likely to fail.

Even if the recognition accuracy was higher, using it with our data is not practical. The spin-image generation and recognition are both lengthy and complex processes. This increases the processing time enough for the the system to be useless for real-time applications. If the input data is captured using more accurate means such as laser scans, structured light scans or even specially designed stereo cameras, then the resulting 3D models are far more accurate than those used in this work. It is believed that this would increase the accuracy enough to make the system comparable with other 3D face recognition systems. Further, if the initial data is very accurate, then downsampling and smoothing would reduce the noise content of the models rather than the information content. This would speed up the processing time further. It may also make it possible to use at least one of the variants, MA2, of the algorithm to compress the model spin-images. This would eliminate much of the redundancy in the spin-images of faces. Even if the compression algorithms cannot be implemented, a smaller fraction of the points may be chosen for matching the test and the training set images. This would also reduce the processing time without affecting the accuracy significantly. If these changes can be implemented, then the system is likely to perform recognition in real-time,

as claimed by its authors.

In conclusion, for accurate 3D face recognition using spin images, it is very important that the data is as error-free as possible. Hence, generating depth information using noisy stereoscopy is not an option. Another, more robust recognition algorithm is required if stereo images are to be used. Alternatively, a more accurate method of data acquisition should be employed.

Further investigations need to be carried out using more accurate input data to draw any concrete conclusions about the algorithm's suitability for face recognition.

7.7 Summary

This chapter described Johnson's (Johnson 1997) spin-image representation and recognition algorithm for 3D objects. Its relative merits and drawbacks with respect to face recognition were discussed, and finally it was tested on the 3D models from the Sheffield Database.

The best values for the algorithm parameters were determined using dataset D_2. Once these were established, two experiments were conducted to test the algorithm: Leave-One-Out cross validation, which uses all but one image from the dataset to train the classifier, and recognition using a reduced training set, which uses only one model per class to train the classifier. The results for 1-NN were 18.42% and 17.9% respectively. These results were lower than expected, and this was attributed to the fact that the input models were extremely noisy. If more accurate training models are used then the classifier accuracy may see significant improvement. The small difference in the recognition rates between the two experiments are indicative of the superiority of object-centred 3D representation systems.

The results obtained in this chapter are compared with the results of the two benchmark algorithms. These algorithms are described in Chapter 8 and the results of the experiments are presented in Chapter 9.

CHAPTER 8

Algorithms for Face Recognition

8.1 Introduction

This chapter details the face recognition algorithms used for the 2D, $2\frac{1}{2}$D and the composite images. Again, two algorithms are investigated - Turk and Pentland's Principal Component Analysis (PCA) based Eigenfaces technique (Turk & Pentland 1991*a*) and a simpler version of Spies and Ricketts' Fourier space nearest-neighbours technique (Spies & Ricketts 2000). The original Eigenfaces technique is often used as benchmark and has been combined with many other approaches and applied to face recognition (del Solar & Navarrete 2005). The second algorithm is based on an extremely simple idea and is reported by its authors as being more accurate than the Eigenfaces algorithm. In this work, both these algorithms are investigated and their performance is compared with the 3D matching algorithm presented in Chapter 7.

Sections 8.2 and 8.3 detail the Eigenfaces technique and the nearest-neighbours in the Fourier space algorithm. The chapter concludes with a brief summary in Section 8.4.

8.2 Principal Component Analysis and Eigenfaces

Principal Component Analysis (PCA) is a dimensionality reduction technique based on extracting the desired number of *principal components* (or "feature axes") from the multi-dimensional data. These feature axes are ordered according to the proportion of variance they explain in the set of faces analysed (training set). PCA is closely related to Karhunen-Loève Transform (KLT) (Loève 1955). It is shown in (Gerbrands 1981) that under the assumption of zero-mean, the formulations of PCA and KLT are identical.

8.2.1 Calculating Eigenfaces

The mathematical details here are taken almost entirely from Turk and Pentland's original papers (Turk & Pentland 1991*a*,*b*). Let a face image $I(x, y)$ be a 2D ($N \times N$) array of intensity values. Let there be M such images. An image can also be considered as an N^2-dimensional vector, or equivalently, as a point in the N^2-dimensional space. Since all face images belong

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

(f) 6 (g) 7 (h) 8 (i) 9 (j) 10

Figure 8.1: The first 10 Eigenfaces for the Sheffield Dataset. These Eigenfaces highlight the discriminatory power of the technique as it captures the different features of the dataset. Particularly noticeable is the capture of pose information in the third Eigenface. Lack of distinct features such as the eyes and the lips is indicative of the input images not having been normalised to align facial features across the dataset.

to the same class of images, they will not be randomly distributed in this space, and can be described by a relatively low dimensional subspace. PCA aims to find those vectors that best account for the distribution of face images within this image space. These vectors define the subspace of face images, or the *face space*. It transpires that these vectors are the eigenvectors of the covariance matrix corresponding to the original face images. And they have a face-like appearance, hence, *Eigenface*. Figure 8.1 depicts the first 10 Eigenfaces for the Sheffield Dataset.

- Let the Training set of face images be:

$$\{\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \dots\} \quad (N^2 \times M) \tag{8.1}$$

- Mean face of the set:

$$\mathbf{\Psi} = \frac{1}{M} \sum_{n=1}^{M} \mathbf{\Gamma}_n \quad (N^2 \times 1) \tag{8.2}$$

- Mean subtracted face image:

$$\mathbf{\Phi}_i = \mathbf{\Gamma}_i - \mathbf{\Psi} \quad (N^2 \times 1) \tag{8.3}$$

  $\mathbf{\Phi}_i$ represents how much each face differs from the mean

- Form the Covariance matrix:

$$\mathbf{C} = \frac{1}{M} \sum_{n=1}^{M} \mathbf{\Phi}_n \mathbf{\Phi}_n^T \tag{8.4}$$

which is equivalent to:

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T \quad (N^2 \times N^2) \tag{8.5}$$

where

$$\mathbf{A} = [\Phi_1, \Phi_2, \ldots \Phi_M] \quad (N^2 \times M) \tag{8.6}$$

The covariance matrix $\mathbf{C}$ is subject to PCA, and yields a set of M orthonormal vectors $\mathbf{u}_n$ and their associated eigenvalues λ_k. These give the best description of the distribution of the images in the face space. $\mathbf{C}$ is a matrix of size $(N^2 \times N^2)$, and finding the eigenvalues and eigenvectors of this is computationally very intensive.

If $(M \ll N^2)$, then the Eigensystem of the smaller $(M \times M)$ matrix $\mathbf{A}^\top\mathbf{A}$ can be found. This means that

$$\mathbf{A}^\top\mathbf{A}\mathbf{v}_i = \mu_i\mathbf{v}_i, \tag{8.7}$$

with eigenvalue μ_i associated with eigenvector $\mathbf{v}_i$. Pre-multiplying both sides by $\mathbf{A}$ gives

$$\mathbf{A}\mathbf{A}^\top\mathbf{A}\mathbf{v}_i = \mu_i\mathbf{A}\mathbf{v}_i \tag{8.8}$$

From this it can be seen that $\mathbf{A}\mathbf{v}_i$ are the eigenvectors of $\mathbf{C} = \mathbf{A}\mathbf{A}^\top$. Based on this, an $(M \times M)$ matrix $\mathbf{L}$ can be constructed.

$$\mathbf{L} = \mathbf{A}^\top\mathbf{A} \quad \text{where} \quad \mathbf{L}_{mn} = \Phi_m^\top\Phi_n \tag{8.9}$$

The eigenvectors $\mathbf{v}_i$ of $\mathbf{L}$ determine linear combinations of the M training set face images to form the Eigenfaces $\mathbf{u}_l$:

$$\mathbf{u}_l = \sum_{k=1}^{M} \mathbf{v}_{lk}\Phi_k, \quad l = 1, \ldots, M \tag{8.10}$$

Thus, the calculations are greatly reduced from the order of the number of pixels in the images (N^2) to the order of the number of images in the training set (M). The associated eigenvalues allow the eigenvectors to be ranked according to their usefulness in characterising the variation among the images. The Eigenfaces $\mathbf{u}_l$ span the basis set with which the images in the face space can be described.

PCA may also be implemented via Singular Value Decomposition (SVD) (Shakhnarovic & Moghaddam 2004). The SVD of an $(N^2 \times M)$ matrix $\mathbf{A}$ (8.6) is given by

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{8.11}$$

where $(N^2 \times M)$ matrix $\mathbf{U}$ and the $(M \times M)$ matrix $\mathbf{V}$ have orthonormal columns, and the $M \times M$ diagonal matrix $\mathbf{D}$ contains the singular values of $\mathbf{A}$ on its main diagonal. A singular value of a matrix $\mathbf{A}$ is the square root of an eigenvalue of $\mathbf{A}\mathbf{A}^\top$. It can be shown that the columns of $\mathbf{U}$ are the orthonormal basis vectors $\mathbf{u}_l$ (equation 8.10), so that SVD allows efficient and robust computation of PCA without the need to estimate the data covariance matrix (equation 8.4). When the number of training images M is much smaller than the size of the image vector (N^2), this is a crucial computational advantage.

The Eigenfaces span an M'-dimensional subspace of the original N^2 image space. The M' significant eigenvectors of the matrix L are chosen as those with the largest associated eigenvalues.

Swets and Weng (Swets & Weng 1996*b*) suggest the following to determine the number of significant eigenvectors to use. Rank the eigenvalues μ_i in non-increasing order. The residual

mean-square error ϵ of using $M' < M$ eigenvectors is the sum of the eigenvalues not used, $\sum_{i=M'+1}^{M} \mu_i$. So M' can be chosen such that the sum of these un-used eigenvalues is less than some fixed percentage P of the sum of the entire set. Let m satisfy

$$\frac{\sum_{i=M'+1}^{M} \mu_i}{\sum_{i=1}^{M} \mu_i} < P \tag{8.12}$$

If $P < 5\%$, a good reduction in the number of eigenvectors is achieved while still retaining a large proportion of the variance present in the original feature vectors (Turk & Pentland 1991*b*, Jain & Dubes 1988). Hence, very little of the original population-capturing power is lost. In practise, the choice of the number of Eigenfaces to use is also guided by computational constraints related to the cost of matching, the size of the training set, etc.

When a new face Γ is to be recognised, it is first transformed into its Eigenface components (projected into the "face space"):

$$\omega_k = \mathbf{u}_k^T(\Gamma - \Psi), \quad k = 1, \ldots, M' \tag{8.13}$$

The weights ω_k form a vector $\Omega^T = [\omega_1, \omega_2, \ldots, \omega_{M'}]$ that describes the contribution of each Eigenface in representing the input image face, treating the Eigenfaces as a basis set for face images. This can then be used to establish which, if any, of the pre-determined face classes best describes the face. The simplest way is to find the face class k that minimises the Euclidean distance

$$d_E = \|\Omega - \Omega_k\|^2, \quad \text{where } \Omega_k \text{ is a vector describing the } k^{th} \text{ face class} \tag{8.14}$$

The face classes Ω_l are calculated by averaging the results of the Eigenface representation over a small number of face images of each individual. A face is classified as belonging to class k when the minimum d_E is below some chosen threshold θ_{d_E}. Otherwise the face is classified as unknown.

Creating the vector of weights is equivalent to projecting the original face image onto the low dimensional face space. Consequently, there will be many images that will project onto a given Eigenface vector. This is not necessarily a problem since the distance d_E between the image and the face space is simply the squared distance between the mean adjusted input image, $\Phi = \Gamma - \Psi$ and $\Phi_f = \sum_{i=1}^{M'} \omega_i \mathbf{u}_i$, its projection onto face space:

$$\epsilon^2 = \|\Phi - \Phi_f\|^2 \tag{8.15}$$

8.2.2 Distance Metrics

In the experiments conducted in this work, two distance measures are used: the Euclidean distance d_E given by equation 8.14, or equivalently by:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}, \tag{8.16}$$

and the Mahalanobis distance d_M given by

$$d_M = \sqrt{(x - y)^\top \Sigma^{-1} (x - y)} \tag{8.17}$$

Equations 8.16 and 8.17 are equivalent if the covariance matrix Σ^{-1} is the identity matrix I. In case of the present application, since the distance is computed over uncorrelated data, the covariance matrix Σ^{-1} is a diagonal matrix whose elements are given by the eigenvalues μ_i. Hence equation 8.17 can be simplified as:

$$d_M(x,y) = \sqrt{\sum_{i=1}^{n} \mu_i^{-1}\left(x_i - y_i\right)^2}, \tag{8.18}$$

where x and y are two feature vectors. Subscript i denotes the i^{th} element of these vectors and μ_i is the i^{th} eigenvalue. The Mahalanobis distance is often used in conjunction with the Eigenfaces algorithm because of its superior discriminatory power (Craw et al. 1999). This claim is confirmed by the results obtained in this work. It measures the distance between two feature vectors in terms of standard deviations from the centroids in each class. Its main advantage is that it can represent non-linear boundaries between features. This is particularly useful when the feature space may be too complex to be divided linearly. The Euclidean distance metric is unable to do this.

8.3 Nearest Neighbours in the Fourier Space (Fourier K-NN)

Spies and Ricketts (Spies & Ricketts 2000) propose a technique based on the Fourier spectra of facial images. Similar to the PCA technique, this technique also relies on a global transformation, i.e. every pixel in the image contributes to each value of its spectrum (Spies & Ricketts 2000). The Fourier spectrum is a plot of the energy against spatial frequencies. The spatial frequencies in this case relate to the spatial relations of intensities in the image. Specifically for face images, this translates to distances between areas of particular brightness such as the overall size of the head, or the distance between the eyes (Spies & Ricketts 2000). Low frequency components contribute to the global description of the image, while higher frequencies describe the finer details in the image.

In (Spies & Ricketts 2000), Spies and Ricketts found the higher frequencies less useful for identification of a person, similarly to the human ability to recognise a face from a brief look without the need to focus on the minute details. Consequently, they chose a small number of frequencies from the entire spectrum and report a 98% recognition rate on the Olivetti Research Ltd (ORL) face database [1]. They also find the real part of the spectra to be more useful for the identification than the imaginary part. In (Zhang 2003) Zhang et. al. also describe the role of spatial frequency in the recognition task similarly. However, they report the high frequency components of images to be of equal importance in aiding the recognition task. In light of this, all the frequency components from the face images are retained in this work. Furthermore, the dataset used in this work is more challenging than the ORL dataset, and it was felt that the retained information would serve to improve the performance of the system.

Spies and Ricketts were able to further reduce their search dimension as all the images they use are real images. Consequently, both the real and the imaginary spectra are symmetric around the origin. This means that only half of each spectra carries valuable information. They use this observation and the variance of the frequencies to select the frequencies that

[1] www.cam-orl.co.uk/facedatabase.html

vary the most. They report good results when comparing their approach with other competing techniques such as Eigenfaces and neural networks on the same dataset. For this work, it was felt that this is not a viable option. Although the above is true for real images (see Figure 8.2), this is not necessarily the case for "complex" images. A real image is defined here as an image matrix whose elements are real numbers, while complex images are image matrices whose elements have both real and imaginary components. In this work complex images are generated during the image matching process. Magarey's image matching algorithm represents the disparity at each pixel as a complex variable. The real and imaginary parts of the complex number represent the horizontal and vertical disparities respectively. Hence the disparity field, or the $2\frac{1}{2}$D image is expressed as a complex-valued matrix.

In complex images, while majority of the information is still concentrated around the centre, there is significant amount of valuable information away from the centre, unlike the Fourier transform of a real image (see figure 8.2(b)).

8.3.1 Fourier Transformation and Recognition

Let the dimensions of an image be $(N \times M)$. There are two frequencies ω_x and ω_y corresponding to the x and the y co-ordinates respectively. If $I_{x,y}$ is the intensity value at location (x, y), then the two-dimensional discrete Fourier transform values f'_{ω_x,ω_y} are given by:

$$f'_{\omega x,\omega y} = \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} I_{x,y} e^{-2\pi j\left(\frac{x\omega_x}{N}+\frac{y\omega_y}{M}\right)} \tag{8.19}$$

Note that the actual implementation of the Fourier Transform is achieved using the Fast Fourier Transform (FFT) function in Matlab.

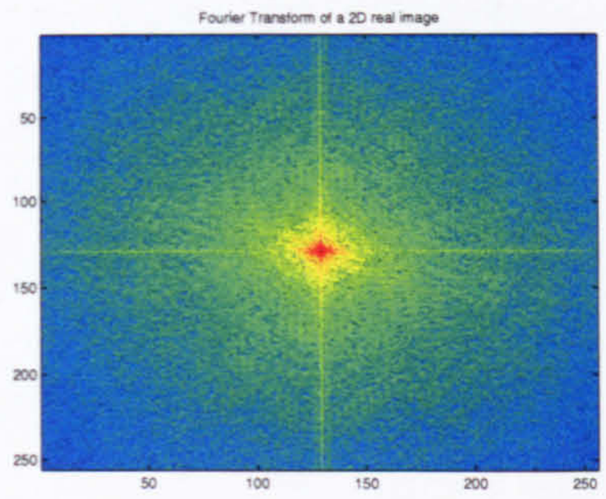
A nearest-neighbour approach is taken to perform the classification and the Frobenius norm is used to give a distance metric between the Fourier transforms of the input images. The Frobenius norm of a matrix A of size $(M \times N)$ is defined as the square root of the sum of the absolute squares of its elements (Golub & Van-Loan 1996):

$$\| \mathbf{A} \|_F = \sqrt{\sum_{i=0}^{N} \sum_{j=0}^{M} |a_{ij}|^2} \tag{8.20}$$

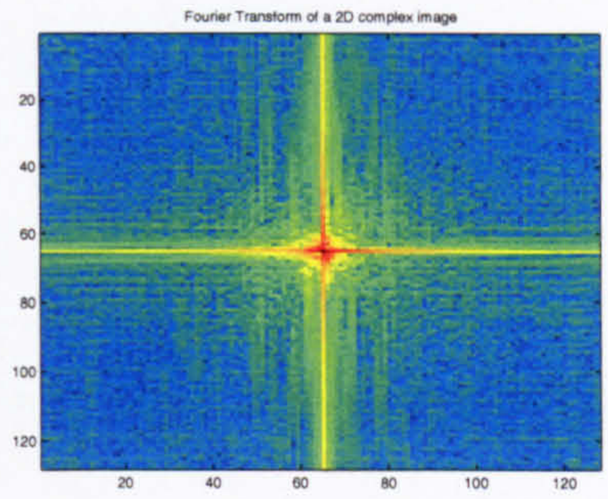
a_{ij} in this case is the difference between the Fourier transforms of the test and the training images.

8.4 Summary

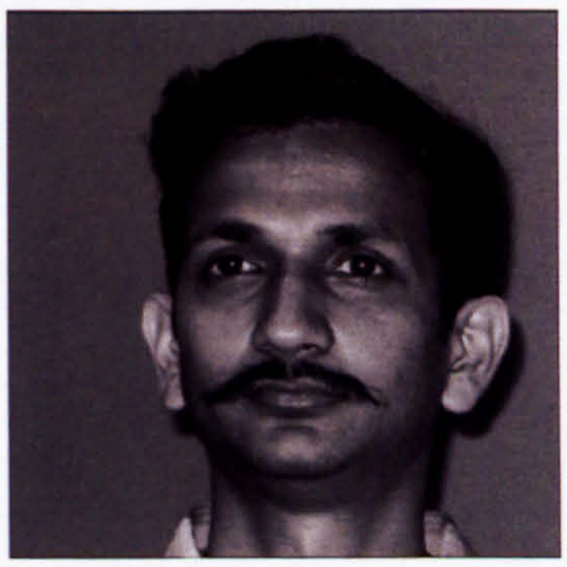
This chapter has detailed the two algorithms used in this work for face recognition. The Eigenfaces algorithm is a well-known algorithm and has been used as a benchmark algorithm in the field since it was pioneered in 1991. The other algorithm is a simple implementation of nearest-neighbour approach in the Fourier space. The three recognition algorithms (spin-image matching, Eigenfaces and Fourier K-nearest neighbour (K-NN)) can now be tested using the Sheffield Dataset. The results of the experiments are presented in the next chapter.



(a)



(b)



(c)

Figure 8.2: Log of absolute values of the Fourier transform of: **(a)** a real 2D image, showing most of the information concentrated around the centre, and **(b)** a complex 2D image. Unlike the Fourier transform of a 2D real-valued face image, there is significant amount of information away from the centre in the Fourier transform of the complex-valued image of the same face. Consequently, selecting a limited number of frequencies to represent such images may result in valuable information being lost. The 2D intensity image corresponding to both Fourier transforms is depicted in **(c)**.

CHAPTER 9

2D Face Recognition: Results and Analysis

9.1 Introduction

Results of face recognition experiments in the 2D space using the benchmark algorithms are presented in this Chapter. These are compared with the results of 3D face recognition using Johnson's (Johnson 1997) spin-image representation from Chapter 7. Turk and Pentland's (Turk & Pentland 1991*a*) PCA based Eigenfaces approach, and Spies and Ricketts' (Spies & Ricketts 2000) Fourier space based nearest-neighbours approach (Fourier K-NN), both described in Chapter 8 are used as benchmarks.

Eigenfaces is a well established technique and is frequently used as a benchmark algorithm. Since the spin-images representation employs a sophisticated nearest-neighbours approach to perform the recognition task, it is more meaningful to compare it with a nearest-neighbours type algorithm in the 2D and the $2\frac{1}{2}$D spaces as well. As in Chapter 7, the first nearest-neighbour (denoted by 1-NN) and the recognition rate of the first 5 matches (denoted by Top5-NN) are analysed and compared with the results in Chapter 7.

9.2 A Note about Eigenfaces

As stated in Section 8.2, the number of Eigenfaces to use in the recognition task is determined using the residual mean-square error measure (equation 8.12) (Swets & Weng 1996*b*). The number of Eigenfaces M' to be used is chosen such that the error measure $\epsilon_{M'}$ is less than 5%. ϵ_m is defined as:

$$\epsilon_{M'} = \frac{\sum_{i=M'+1}^{M} \mu_i}{\sum_{i=1}^{M} \mu_i} \tag{9.1}$$

The residual mean-square error is the sum of eigenvalues not used, $\sum_{i=M'+1}^{M} \mu_i$. M is the total number of eigenvalues.

9.3 Leave-One-Out Cross-Validation on D_1

Leave-One-Out cross-validation is performed on all the 2D images from the Left stereo channel in dataset D_1, consisting of 692 images of 22 individuals (11 males and 11 females). Note that as before, during cross-validation and the computation of the recognition rates, images of an individual with glasses and without glasses are treated as belonging to two different classes. It is believed that this gives a better idea of the discriminatory power of the recognition algorithms. The recognition rate is defined as the number of correctly identified individuals divided by the total number of images in the dataset.

LOO cross-validation is performed on D_1 using both Eigenfaces and Fourier K-NN algorithm. Figure 9.1 shows a plot of the eigenvalues for this dataset. They have been normalised to lie between 0 and 1. Here onwards, the term eigenvalues will refer to normalised eigenvalues.

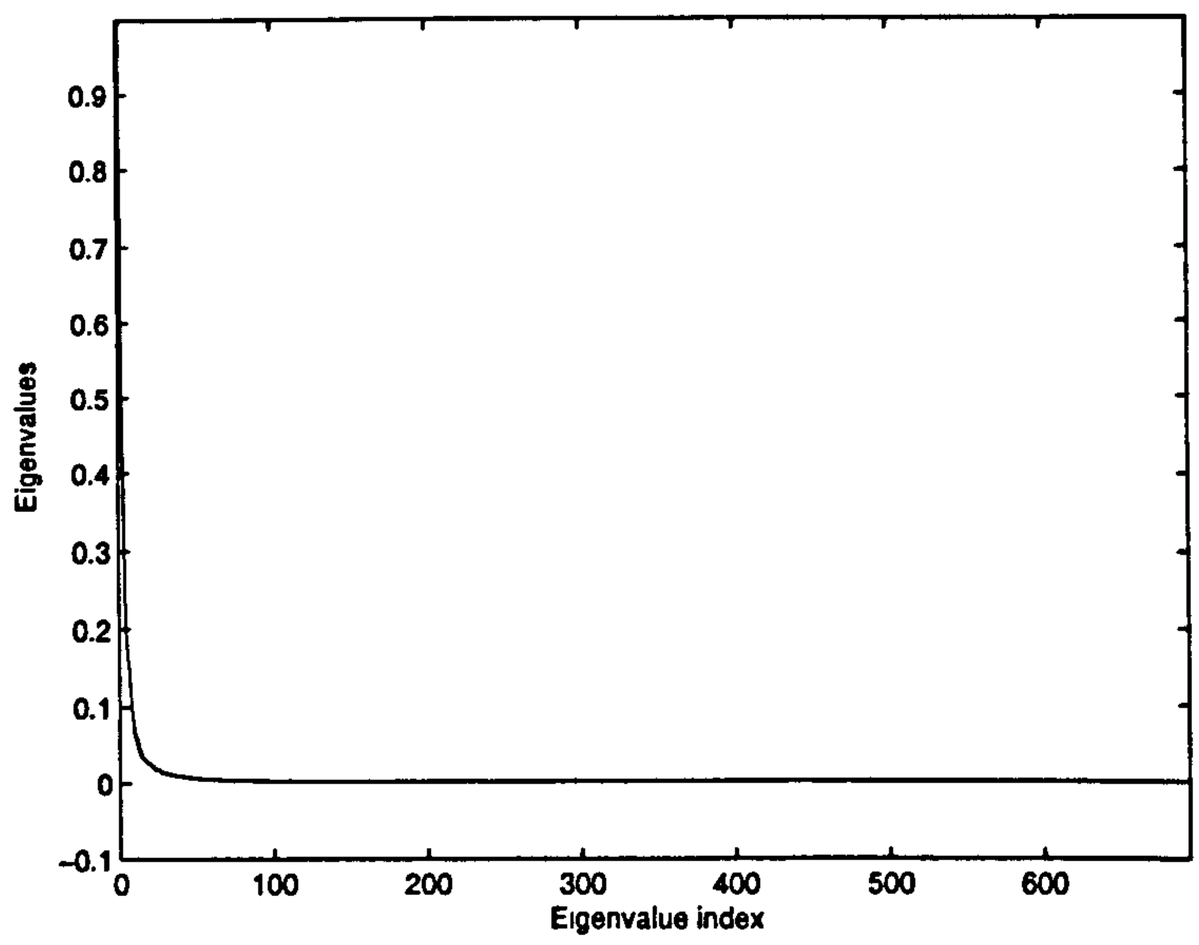
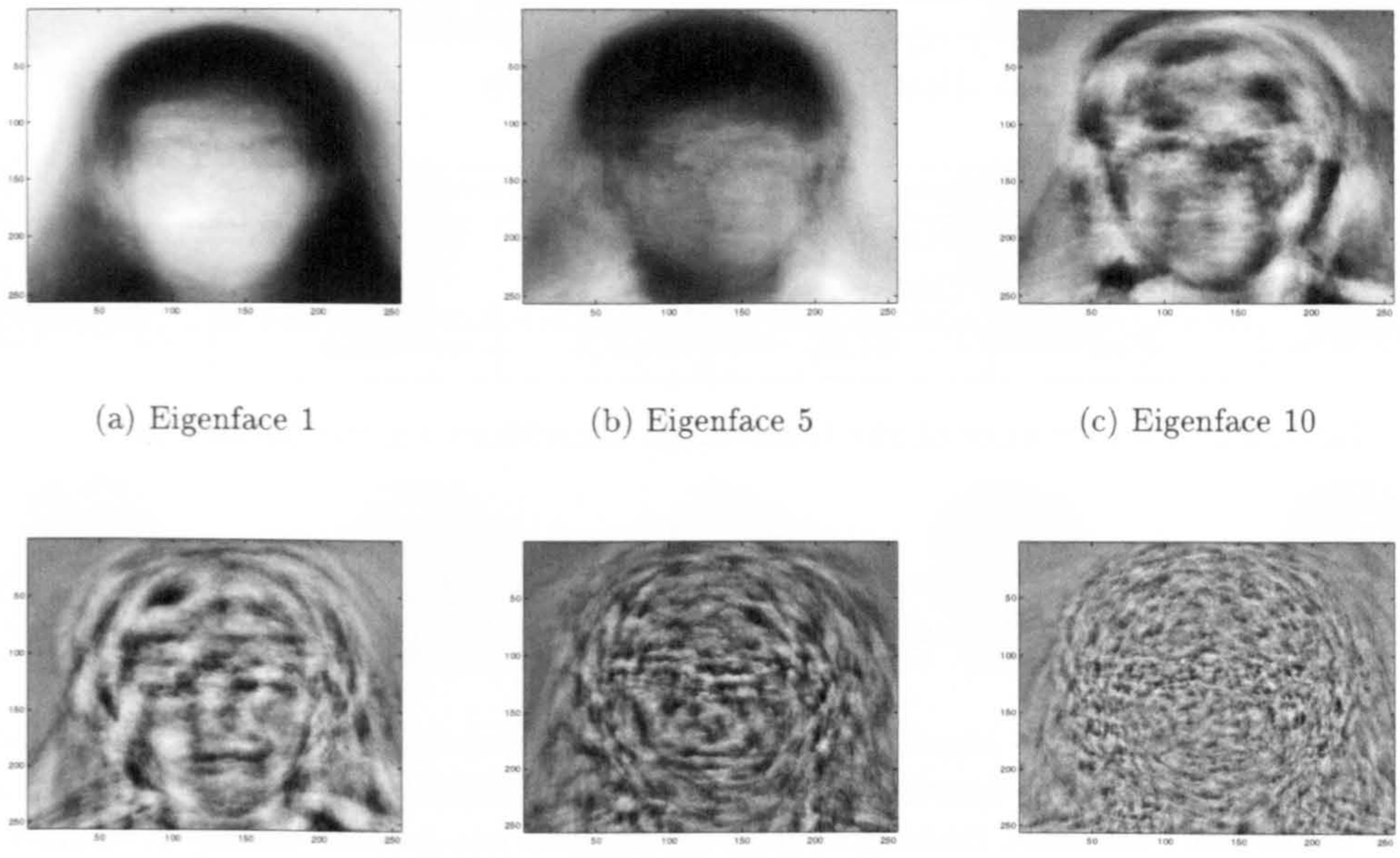


Figure 9.1: Eigenvalues of 2D images in dataset D_1.

Eigenvalues are indicative of the amount of variation in the corresponding eigenvectors, and hence the Eigenfaces. Proportionally, more of this variance is retained in the first few Eigenfaces than in the subsequent ones, and it is this variance that allows the faces to be identified accurately. This is illustrated in Figure 9.2, which compares the set of Eigenfaces $\{1, 5, 10, 100, 250, 500\}$. This observation is confirmed by the inverse ι_i, of the error values listed in Table 9.1.

9.3.1 Recognition Rates

The first 250 Eigenfaces are used for recognition, giving an error measure ϵ_{250} of 4.38%. The corresponding recognition rate is 55.2% for the Euclidean distance metric and 67.05% for



(a) Eigenface 1 (b) Eigenface 5 (c) Eigenface 10

(d) Eigenface 100 (e) Eigenface 250 (f) Eigenface 500

Figure 9.2: Eigenfaces $\{1, 5, 10, 100, 250, 500\}$ of 2D images of D_1.

Eigenface Index i	1	5	10	100	250	500
Inverse error-measure ι_i (%)	24.46	4.09	1.49	0.07	0.02	0.008
Error Measure ϵ_i (%)	75.54	45.58	33.47	10.36	4.38	1.02

Table 9.1: Error measure values and their inverses for the Eigenfaces $\{1, 5, 10, 100, 250, 500\}$, illustrated in Figure 9.2.

the Mahalanobis distance metric. This clearly highlights the superiority of the Mahalanobis distance measure in classifying faces in the Eigenface space.

By contrast, the recognition rate for the Fourier K-NN algorithm was 68.9% for the 1-NN and 86.6% for the Top5-NN[1]. This is in agreement with Spies and Ricketts' findings. They too report a better performance for the Fourier K-NN compared with the Eigenfaces in (Spies & Ricketts 2000). They report recognition rates of 98% and 94% for Fourier K-NN and Eigenfaces (with Euclidean distance metric) respectively.

Eigenfaces		Fourier K-NN	
Euclidean Distance (%)	Mahalanobis Distance (%)	1-NN (%)	Top5-NN (%)
55.2	67.05	68.9	86.6

Table 9.2: Performance of the benchmark algorithms on the dataset D_1.

These benchmark results are much lower than those reported by Spies and Ricketts (Spies & Ricketts 2000). This inconsistency in the recognition rates could be due to two reasons:

1. **Challenging dataset:** Most 2D face recognition algorithms reported in the literature use publicly available datasets. If this is not the case then the data they use is usually very controlled. Sheffield Dataset on the other hand has greater variety of images per class, and imposes very few constraints on subjects' pose and expression. Hence, it is possible that on a simple dataset, the algorithms would give a better performance.

2. **Presence of "bad" images in dataset:** There may be some image classes in the dataset that may be placed in the face-space in such a way that a high proportion of individuals are assigned to that class.

Both possibilities were investigated in order to establish the cause(s) for such poor results. The difficulty level of the dataset was investigated by running the code for the Fourier K-NN and Eigenfaces on a publicly available dataset. The Yale Database[2] containing 165 images (15 subjects, 11 images per subject) was chosen for this. The main considerations behind this choice were the size of the database and the variety of images available for each subject. The database is small enough for the code to run fast and has an acceptable mix of image types. The dataset is somewhat limited in terms of the individuals in the database (e.g. there is only one female subject). Each subject is imaged both with and without glasses. There are five images with varying expressions: happy, sad, sleepy, surprised and wink. Three images are captured by moving the light source to the left, the right and the centre of the face, and one image is neutral (no expression, no particular illumination direction). Figure 9.3 shows the different subjects in the Yale Database and 9.4 shows the images for one of the classes in the database.

[1]Top5-NN refers to the recognition rate when the correct match is identified among the top five matches, rather than the conventional 5-nearest-neighbours definition

[2]http://cvc.yale.edu/projects/yalefaces/yalefaces.html

Figure 9.3: Subjects in the Yale Database.

Figure 9.4: Subset of images from the Yale Database.

9.3.2 LOO cross-validation on the Yale Database

The first 60 Eigenfaces of the Yale Database are used, giving an error measure ϵ_{60} of 3.77%. Figure 9.5 shows the first 10 Eigenfaces from the Yale database. Combination of strict controls being imposed during image capture and the relatively small size of the database means that the features such as the eyes in these Eigenfaces are far more pronounced than say in Figure 8.1. The features in Sheffield dataset Eigenfaces (Figure 8.1) are not very discernible at all.

Figure 9.5: The first 10 Eigenfaces of the Yale Database.

Table 9.3 presents the results of recognition using the two benchmark algorithms in the Yale Database. In (Belhumeur et al. 1997), Belhumeur et. al. report similar results for the Eigenfaces technique applied to the Yale Database.

Eigenfaces		Fourier K-NN	
Euclidean Distance (%)	Mahalanobis Distance (%)	1-NN (%)	Top5-NN (%)
79.3	98.8	80.4	84.9

Table 9.3: Performance of the benchmark algorithms on the Yale Database.

This experiment with the Yale Database also highlighted some issues regarding test and training data. It is obvious that D_1 is much bigger and far more varied than the Yale Database, both in terms of the image types and the subjects. The Yale Database is very restrictive in that all its images are strictly controlled, and hence not very realistic. Such datasets result in good accuracy for algorithms that may not be suitable for operation in uncontrolled environments (such as crowd surveillance or identification). It highlights the point that recognition rates reported in the literature should always be considered in the context of the test data used, and if possible, the same dataset should be used when comparing algorithms.

9.3.3 Confusion Matrices

If any one class, α, in the image database is more prone to confusion and miscalssification, then it is possible that a lot of images may be wrongly classified as belonging to that class. This can decrease the recognition rate for the entire classifier. It does not necessarily mean that the classifier is poor, just that the data or a certain subset of the data is poor.

In order to detect the presence of such image classes in the dataset, *confusion matrices* were used. Confusion matrices are used for checking the accuracy of classifiers. For n classes in the dataset, an $(n \times n)$ zero matrix $\mathbf{C}$ is initialised. The rows of $\mathbf{C}$ depict the actual class labels and the columns depict the class labels assigned by the classifier. The elements of this matrix act as bins. The bins count the number of class i images that are classified as belonging to class $1, 2, \ldots, n$. A graphical representation of this matrix is quick and an easy visual tool for identifying any classes that may be responsible for the classifier's poor performance. Figure 9.6 shows the confusion matrix for a perfect classifier.

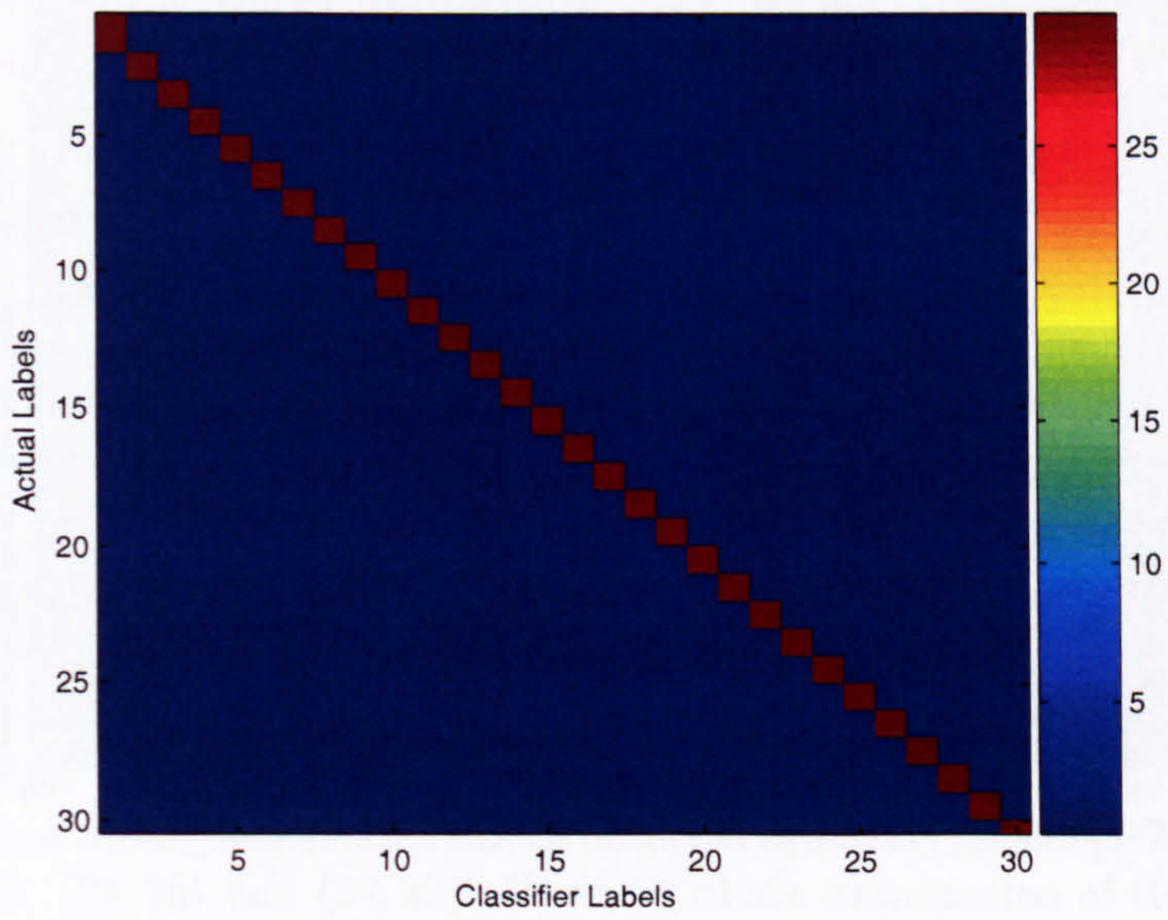


Figure 9.6: Confusion Matrix for a perfect classifier is a diagonal matrix as all images are classified correctly.

The confusion matrix is diagonal, i.e. all class i images are correctly classified as belonging to class i. The off-diagonal entries in the matrix are incorrectly classified individuals, which in the case of a perfect classifier are all 0. Figures 9.7 and 9.8 show the confusion matrices for the LOO cross-validation on D_1 using Eigenfaces and Fourier K-NN respectively.

It is evident that the classifiers are not perfect classifiers, as the matrices are not strictly diagonal. However, the distribution of the misclassifications also appears to be random. It is not possible to identify any particular class of images as such, that may be causing the classifiers to perform badly. The confusion matrices also emphasise the fact that the Mahalanobis distance measure is more powerful than the Euclidean when used in conjunction

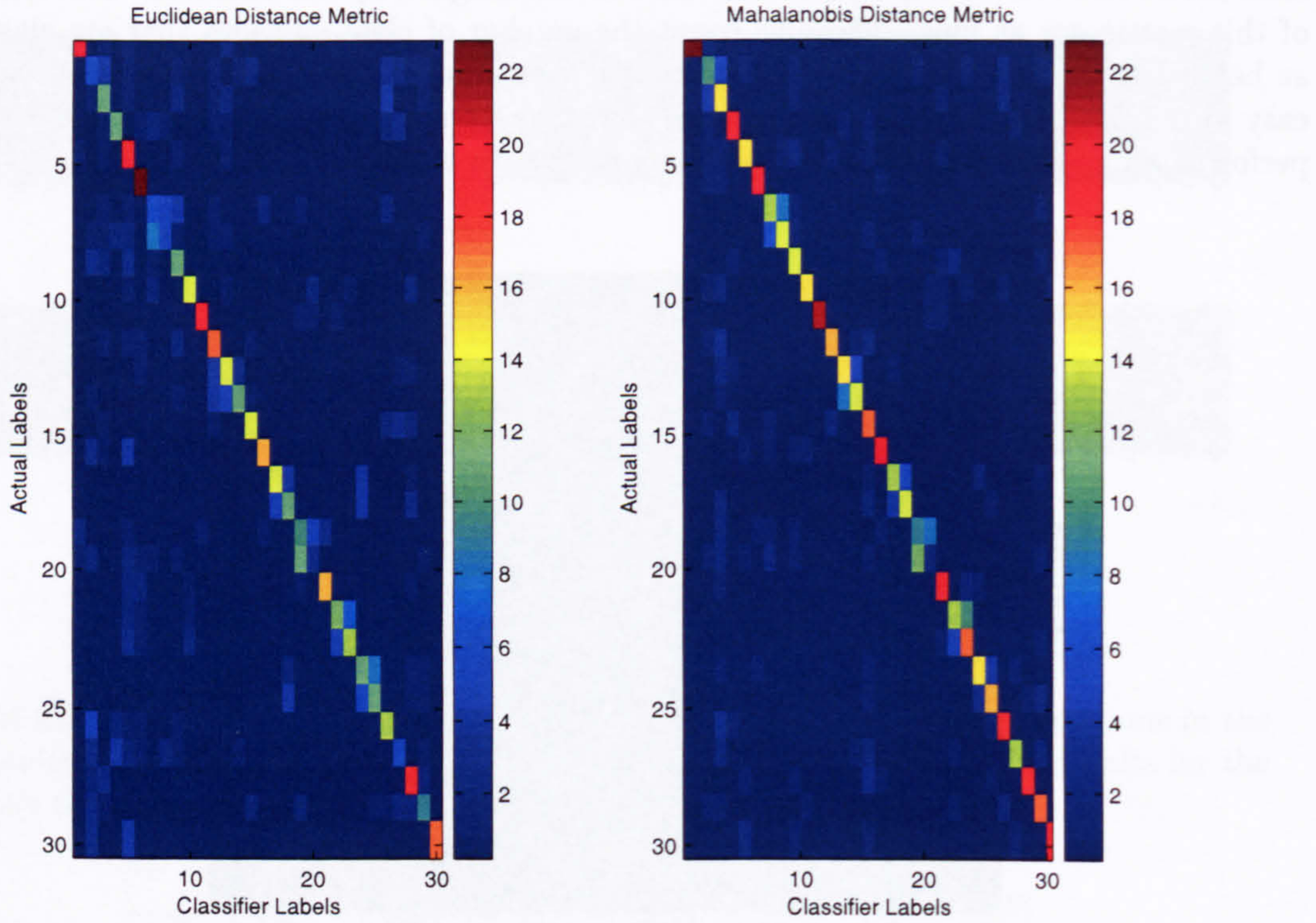


Figure 9.7: Confusion Matrices for the Eigenfaces classifier, using both the Euclidean and the Mahalanobis distance metric. Although the matrices are not diagonal, there is no emerging pattern. Hence, there are no classes in the dataset that are particularly "bad", indicating a random distribution of errors.

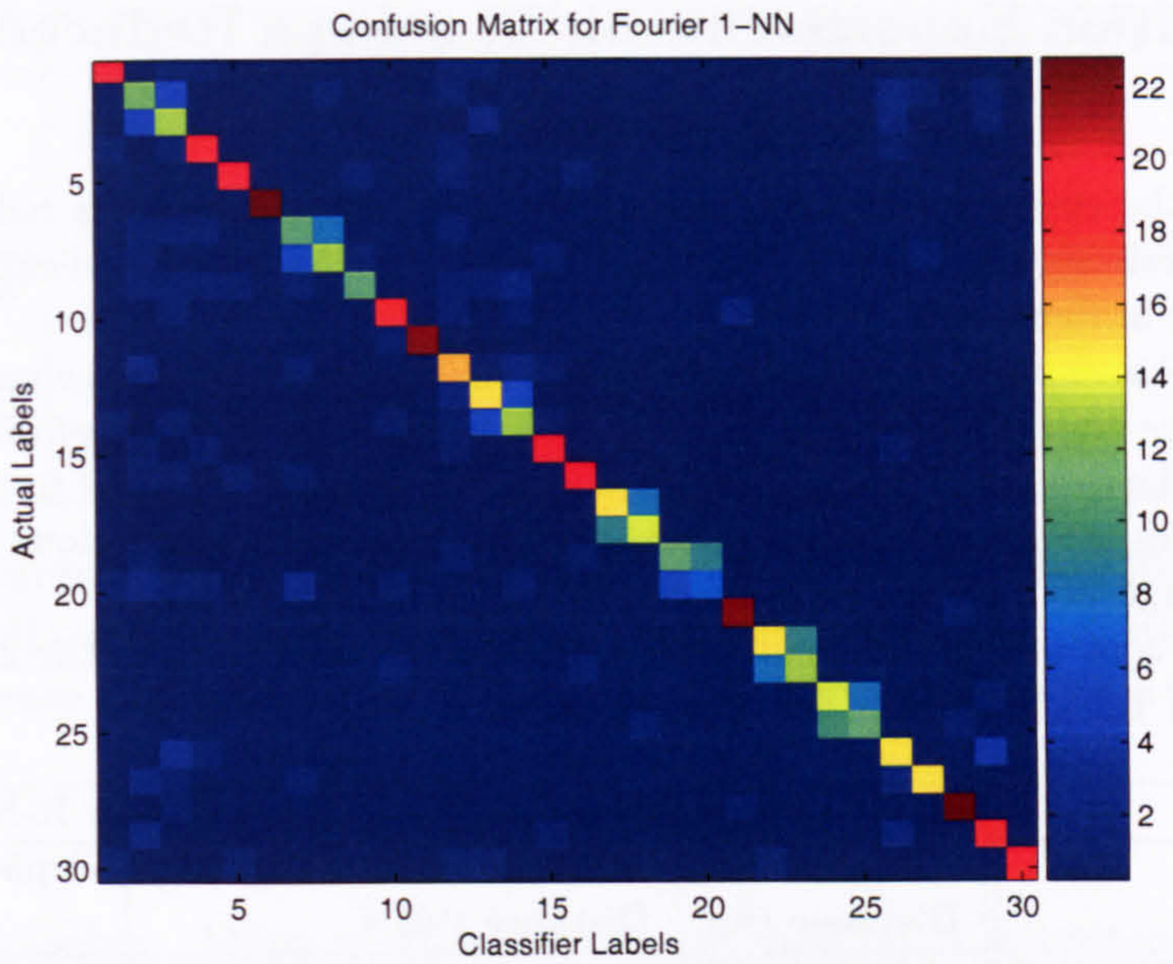


Figure 9.8: Confusion Matrix for the Fourier 1-NN classifier. Again, the matrix is not entirely diagonal, but there are fewer values off diagonal than in Figure 9.7.

with Eigenfaces. Its confusion matrix is relatively more diagonal than that of the Euclidean measure. This is depicted by the greater number of squares on the diagonal with shades of red, and relatively fewer squares off diagonal with a value greater than 4. Similarly, Figure 9.8 shows that Fourier K-NN is a better classifier than the Eigenfaces.

Also noticeable is that certain elements in the confusion matrices appear to be more erroneous than others. This is depicted by sub and super-diagonal elements with relatively high number of misclassified elements. This can be observed around elements $\{2,3\}, \{7,8\}, \{13,14\}$, $\{17,18\}, \{19,20\}$, $\{22,23\}$ and $\{24,25\}$. However, closer examination of the dataset reveals that this is in fact due to the nature of the dataset. The elements in the above list represent adjacent classes of images where one class is an individual with glasses and the other is the same individual without glasses. Classes 22-25 contain images of the same individual with and without glasses and head-scarf. But notice how classes $\{22,23\}$ ($\{24,25\}$) are only confused between themselves and not with $\{24,25\}$ ($\{22,23\}$). The classifiers treat subjects 22 and 23 (without heads-scarf) as the same, and subjects 24 and 25 as the same (with head-scarf), but $\{22,23\}$ and $\{24,25\}$ as different from each other. The classifiers are able to cope with accessories such as glasses but not head-scarves, which significantly alter the appearance of the face.

Hence it can intuitively be seen that more "misclassifications" are possible when these image classes are treated as being two distinct individuals. If however, they are treated as the same individual and the classifiers are trained appropriately, then the confusion matrices would have a very pronounced diagonal with fewer erroneous values distributed randomly off-diagonal, as expected.

9.4 Recognition Experiments on D_1 using a Reduced Training Set

As in Chapter 7, the classifiers are also tested on the dataset D_1 using a reduced training set. The reduced training set consists of the full frontal images with neutral expressions only. These are depicted in figure 4.2.

In Chapter 9, this experiment was conducted to see how the spin-image representation's recognition rates are affected when the training set for each class consists of only one image. It was found that the drop in recognition rates was only marginal. This led to the conclusion that the spin-images are a powerful representation and recognition technique when dealing with 3D data, though not necessarily for faces.

Here, the same is investigated using the two benchmark algorithms in the 2D space. The results are given in Table 9.4.

	Eigenfaces		Fourier K-NN	
	Euclidean Distance (%)	Mahalanobis Distance (%)	1-NN (%)	Top5-NN (%)
LOO cross-validation	55.2	67.05	68.9	86.6
Reduced Training Set	46.24	48.7	45.09	73.41

Table 9.4: Performance of the benchmark algorithms trained using almost the entire dataset (LOO cross-validation) and using the reduced training set and tested on the dataset D_1.

There is not a significant difference between the recognition rates achieved using the Euclidean and Mahalanobis distance metrics. This indicates that although adding a greater variety of images to the training set does improve the performance of the PCA classifier, it may also make the distribution of the classes in the face space more complex. This increased complexity justifies the need for a more sophisticated distance metric such as the Mahalanobis. Also, the Fourier K-NN algorithm performs worse than the Eigenfaces algorithm. This is a clear indication that this is a powerful technique, but as with all nearest-neighbours type algorithms, it does require more training data to optimise its performance.

The following section compares the results of recognition experiments on the dataset D_1 in the 2D space using the benchmark algorithms with the results obtained in Chapter 7.

9.5 Comparing Face Recognition in 2D and 3D spaces

The results of the recognition experiments using the spin-image representation in the 3D space, and using Eigenfaces and Fourier K-NN in the 2D space are shown in Table 9.5.

It was clear in Chapter 7 that the recognition rates in the 3D space were extremely poor and certainly much lower than expected. However, it was not expected that the 2D algorithms would out-perform the 3D algorithms by such a vast margin. The poor results for the spin-image representation were thought to be a consequence of all the errors that were accumulated in the input data prior to the recognition stage (see Chapter 7 for details). These errors are accumulated during the reconstruction process (due to errors in camera calibration, image matching and triangulation processes), the mesh generation process and

	Eigenfaces		Fourier K-NN		Spin-Images	
	Euclidean Distance (%)	Mahalanobis Distance (%)	1-NN (%)	Top5-NN (%)	1-NN (%)	Top5-NN (%)
LOO Cross-Validation	55.2	67.05	68.9	86.6	18.42	32.75
Reduced Training Set	46.24	48.70	45.09	73.41	17.90	25.33

Table 9.5: Comparison of the recognition rates in the 2D space (using Eigenfaces and Fourier K-NN) and the 3D space (using Spin Image representation). Recognition rates for the two experiments - Leave-One-Out Cross-Validation and Recognition using the reduced training set are shown.

in mesh pre-processing. Many of these errors are difficult to eliminate. But it is equally difficult to obtain good recognition rates in the presence of these errors.

Hence, after weighing various pros and cons (see Section 7.5 for details), it was concluded that 3D face recognition with the spin-images was not suitable when using stereo images. This is not to say that it is not a viable technique for face recognition, just that it is neither particularly practical nor accurate when the data is extremely noisy. It would be interesting to investigate in future the performance of this technique when the 3D data is more accurate. These results also offered no conclusive evidence regarding the usefulness of depth information in face recognition.

Finally, the results in Table 9.5 show that the classifier performs better in LOO cross-validation experiments than in the experiments when the classifier has fewer training images. This is to be expected since as a general rule, the amount of training data is directly proportional to the classifier accuracy, processing time and the storage requirements. In addition, there is also the danger of over-training the classifier so that it is unable to handle input data that deviates significantly from the training data.

9.6 Summary

The results of face recognition experiments in the 2D space are presented in this chapter. As in Chapter 7, experiments are performed using Leave-One-Out (LOO) cross-validation and also by training the classifiers on just a single image from each of the classes (reduced training set). The recognition rates are better for the LOO cross-validation experiment. The 2D recognition rates, although better than the 3D, are still very disappointing compared to those reported in the literature. LOO cross-validation experiments are performed on a publicly available dataset, the Yale Database, and results are compared with published results. This ascertains that Sheffield Dataset is more challenging than some of the publicly available datasets, and low accuracy of the benchmark algorithms can be attributed to this. Confusion matrices are used to establish that the miscalssifications are more likely between images of subjects with and without glasses. For other subjects, there are no strong patterns in the misclassifications.

CHAPTER 10

Face Recognition Using $2\frac{1}{2}$D and Composite Images

10.1 Introduction

$2\frac{1}{2}$D images are defined as 2D images that encode depth information. In literature, $2\frac{1}{2}$D images are generally constructed using the actual depth values instead of the intensity values in the 2D images. This approach implicitly assumes that extracted depth values are accurate. However, when depth values are extracted from stereo images, this is rarely the case, unless accuracy in image matching, camera calibration and triangulation can be guaranteed. In most real settings, cameras are susceptible to perturbations, which can result in incorrect camera matrices, and hence depth values. This is one of the main reasons for the unpopularity of stereo images for face recognition. In this chapter, an alternative encoding of depth information is presented for $2\frac{1}{2}$D images. Also presented is a simple and effective way of combining this depth information along with the texture or the intensity information from the 2D images to form *composite images.*

Face recognition experiments are conducted on both $2\frac{1}{2}$D and composite images. The results are compared with those of 2D face recognition. Using the same dataset and classification algorithms in all the spaces makes the comparison of results more meaningful. Again, Eigenfaces and the Fourier space based nearest neighbours algorithm (Fourier K-NN) are used for classifying the faces. Results and analysis are presented, followed by conclusions and a summary of the chapter.

10.2 $2\frac{1}{2}$D Image Face Recognition

Constructing $2\frac{1}{2}$D images using depth values (depth maps) is a simple and popular means of encoding relief information in a 2D image, as all the existing image processing and 2D face recognition techniques can be used. However, if these depth values are obtained through stereoscopy, then the existing problems with noise in camera matrices, correspondences and triangulation remain. Obviously, for accurate recognition results, the depth information

needs to be as error-free as possible.

Inspired by human binocular vision, this work uses disparity information as proxy for depth. Disparity values contain depth information, proportional up to parameters of the camera matrices. The camera matrices only serve to give the placement of the object being reconstructed in the 3D space, relative to other objects, including the cameras themselves. So, the actual depth information is contained in the disparity values, which in theory at least, should be as accurate as the depth values. Using the disparity values bypasses the noisy camera calibration and the triangulation processes, while still retaining all the depth information. Only the errors from the image matching process influence the disparity values.

In this work, the $2\frac{1}{2}$D images encode the depth information in the form of *disparity* values rather than actual *depth* values. For each pair of 2D images that are matched, two $2\frac{1}{2}$D images are produced. These are denoted by LR and RL. LR (RL) represents the $2\frac{1}{2}$D image that is generated by holding the left (right) image as the reference image and looking for the corresponding matches in the right (left) image, i.e. LR (RL) is the left-right (right-left) disparity.

As explained before, in his work, Magarey (Magarey 1997) represents the disparity values at each pixel by a complex number. The real and imaginary parts correspond to disparity in the horizontal and vertical directions respectively. 2D intensity images only have a real intensity value at each pixel location, and similarly depth maps only have a real depth value at each pixel location. Ways of extracting meaningful information from the pair of disparity values (real for horizontal and imaginary for vertical disparity) are also investigated.

Five simple representations of the complex number are investigated:

1. **Complex:** Disparity values are represented as complex numbers as this retains the displacement in the horizontal and vertical directions as such. It allows the actual depth values to be extracted, given the camera matrices, but has the disadvantage of the image not being in a form that is easy to visualise.

2. **Real:** Only the horizontal disparity values are used and the $2\frac{1}{2}$D image can be easily visualised.

3. **Imaginary:** Similarly, only the vertical disparity values are used and again, the $2\frac{1}{2}$D image can be easily visualised.

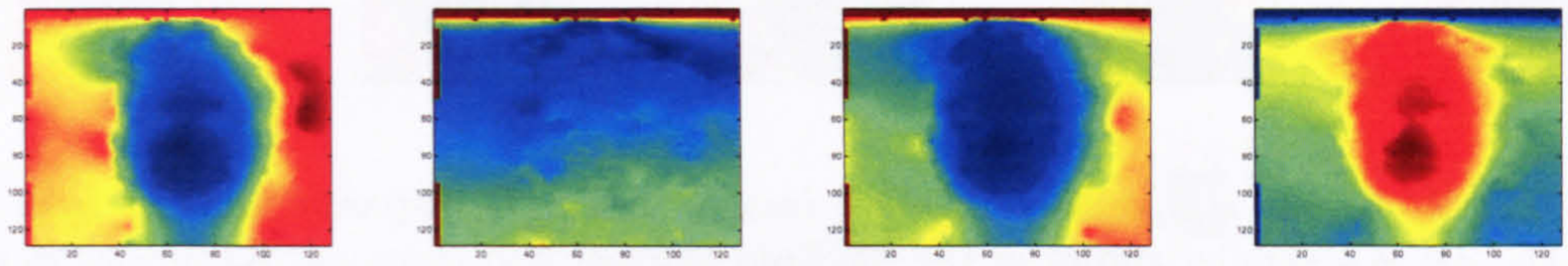
4. **L_1-Norm:** This is a simple way of combining the information in the real and the imaginary components. The L_1-norm of a complex number $c = a + ib$ is computed using

$$L_1 = |a| + |b| \tag{10.1}$$

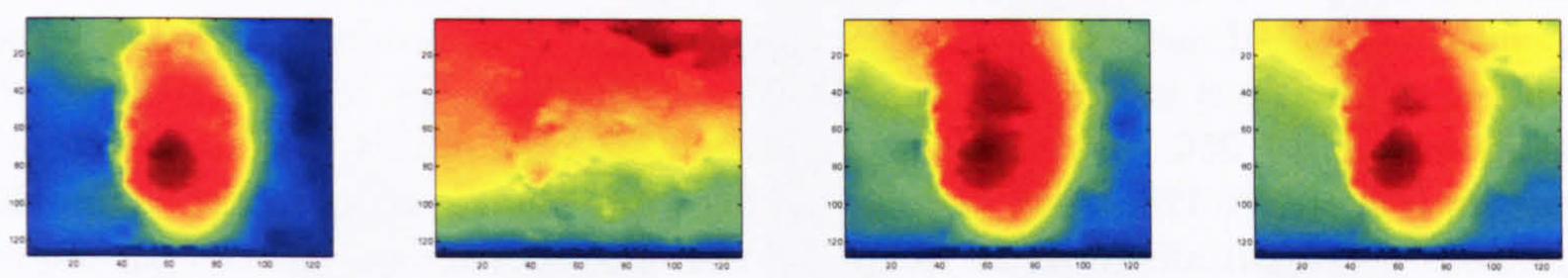
5. **L_2-Norm:** This also combines disparity information in both the directions, and is commonly used to represent the magnitude of a complex number. The L_2-norm of a complex number $c = a + ib$ is computed using

$$L_2 = \sqrt{(a^2 + b^2)} \tag{10.2}$$

Although the L_1 and the L_2 norms encode all the disparity information, its components in the horizontal and the vertical directions are lost. Figure 10.1 shows the $2\frac{1}{2}$D images using the representations 2-4.



(a) Real LR (b) Imaginary LR (c) L_1 LR (d) L_2 LR



(e) Real RL (f) Imaginary RL (g) L_1 RL (h) L_2 RL

(i) 2D intensity image

Figure 10.1: LR and RL $2\frac{1}{2}$D images generated using the four non-complex representations of the disparity information: Real, Imaginary, L_1-norm and the L_2-norm, and the corresponding 2D image.

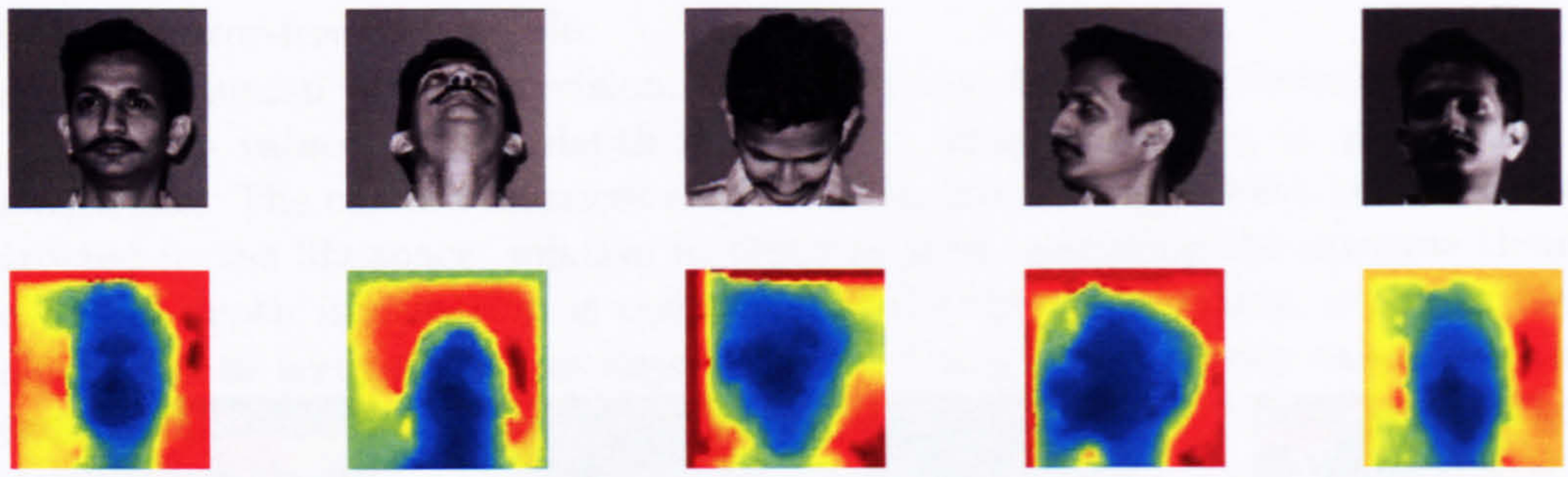


Figure 10.2: A sample image class (2D left images and $2\frac{1}{2}$D disparity images) from the dataset D_2. D_2 is a smaller subset of the Sheffield Dataset and contains 165 images (5 or 6 images per class).

Recall from Chapter 6 that Magarey's algorithm requires the disparity field at each level to be interpolated till it is of the same size as the input images. This produces a disparity vector for each point in the input images. As mentioned in Chapter 4, the images in the Sheffield Database measure 256×256. However, the disparity maps used in this work to represent the $2\frac{1}{2}$D images only measure 128×128. Having smaller input images reduced the computation time significantly without affecting the quality of the images or indeed the recognition rates in pilot experiments. In addition, it was felt that unnecessary interpolation had the potential to introduce unnecessary errors. Hence the disparity maps are not interpolated to achieve full resolution.

10.2.1 Feasibility Study

Initially, a feasibility study was carried out to investigate whether or not using the disparity values as a proxy for depth was effective and if it merited further investigation.

The feasibility study was carried out using dataset D_2, using LOO cross-validation. Only the Eigenfaces algorithm with the Euclidean distance measure was investigated in this study since earlier experiments showed that this classifier performed the worst. If the Eigenfaces algorithm performed well in this study, then it would be reasonable to assume that the Fourier K-NN algorithm would also give acceptable results on the $2\frac{1}{2}$D images.

The results of the feasibility study are presented in Figure 10.3. 100 Eigenfaces are used for both 2D and $2\frac{1}{2}$D images. This corresponds to an error value of ϵ_{100} of 3.73% for the 2D images. The mean error value for the $2\frac{1}{2}$D images is $< 1\%$.

It can be seen from Figure 10.3 that on average, LR and RL images perform similarly. Recall from Chapter 6 that the RMS errors (in the image matching process) for LR and RL images were 41.27 and 28.89 respectively. Then, the absolute mean difference in the recognition rates between LR and RL images of 0.85% is surprising. It would appear that errors in the image matching process are not necessarily passed on to the recognition stage.

The imaginary component of the complex disparity value, i.e. vertical disparity results in the highest classifier accuracy. This is extremely surprising since the imaginary $2\frac{1}{2}$D image (see Figure 10.1) does not appear to contain much information. In fact, it appears very noisy compared to the other images. The most prominent feature in the imaginary $2\frac{1}{2}$D images is

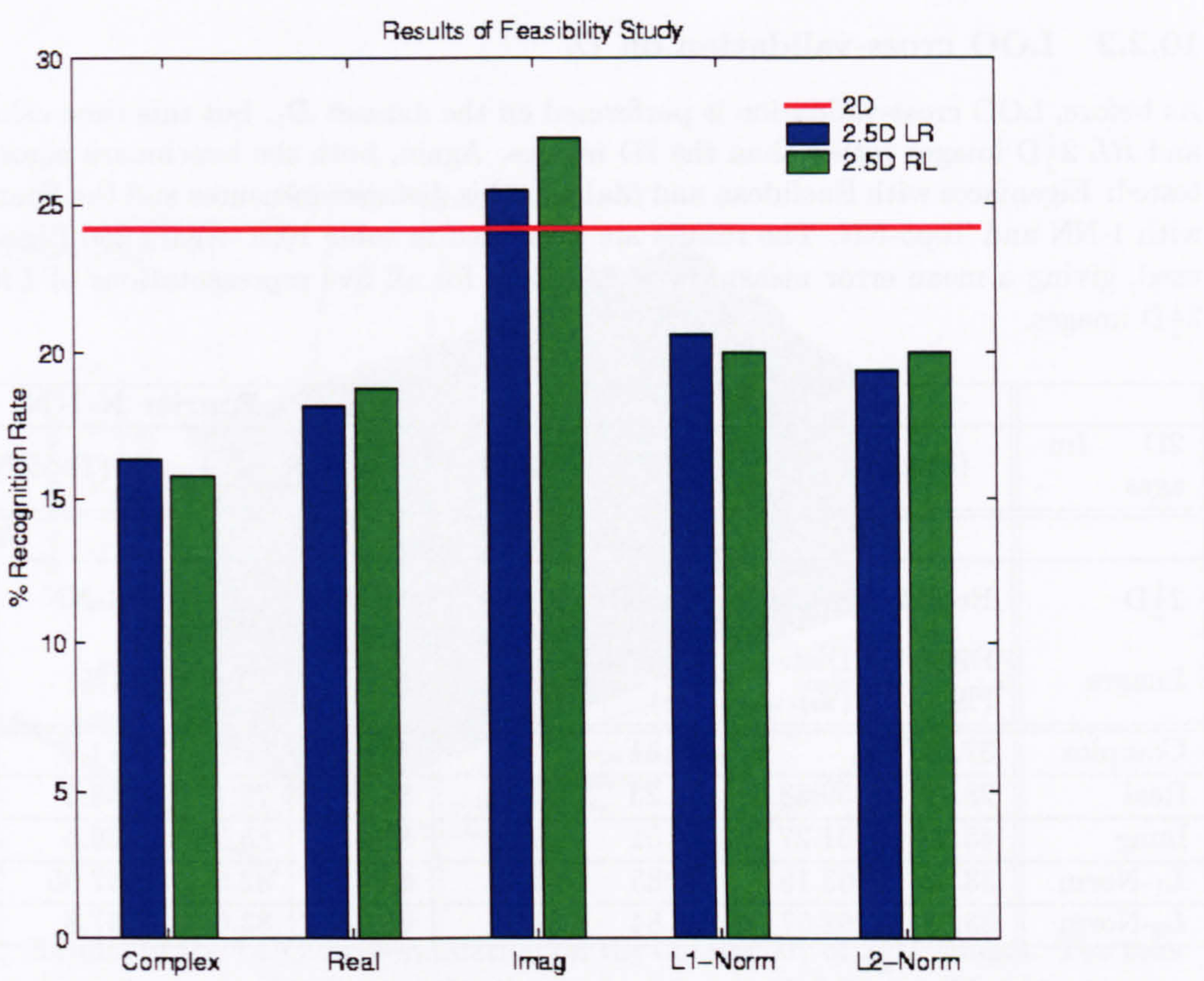


Figure 10.3: Results of the feasibility study conducted to ascertain whether depth can be represented using 2D disparity maps. LOO cross-validation is performed on D_2 using the Eigenfaces algorithm and the Euclidean distance measure. The bars depict the results of the five representations of the $2\frac{1}{2}$D images, both LR and RL. The recognition rate for the corresponding 2D images is shown by the red line.

the eyes, and it would seem that this information is sufficient for classification. Recognition using the imaginary $2\frac{1}{2}$D image outperforms the 2D recognition by up to 3%. This is not significant considering the other forms of $2\frac{1}{2}$D images under-perform by about 5.5%.

The results of the feasibility study were promising, and consequently, LOO cross-validation was performed on the larger dataset D_1. The results are presented in Section 10.2.2. The classifiers are also tested by training them on a reduced dataset, as in Section 9.4. The results for this experiment are presented in Section 10.2.3.

10.2.2 LOO cross-validation on D_1

As before, LOO cross-validation is performed on the dataset D_1, but this time using the *LR* and *RL* $2\frac{1}{2}$D images rather than the 2D images. Again, both the benchmark algorithms are tested: Eigenfaces with Euclidean and Mahalanobis distance measures and the Fourier K-NN with 1-NN and Top5-NN. The results are presented in table 10.1. Again 250 Eigenfaces are used, giving a mean error measure ϵ_{250} of 0.36% for all five representations of *LR* and *RL* $2\frac{1}{2}$D images.

	Eigenfaces				**Fourier K-NN**			
2D Images	(Euclid.) 55.2		(M'bis) 67.05		(1-NN) 68.9		(Top5-NN) 86.6	
	LR		*RL*		*LR*		*RL*	
$2\frac{1}{2}$D Images	Euclid. Dist. (%)	M'bis Dist. (%)	Euclid. Dist. (%)	M'bis Dist. (%)	1-NN (%)	Top5-NN (%)	1-NN (%)	Top5-NN (%)
Complex	37.57		39.31		69.9	86.0	71.4	84.4
Real	28.76	59.68	32.23	61.71	64.5	79.1	68.5	80.8
Imag	45.52	61.27	44.51	59.68	67.6	85.3	66.5	84.7
L_1-Norm	33.53	63.15	36.85	61.42	65.2	82.5	67.05	81.2
L_2-Norm	33.09	62.57	35.84	61.27	65.3	83.0	67.8	82.1

Table 10.1: Comparison of the recognition rates in the $2\frac{1}{2}$D space using Eigenfaces and Fourier K-NN. Recognition rates for all the representations of the $2\frac{1}{2}$D images are shown.

Compare the values in Table 10.1 with those in Table 9.5. It is easy to see that the recognition rates for the $2\frac{1}{2}$D images are considerably lower than those obtained for the 2D images, on the same dataset and using the same algorithms. A graphical depiction of this is presented in Figure 10.4.

Even with the low recognition rates obtained with the Eigenfaces algorithm, the recognition in the 2D space still outperforms recognition in the $2\frac{1}{2}$D space. These results contradict those obtained in the feasibility study, as at least one form of the $2\frac{1}{2}$D images outperformed the 2D image recognition. Admittedly, it was only by a small margin. However, adding more images to the training set was expected to enhance the performance of the classifier rather than deteriorate it. Instead, the recognition rates have fallen by an average of 17.5% across all image representations and distance measures. The drop in the recognition rates achieved

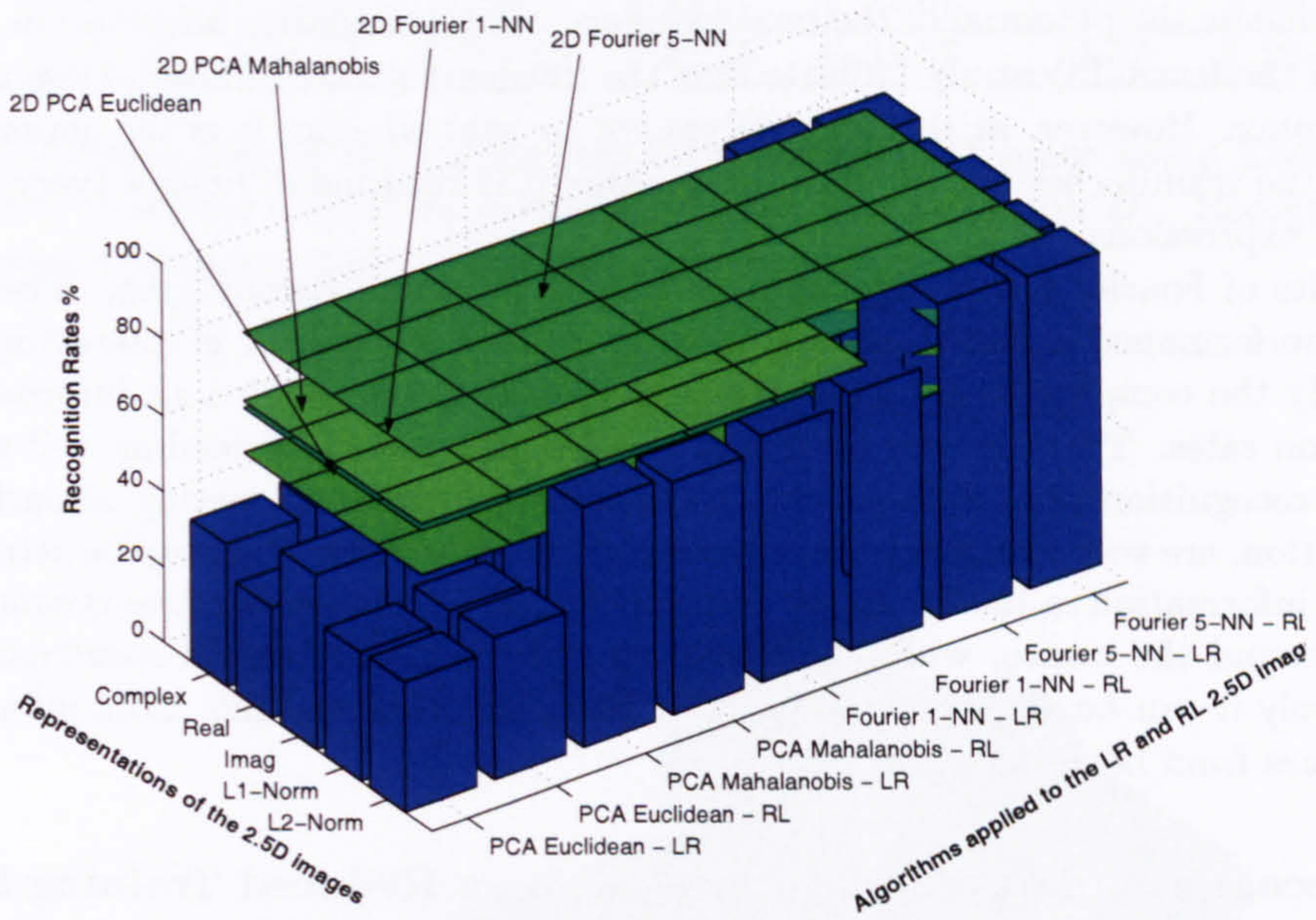


Figure 10.4: Results of the LOO cross-validation on the dataset D_1 of $2\frac{1}{2}$D images. The bars in the chart show the recognition rated for the various forms of LR and RL $2\frac{1}{2}$D images. Superimposed on the bars are four surfaces representing the results of applying the same algorithms to 2D images.

by the Euclidean distance metric is far more pronounced and is mostly responsible for this overall steep decline in performance. The recognition rates for the complex images using the Mahalanobis distance measure are omitted from the table since Mahalanobis distance cannot be defined for complex feature vectors and any accuracy measure defined as such would be meaningless. The remainder of the $2\frac{1}{2}$D images display a much smaller deterioration in the recognition rates - a mean fall of 5.7%. Note that here too the Mahalanobis distance classifier performs better than the Euclidean in both 2D and the $2\frac{1}{2}$D spaces. However, recognition in the 2D space is still more accurate than recognition in the $2\frac{1}{2}$D space.

Although disappointing, these results are not entirely surprising. When one compares the $2\frac{1}{2}$D images with the 2D ones (see Figure 10.1), it is easy to see that in terms of intensity information, the $2\frac{1}{2}$D images are very limited. Their information content is more subtle than the 2D images. In view of this, it is in fact encouraging that recognition rates are as they are, and emphasise the potential of the two classifiers and the disparity information. Further, the results of the feasibility study indicate that the images contained in each class may be of some significance. However, at this stage it cannot be said whether it is the actual number of images in the training set of each class, or whether it is the kind of images (varying poses, illumination, expressions, etc.).

The results of Fourier K-NN algorithm, as before, are more encouraging. Based on the algorithm's performance in the 2D space, this was to be expected. It is interesting to note that it is only the complex form of the disparity map that has yielded an improvement in the recognition rates. The other forms of $2\frac{1}{2}$D images have not performed as well as the 2D images. The recognition rates obtained by the imaginary form, representing disparity in the vertical direction, are very close to those achieved by the 2D images. This can be attributed to the fact that information in the Fourier transforms of real valued face images is concentrated very much around the centre, while in complex images, a larger part of spectrum is used. Also, intuitively it can be seen that the complex representation contains more information - disparity values from both the directions

10.2.3 Recognition Experiments on D_1 using a Reduced Training Set

The results for this experiment are shown in Table 10.2 and Figure 10.5.

Again, the results in the $2\frac{1}{2}$D space do not measure up to those in the 2D space. It is however interesting to note that the imaginary $2\frac{1}{2}$D images perform particularly well in this experiment. They yield the highest recognition rates for the Fourier K-NN classifier and for Eigenfaces with the *LR* $2\frac{1}{2}$D images. This has not been observed previously. The improved performance of the imaginary images indicates that in the absence of sufficient training data, the vertical disparity values hold more classification information than the horizontal or the combined disparity values. However, at this stage it is not possible to say whether or not this would be the case for all datasets. It may be that this phenomenon is peculiar to the Sheffield Dataset and the way it was collected. More research needs to be conducted to ascertain whether this is true of all datasets.

10.3 Wavelets-Based Pre-Processing for Eigenfaces

Wavelets based pre-processing was investigated to see if the performance of the Eigenfaces classifier in the $2\frac{1}{2}$D space can be enhanced. This was motivated by the results of the earlier

	Eigenfaces				**Fourier K-NN**			
2D Images	(Euclid.) 46.24		(M'bis) 48.70		(1-NN) 45.09		(Top5-NN) 73.41	
	LR		*RL*		*LR*		*RL*	
$2\frac{1}{2}$D Images	Euclid. Dist. (%)	M'bis Dist. (%)	Euclid. Dist. (%)	M'bis Dist. (%)	1-NN (%)	Top5-NN (%)	1-NN (%)	Top5-NN (%)
Complex	31.65		31.36		31.94	64.16	31.79	64.86
Real	28.03	38.58	28.90	41.47	28.47	57.51	28.61	57.23
Imag	33.24	38.15	33.53	36.13	34.10	66.62	33.53	66.91
L_1-Norm	30.35	36.42	30.06	40.17	31.07	58.53	29.91	60.69
L_2-Norm	29.19	35.40	30.20	40.75	31.07	58.53	29.91	60.69

Table 10.2: Comparison of the recognition rates in the $2\frac{1}{2}$D space using Eigenfaces and Fourier K-NN classifiers trained on a reduced dataset and tested on D_1. Recognition rates for all the representations of the $2\frac{1}{2}$D images are shown.

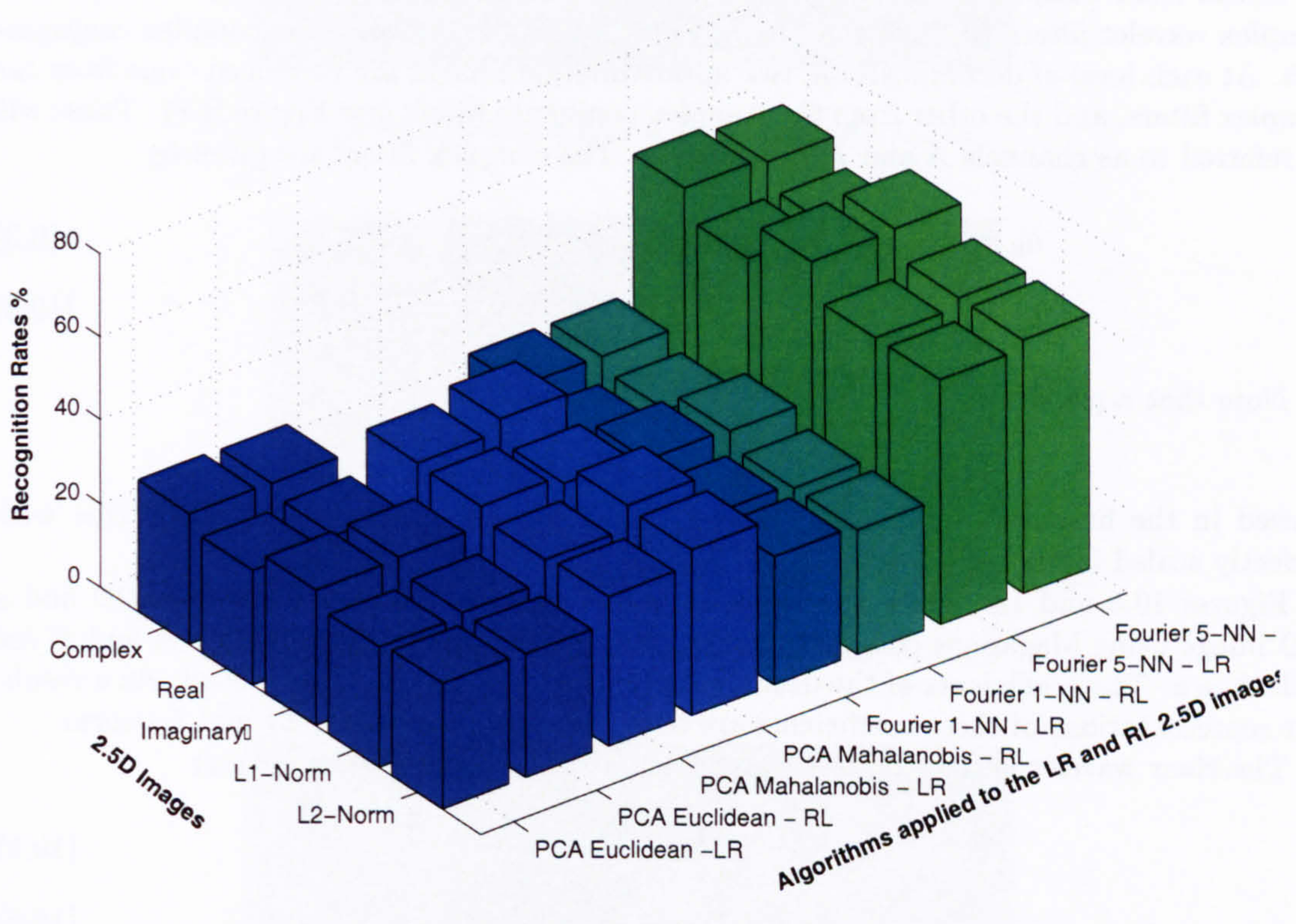


Figure 10.5: Performance of the benchmark algorithms in $2\frac{1}{2}$D space, trained using the reduced dataset and tested on the dataset D_1.

experiments that highlighted the benefits of using Fourier decomposition of the intensity images rather than the images themselves. However, this idea was not explored in detail since it was decided at the outset that pre-processing would be avoided and that the classifiers and the images would be tested in their "raw" form.

The effects of two wavelets were investigated. The Haar or the Daubechies-1 (Db1) wavelet has been used to this end by Chen et. al. in (Chen et al. 2003) and by Chien et. al in (Chien & Wu 2002). They have reported reasonable results for face recognition. Haar wavelets have also been used for image querying by Jacobs et. al in (Jacobs et al. 1995) with promising results. In Chapter 6 Magarey's complex wavelets performed extremely well in matching face images. These wavelets have very good orientation selectivity, which makes them ideal for image processing.

As explained in Appendix A, 2D wavelets decompose the input images into four subimages: a low frequency Approximation image and three high-frequency Detail images that extract the horizontal, vertical and diagonal features from the images. Most of the information useful for recognition is contained in the low-pass approximation images (Magarey 1997, Chien & Wu 2002). Hence, only these images are used for the recognition task. However, a recent publication by Ekenel and Sankur (Ekenel & Sankur 2005) reports that using the horizontal detail images achieves greatest robustness to varying illumination, while the approximation images are the most expression-invariant when Daubechies-4 (Db4) wavelets are used.

Recall from Chapter 6 that Magarey's wavelets are 2D complex wavelets. Two pairs of complex wavelet filters $\{h_0, h_1\}$ and $\{h_0^*, h_1^*\}$ are used and h^* denotes the complex conjugate of h. At each level of decomposition, two approximation images are produced - one from the complex filters, and the other from the complex conjugate filters (see Figure B.1). These will be referred to as channels A and B respectively. The complex filters are given by

$$h_0 = \begin{bmatrix} 1-j & 4-j & 4+j & 1+j \end{bmatrix}/10 \tag{10.3}$$

$$h_1 = \begin{bmatrix} -1-2j & 5+2j & -5+2j & 1-2j \end{bmatrix}/14 \tag{10.4}$$

Note that a pre-filter

$$f = \begin{bmatrix} -j & 5 & j \end{bmatrix}/5$$

is used in the first level of decomposition to simulate an infinitely large DWT tree with perfectly scaled filters.

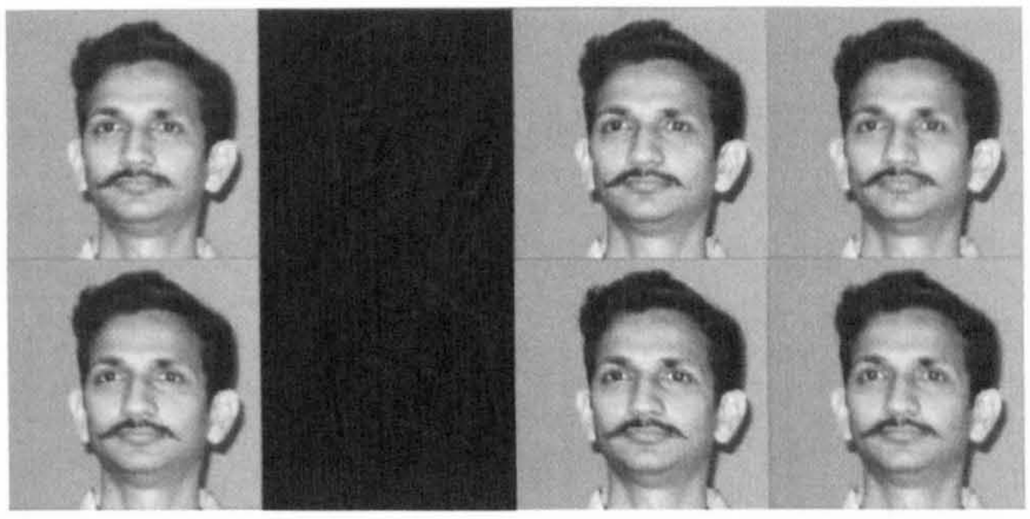
Figures 10.6 and 10.7 show the low-pass or the Approximation images for a 2D and a $2\frac{1}{2}$D image using Magarey's complex wavelet. Note that output from channels A and B are both shown. The coefficients of the decomposed images are also complex valued. As a result, four representations of these coefficients are depicted: real, imaginary, L_1 and L_2-norm.

The Haar wavelet is the simplest wavelet and corresponds to the filter pair

$$h_0 = \begin{bmatrix} 1 & 1 \end{bmatrix}/\sqrt{2} \tag{10.5}$$

$$h_1 = \begin{bmatrix} 1 & -1 \end{bmatrix}/\sqrt{2} \tag{10.6}$$

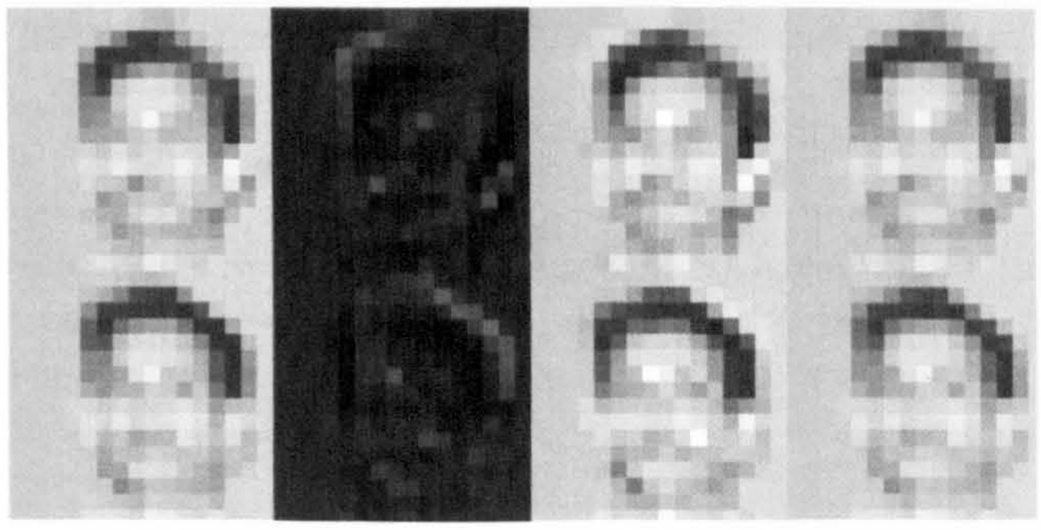
Figures 10.8 and 10.9 show the low-pass or the Approximation images for a 2D and a $2\frac{1}{2}$D image using the Haar wavelet.



(a) Approximation at level 1: Real, Imaginary, L_1, L_2

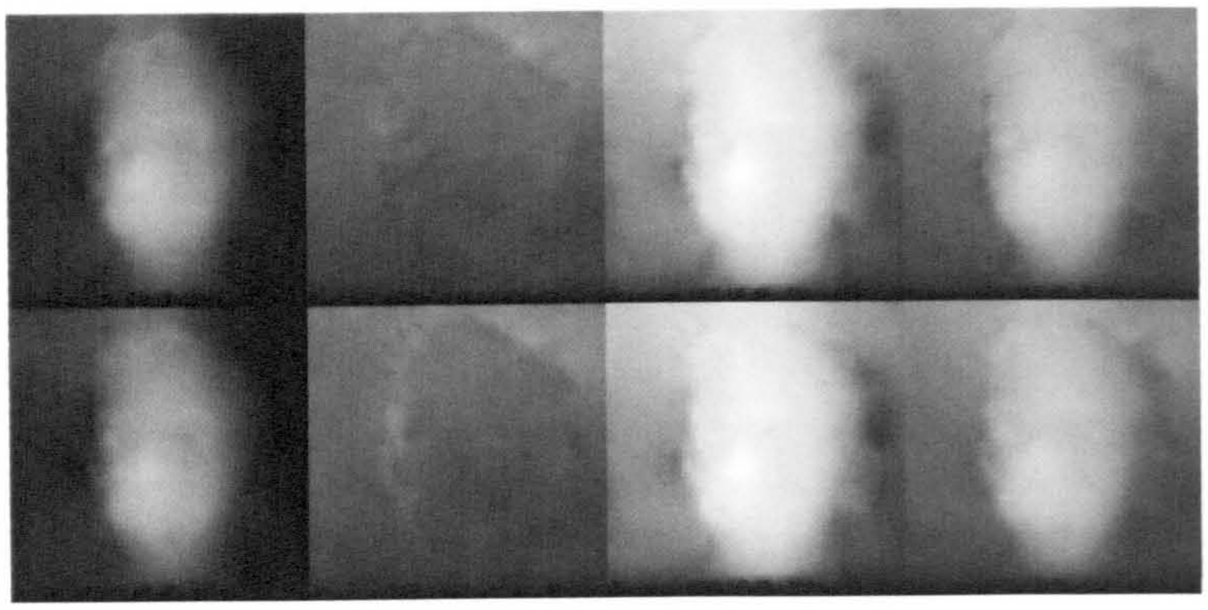
(b) Approximation at level 2: Real, Imaginary, L_1, L_2

(c) Approximation at level 3: Real, Imaginary, L_1, L_2

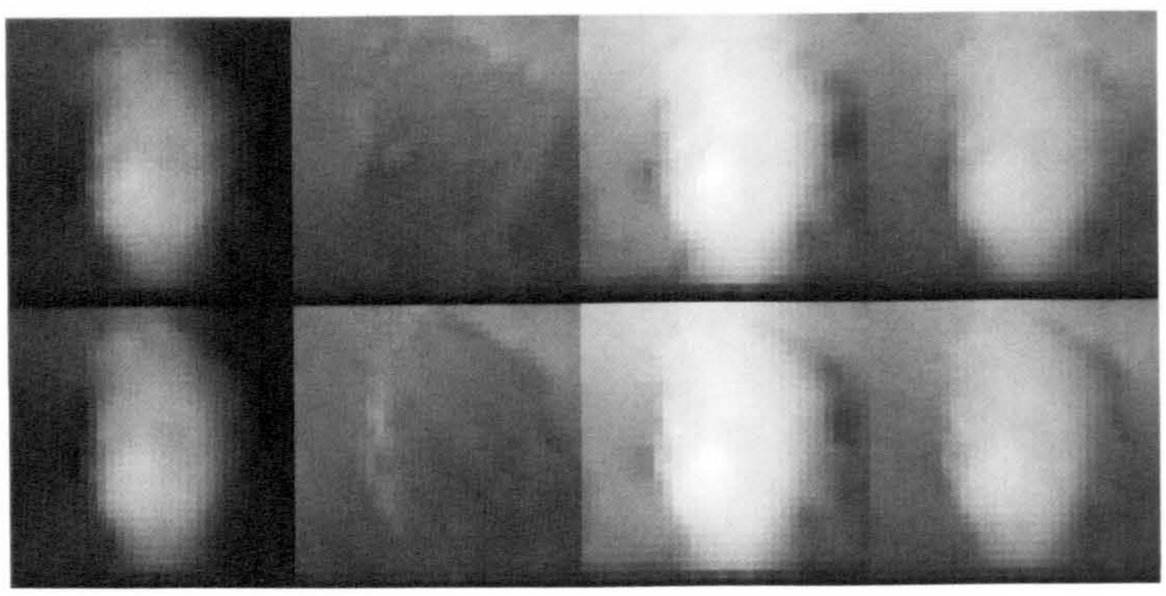


(d) Approximation at level 4: Real, Imaginary, L_1, L_2

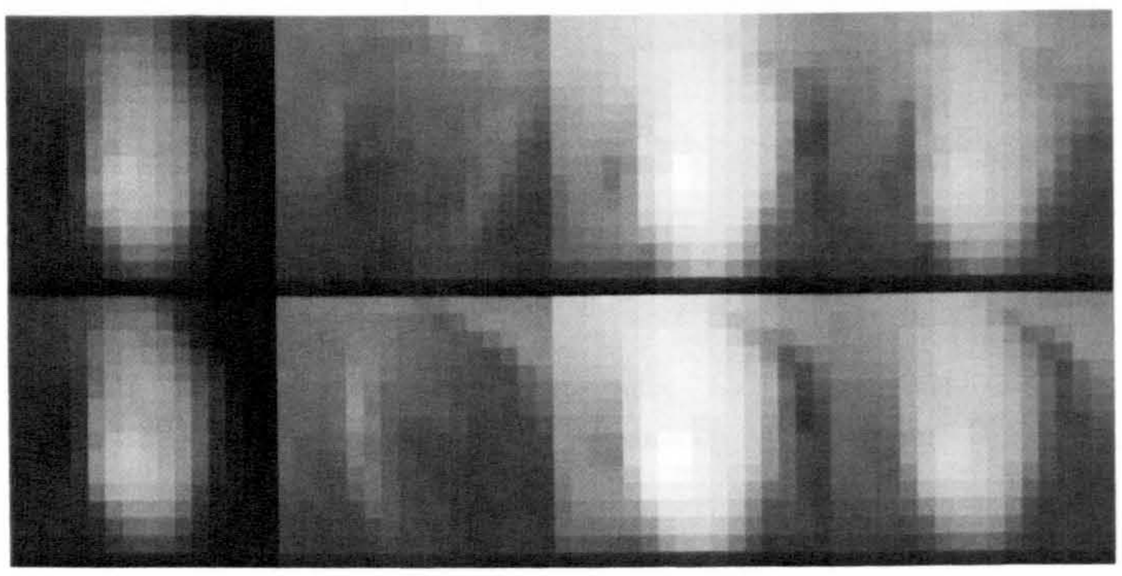
Figure 10.6: The outputs of the 2 low-pass channels, A & B, (Approximation images) for a 2D intensity image that has undergone 4 levels of decomposition using Magarey's complex wavelet. Real, Imaginary, L_1 and L_2-norm representations from both channels are shown.



(a) Approximation at level 1: Real, Imaginary, L_1, L_2

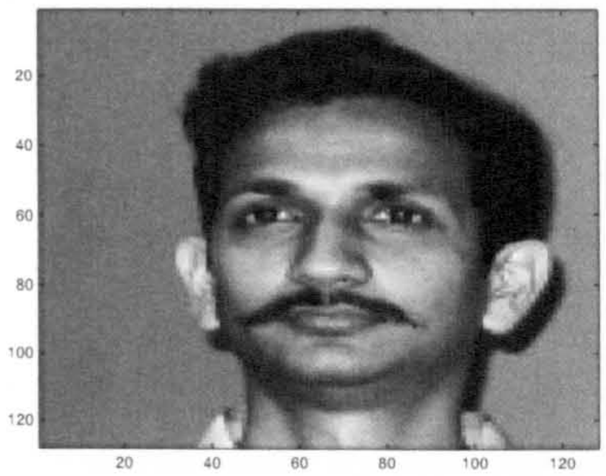


(b) Approximation at level 2: Real, Imaginary, L_1, L_2

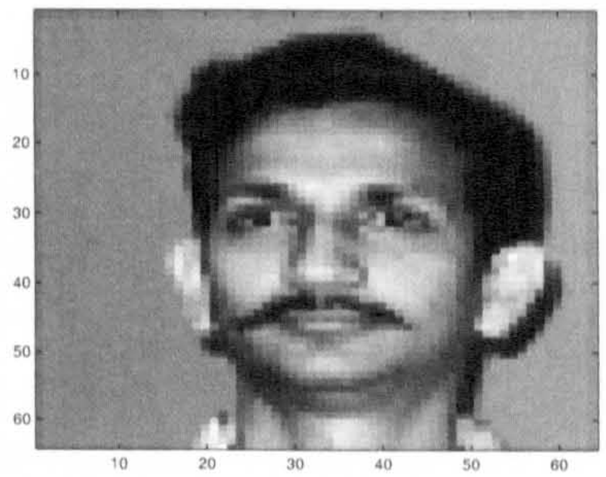


(c) Approximation at level 3: Real, Imaginary, L_1, L_2

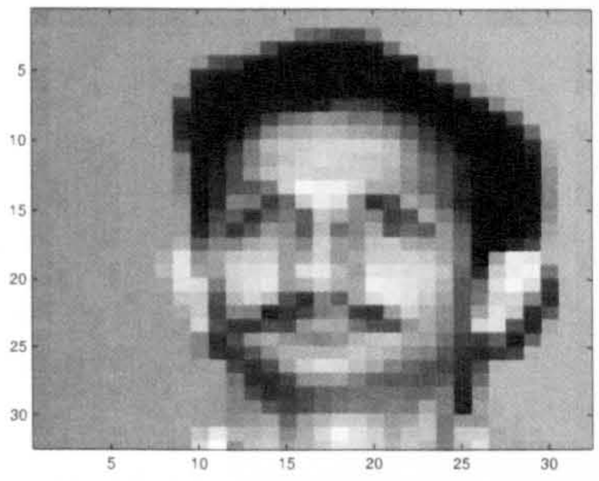
Figure 10.7: The outputs of the 2 low-pass channels, A and B, or the Approximation images, for a $2\frac{1}{2}$D intensity image that has undergone 3 levels of decomposition using Magarey's complex wavelet. Real, Imaginary, L_1 and L_2-norm representations from both channels are shown.



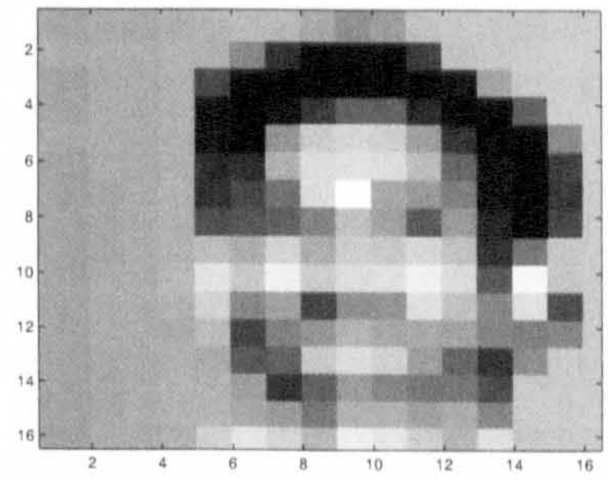
(a) Approximation at level 1



(b) Approximation at level 2

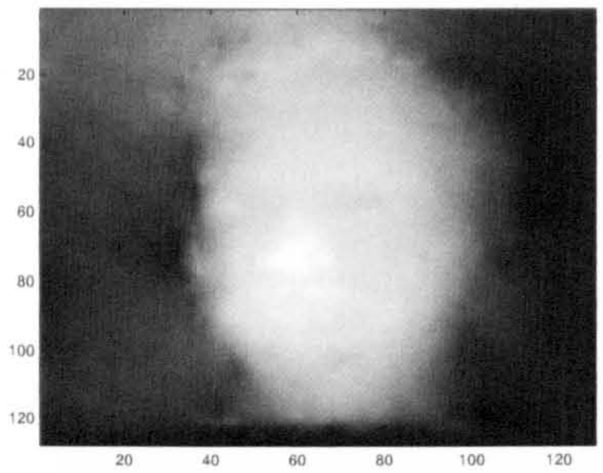


(c) Approximation at level 3

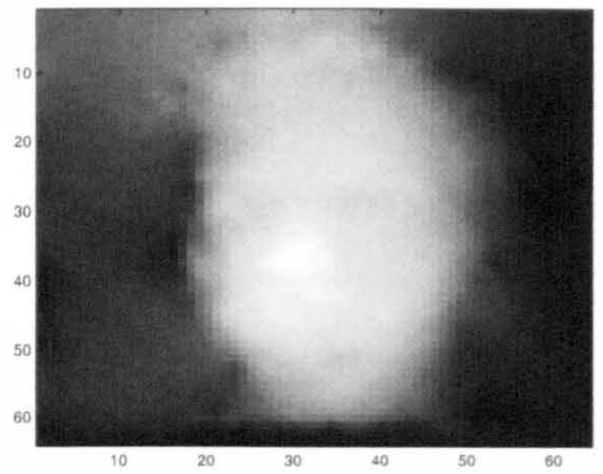


(d) Approximation at level 4

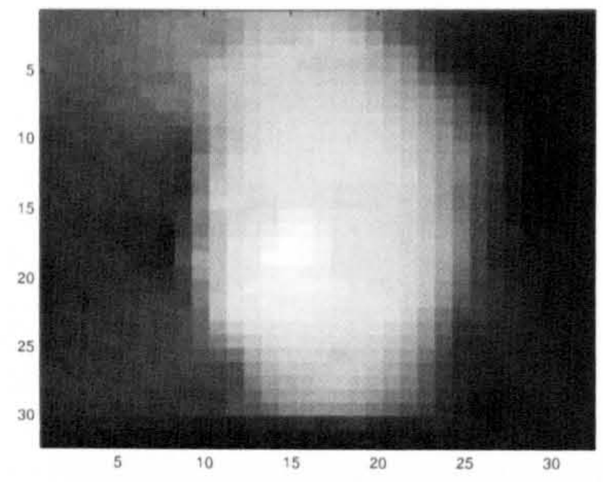
Figure 10.8: The output of the low-pass channel or the Approximation images for a 2D intensity image that has undergone 4 levels of decomposition using the Haar wavelet.



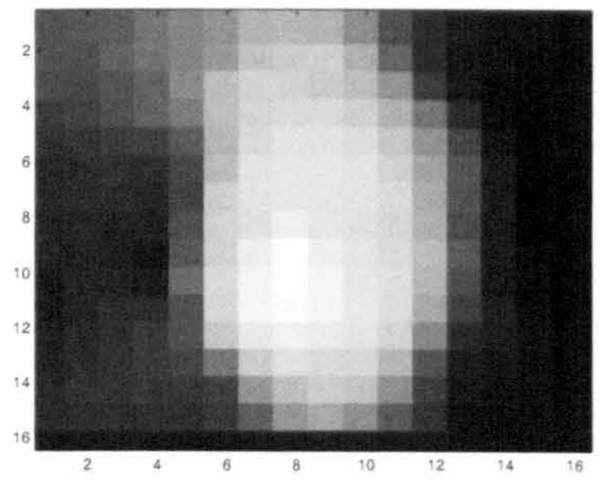
(a) Original $2\frac{1}{2}$D image



(b) Approximation at level 1



(c) Approximation at level 2



(d) Approximation at level 3

Figure 10.9: The output of the low-pass channel or the Approximation images for a $2\frac{1}{2}$D image that has undergone 3 levels of decomposition using the Haar wavelet. The $2\frac{1}{2}$D image corresponds to the 2D image shown in Figure 10.8.

10.3.1 Feasibility Study

A feasibility study was carried out in the first instance to see if the pre-processing led to any significant improvements in the recognition rates. The effects of wavelet pre-processing on dataset D_2 were investigated using the Eigenfaces algorithm in conjunction with the Euclidean distance metric. As before, LOO cross-validation was performed and the first 100 Eigenfaces were used for classification.

The recognition rates for the 2D and the $2\frac{1}{2}$D images in D_2 (from Section 10.2.1) are recalled in Table 10.3. These results were used as benchmark results and compared with the classifier accuracy after pre-processing.

	Recognition Rates (%)
2D	24.24
LR $2\frac{1}{2}$D	25.45
RL $2\frac{1}{2}$D	27.27

Table 10.3: Recognition rates for the 2D and the $2\frac{1}{2}$D images using the Eigenfaces algorithm and the Euclidean distance measure. LOO cross-validation is performed on the dataset D_2 and the input images undergo no pre-processing.

The results of the feasibility study were very promising. The 2D images are decomposed to 4 levels, while the $2\frac{1}{2}$D images are decomposed to 3. Pre-processing using the Haar wavelet improved the accuracy, but only marginally. The 2D recognition rates increased by 0.61%, while the $2\frac{1}{2}$D LR and RL rates increased by 1.22% and 0% respectively. These increased recognition rates were observed in decomposed images of size 64×64. The results with Magarey's wavelets were much better, though this was expected since these wavelets have a greater orientation selectivity and are generally better suited to image processing tasks.

The 2D recognition rate with Magarey's wavelet increased to 32.1% - a rise of almost 8% in the image of size 128×128 from channel B. The LR $2\frac{1}{2}$D images displayed an increase in recognition rates of between 6% and 24%. Improvements were observed in all forms of $2\frac{1}{2}$D images of size 64×64 from channel A. The RL images did not display such significant changes in the recognition rates. The classifier performance rose by a mere 3.6%.

Pre-processing the $2\frac{1}{2}$D images with Magarey's complex wavelets leads to a marked increase in the recognition rates for the LR images. On the basis of these results, LOO cross-validation was carried out on the dataset D_1 using the Eigenfaces algorithm alone (with both Euclidean and Mahalanobis distance metrics).

10.3.2 LOO cross-validation on D_1

As before, LOO cross-validation is performed on the 2D and the corresponding $2\frac{1}{2}$D images in the dataset D_1. Only the Eigenfaces algorithm is tested using both, Euclidean and the Mahalanobis distance measures. The images are pre-processed using the Haar wavelet and Magarey's complex wavelets. The 2D images (256×256) undergo four levels of decomposition and the $2\frac{1}{2}$D images (128×128) are decomposed three times, so that the final images are 16×16.

Table 10.4 summarises the outcome of this experiment. Note that the recognition rates presented here pertain to images of size 64×64. Images smaller than this give lower accuracy

than the benchmark (not pre-processed) images. The bigger images do not display significant improvements in the recognition rates. This is echoed by the findings of Ekenel et. al. (Ekenel & Sankur 2005) as well.

	% Recognition Rates		
	2D	$2\frac{1}{2}$D *LR*	$2\frac{1}{2}$D *RL*
Benchmark	67.05	63.15	61.71
Haar	67.49	63.73	61.42
Magarey	67.77	64.02	63.87

Table 10.4: Comparison of the recognition rates in 2D and $2\frac{1}{2}$D spaces using Eigenfaces algorithm. The input images have been pre-processed using the Haar wavelet and Magarey's complex wavelet. These recognition rates are achieved using the Mahalanobis distance measure. Note that only the highest recognition rates are presented here and no distinction is made between the various forms of the $2\frac{1}{2}$D images and outputs of channels A and B.

Again, the results of the LOO cross-validation on D_1 do not follow the same pattern as the results of LOO cross-validation on D_2. Compared to the results of the feasibility study, these results are extremely disappointing. The improvement in the recognition rates is far less pronounced and the $2\frac{1}{2}$D images still perform worse than the benchmark 2D images. The only advantages of the pre-processing are that the execution time is much faster and less storage space is required since the images are much smaller. These two advantages would make it worthwhile to use wavelet processed 2D images instead of the raw 2D images, especially since there is an improvement rather than deterioration in the classification performance. However, for this work, this line of investigation is not pursued any further. Instead, the reasons for the poor performance of the $2\frac{1}{2}$D images and ways of addressing their shortfalls were sought.

10.4 Composite Image Face Recognition

Classifiers operating in the $2\frac{1}{2}$D and 3D spaces perform worse than those in the 2D space. However, some encouraging results have been obtained for the $2\frac{1}{2}$D images.

2D images contain useful texture information. This information supplements the information from depth values in identifying features and individuals themselves. The importance of this intensity information in automatic face recognition had been gravely underestimated in this work. The consistently better performance of the classifiers in 2D space put this information into perspective.

As seen in the literature review in Chapter 2, many works combine the depth and texture information in different ways to take advantage of information from both these domains. It is known that humans use all available information to recognise faces (Todd 2002, Liu et al. 2000). In (Lam & Suen 1997) Lam and Suen give many examples from literature where combining the decisions of multiple classifiers has led to better recognition results. These findings are confirmed by other works such as (Lin et al. 2003, Tsalakanidou et al. 2003, Chang & Bowyer 2005, Ekenel & Sankur 2005) Consequently, both the texture and the depth information, both readily available from the Sheffield Dataset, were combined in a simple but effective way. This resulted in greater classifier accuracy in comparison with the

classifiers from either modality by itself.

Both, Eigenfaces and the Fourier K-NN classifiers transform the $n \times n$ input image into an n^2 vector before proceeding to the subsequent stages of the algorithm. This is done to simplify the implementation in case of the Fourier K-NN algorithm and to capitalise on certain matrix manipulation rules in case of the Eigenfaces. A simple way of combining the intensity information with the disparity information is to concatenate the two vectors. This combined representation of a 3D image is referred to as a *composite image* in this work. So, an $n \times n$ intensity image combined with an $m \times m$ disparity image would yield a composite image or a composite feature vector of the form

$$\begin{bmatrix} n^2 \text{ 2D image vector} \\ m^2 \ 2\frac{1}{2}\text{D image vector} \end{bmatrix} \tag{10.7}$$

and of length $n^2 + m^2$. It should be noted that this representation cannot be interpreted graphically and serves only as a means of combining and using effectively the intensity and the depth information. Although this is strictly speaking a composite *feature vector*, it is referred to as a composite *image* in this work. Since the $2\frac{1}{2}$D images come from two separate channels LR and RL, a separate composite image is obtained for each of these channels.

The composite image needs to be normalised prior to being input into a classifier since its 2D and the $2\frac{1}{2}$D components are measured in different units. The intensity images typically range from 0 to 255. The disparity images on the other hand measure displacement in sub-pixel values and the typical values range from -25.59 to 20.74. Normalisation of the two sets of values, prior to concatenation, is important in order to weight the two image components equally. If the minimum value is greater than or equal to zero, the matrix is normalised simply by dividing all the values by the maximum. If the minimum is less than zero, the absolute value of the minimum is added to the matrix to shift the minimum to zero, followed by division by the maximum value.

To the author's knowledge, intensity and depth information have not been combined in this way previously. Works such as (Chang & Bowyer 2005, Tsalakanidou et al. 2003) that utilise both shape and texture information in the form of 2D intensity and $2\frac{1}{2}$D range/depth images obtain classification scores for each type of image separately. These scores are then fused together using some kind of classifier combination rule (e.g. sum rule or the product rule). The approach adopted in this work extracts a single feature vector for both the domains and uses this to classify the test image, rather than using two or more separate feature vectors. In addition, normalisation for brightness and geometry is avoided. The implicit assumption that both images are illuminated reasonably uniformly still holds.

10.4.1 LOO cross-validation on D_1

The results of LOO cross-validation on dataset D_1 in the composite space are shown in Figure 10.10. In all, there are 10 composite image types, from the 10 $2\frac{1}{2}$D images (5 representations of LR and RL images). Both Eigenfaces and the Fourier K-NN classifiers are tested. Results for the Mahalanobis distance measure and the 1-NN are shown as these are the most accurate. Note that the recognition rates for complex composite images using the Eigenfaces classifier were not computed owing to time-constraints.

The Eigenfaces classifier produces the least accurate results in the composite space. It performs worse in the composite space than the $2\frac{1}{2}$D space, which was unexpected and

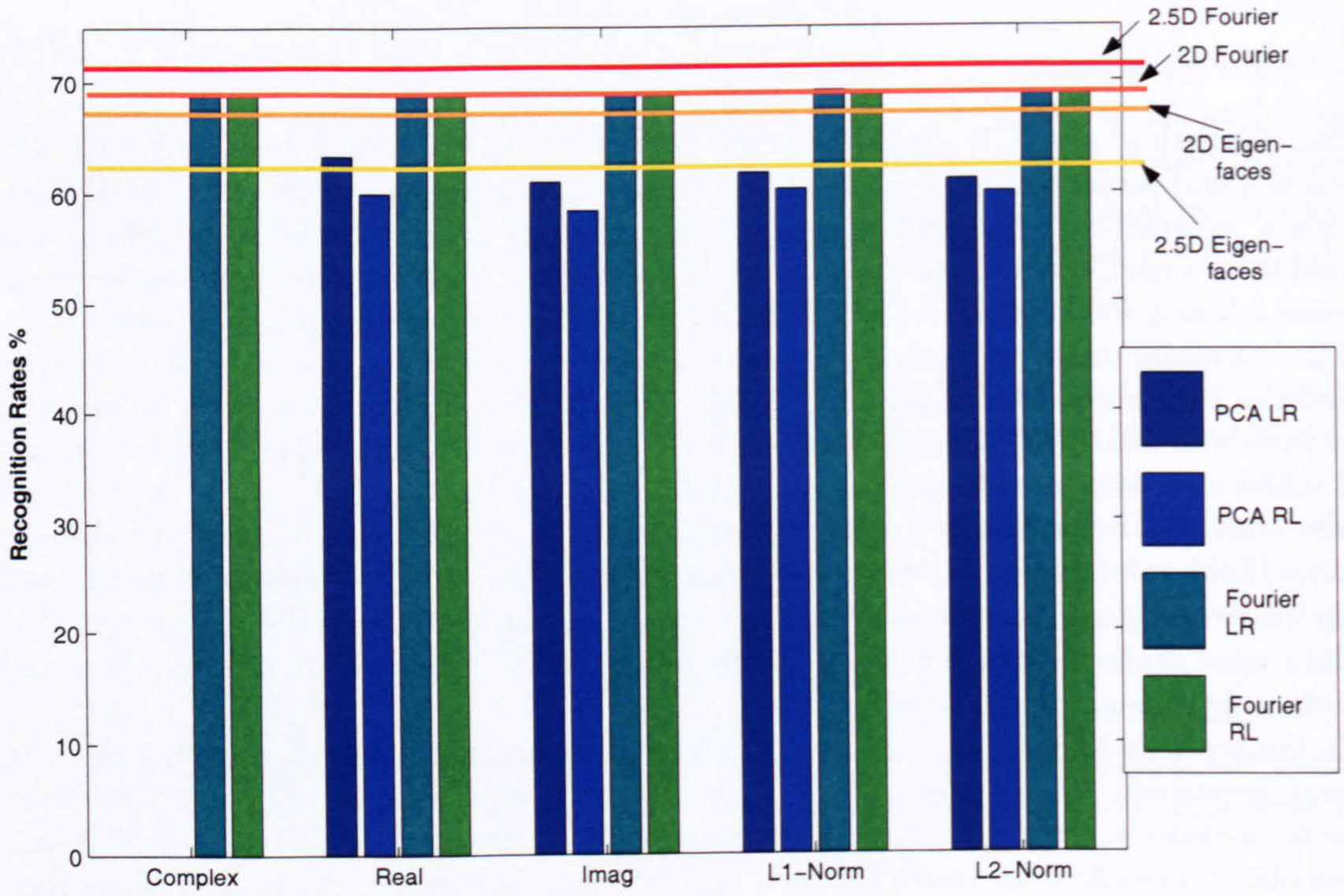


Figure 10.10: LOO cross-validation results for D_1 in the composite image space. Results for Eigenfaces (PCA) classifier using the Mahalanobis distance measure and Fourier 1-NN are shown for both LR and RL images. Also shown are the recognition rates for the 2D and the $2\frac{1}{2}$D spaces using the same classifiers.

disappointing. Fourier K-NN outperforms the Eigenfaces classifier, as it has done in all the previous experiments. However, contrary to the expectation that the classifiers in the composite space would be able to draw on the strengths of 2D and the $2\frac{1}{2}$D images, the Fourier K-NN classifier appears to be bounded by its performance in the 2D space. The highest recognition rates are still those yielded by the Fourier K-NN in $2\frac{1}{2}$D space.

10.5 Breakdown of recognition results in 2D and $2\frac{1}{2}$D spaces

So far, the $2\frac{1}{2}$D images have consistently performed better than the 2D images in the feasibility studies (dataset D_2). However, when tested on the larger dataset D_1, the recognition rates dropped significantly. The composite images, incorporating both 2D and the $2\frac{1}{2}$D images also performed worse than the 2D images.

One possible explanation for this anomaly between the results of the feasibility studies and the larger experiments was that the larger dataset contains images that are more varied than the smaller dataset. The images in D_1 can be partitioned into following *groups*:

Frontal	Smile	EyesClosed
LookUp	LookDown	EyesClosedLookUp
EyesClosedLookDown	Rotation1	Rotation2
RotationXYLook Up	RotationXYLookDown	Expressions
Lighting		

It may be possible that one particular type or types of images from the list above causes the classifier performance to deteriorate sharply in the $2\frac{1}{2}$D space. This would reflect in the overall performance of the classifier. In order to verify this, the LOO cross-validation results for D_1 were categorised according to the image groups. The results are presented in Figures 10.11 and 10.12 for the Eigenfaces (using Mahalanobis distance) and the Fourier 1-NN classifiers respectively.

Note that for the $2\frac{1}{2}$D images, only the highest benchmark results are charted. No distinction is made between the different representations of the $2\frac{1}{2}$D images. Analysis of the breakdown results have shown that for the LR images, the imaginary (46%) and the L_1 (39%) representations perform consistently well, while for the RL images, the real (61%) and the L_1 (31%) representations give good results. For the Fourier K-NN algorithm in the $2\frac{1}{2}$D space, the complex representation consistently gives better results.

Figure 10.11 clearly demonstrates that for most image groups, recognition rates for the 2D and the $2\frac{1}{2}$D spaces are very similar. However, 2D images perform significantly better (between 15% and 25%) for Rotation2 images, indicating that 2D Eigenfaces is more superior when dealing with variations in pose about the y-axis. $2\frac{1}{2}$D images on the other had give much better results than 2D for RotationXYLookUp and the harsh lighting images. It is known that perturbations in illumination causes the 2D Eigenfaces performance to deteriorate considerably. Depth information definitely enhances the performance of the Eigenfaces classifier when the faces have been imaged with uneven illumination.

Breakdown of recognition results using the Fourier K-NN algorithm shows that on the whole, recognition in the $2\frac{1}{2}$D space performs at least as well as, if not better than recognition in the 2D space. The 2D recognition rates are marginally better when the eyes are closed

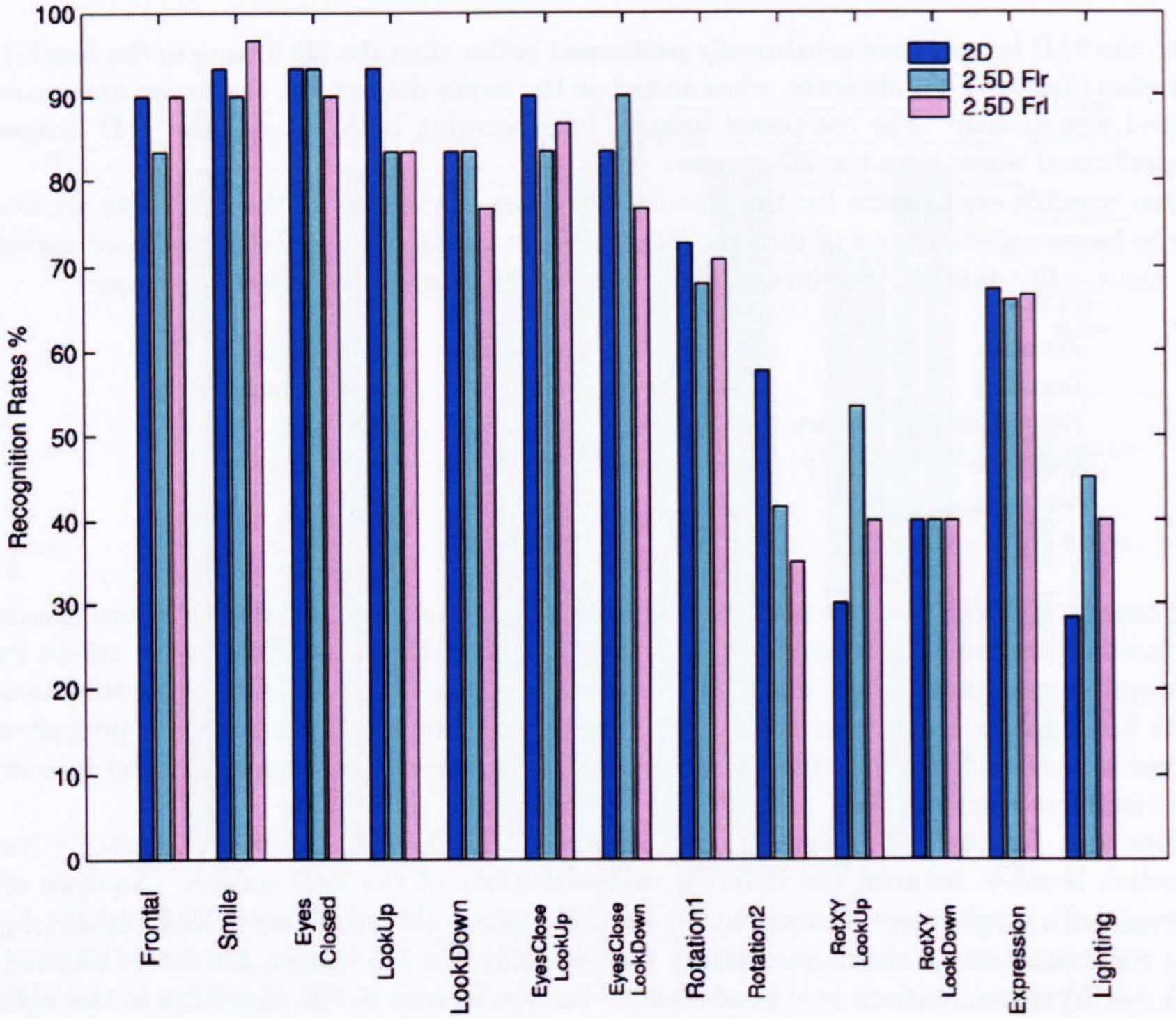


Figure 10.11: LOO cross-validation results for D_1 using the Eigenfaces algorithm with Mahalanobis distance, categorised according to the image types. Comparisons are drawn between the 2D and the $2\frac{1}{2}$D spaces.

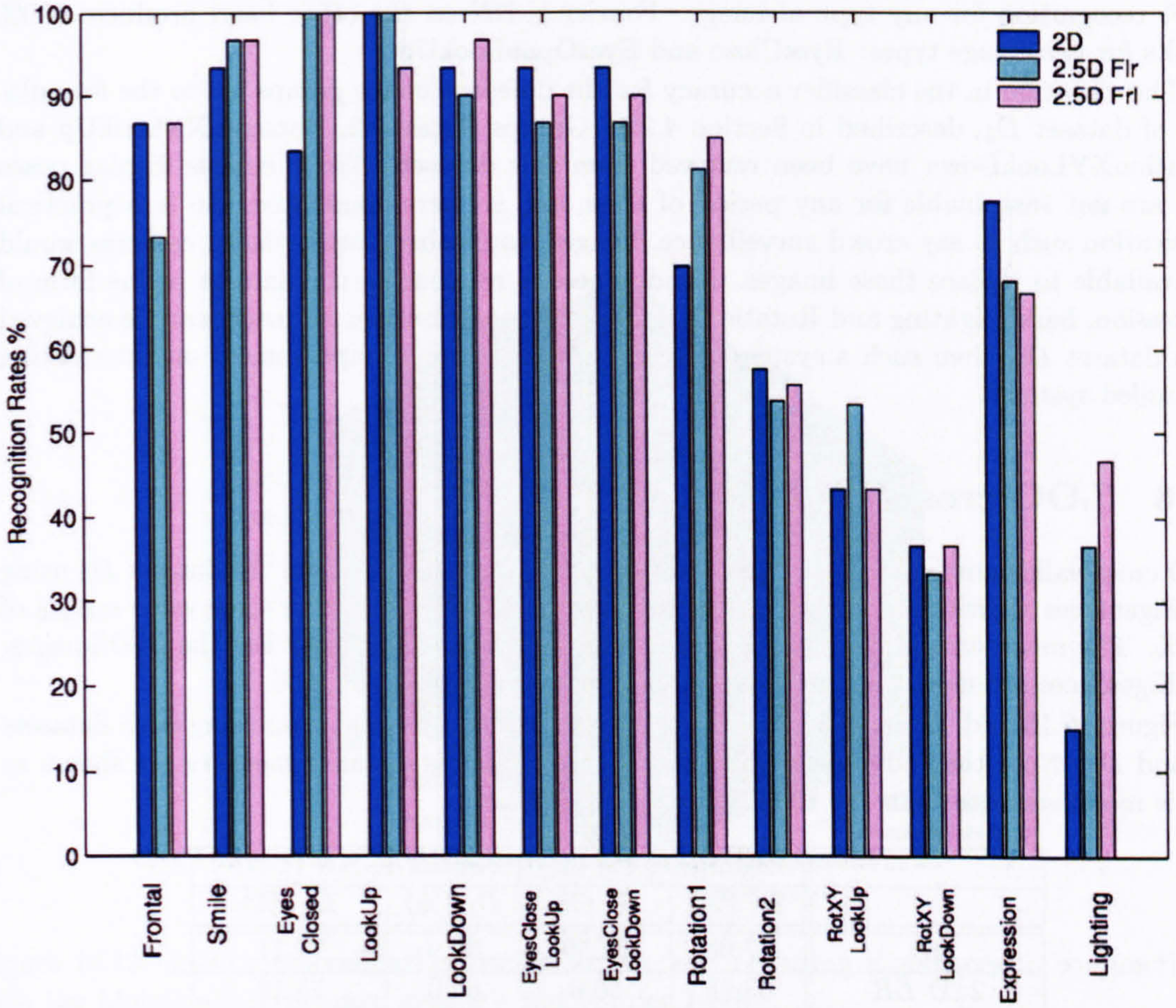


Figure 10.12: LOO cross-validation results for D_1 using the Fourier K-NN algorithm with 1-NN, categorised according to the image types. Comparisons are drawn between the 2D and the $2\frac{1}{2}$D spaces.

and the subject is either looking up or down. 2D Fourier K-NN performs significantly better than $2\frac{1}{2}$D when the subjects display varying expressions. As with the Eigenfaces classifier, in case of extreme lighting conditions, the $2\frac{1}{2}$D images perform better than 2D images by up to 30%. This, coupled with the marginally better performance of the $2\frac{1}{2}$D images for most other image types produces the consistently high recognition rates for the $2\frac{1}{2}$D space when using Fourier K-NN

Comparing Figures 10.11 and 10.12 also shows that the Eigenfaces classifier never achieves 100% recognition for any type of image. Fourier K-NN on the other hand produces 100% results for two image types: EyesClose and EyesOpenLookUp.

The variation in the classifier accuracy for the different image groups led to the formulation of dataset D_3, described in Section 4.2.1. Groups Rotation2, RotationXYLookUp and RotationXYLookDown have been removed from this dataset. These images display poses that are not sustainable for any period of time, and so were eliminated. So in a practical application such as say crowd surveillance, images from other better-placed cameras would be available to replace these images. Randomness is retained in the dataset in the form of expression, harsh lighting and Rotation1 images. If good classifier accuracy can be achieved with dataset D_3, then such a system would be a considerable improvement on the existing controlled systems.

10.6 LOO cross-validation on D_3

LOO cross-validation is performed 2D, $2\frac{1}{2}$D and composite images from the dataset D_3 using the Eigenfaces algorithm. 250 Eigenfaces are used for the 2D, giving an error value $\epsilon_{2D,250}$ of 3.31%. The mean error value $\epsilon_{comp,250}$ for composite images is 2.90%. For the $2\frac{1}{2}$D images, 100 Eigenfaces are used and the corresponding error value ϵ_{100} is 1.00%.

Figure 10.13 and Table 10.5 show the results of LOO cross-validation using both datasets D_1 and D_3. Note that only the results for the Mahalanobis distance measure are shown as this is more accurate than the Euclidean distance measure.

	Eigenfaces (M'bis)		Fourier K-NN (1-NN)	
	D_1 (%)	D_3 (%)	D_1 (%)	D_3 (%)
2D	67.05	62.59	68.9	71.11
$2\frac{1}{2}$D *LR*	63.15	50.0	69.9	75.19
$2\frac{1}{2}$D *RL*	61.71	48.89	71.4	77.22
Composite *LR*	63.44	71.48	68.9	71.11
Composite *RL*	60.04	70.0	69.1	71.11

Table 10.5: Comparison of LOO cross-validation results for datasets D_1 and D_3 using Eigenfaces with Mahalanobis distance and the first nearest-neighbour of the Fourier K-NN classifier.

Removing the images from the dataset that can be seen as containing relatively less useful recognition information deteriorates the performance of the Eigenfaces classifier in the 2D and the $2\frac{1}{2}$D spaces by an average of 10% (when using the Mahalanobis distance). The performance of the Eigenfaces classifier in the composite space shows an 8% improvement,

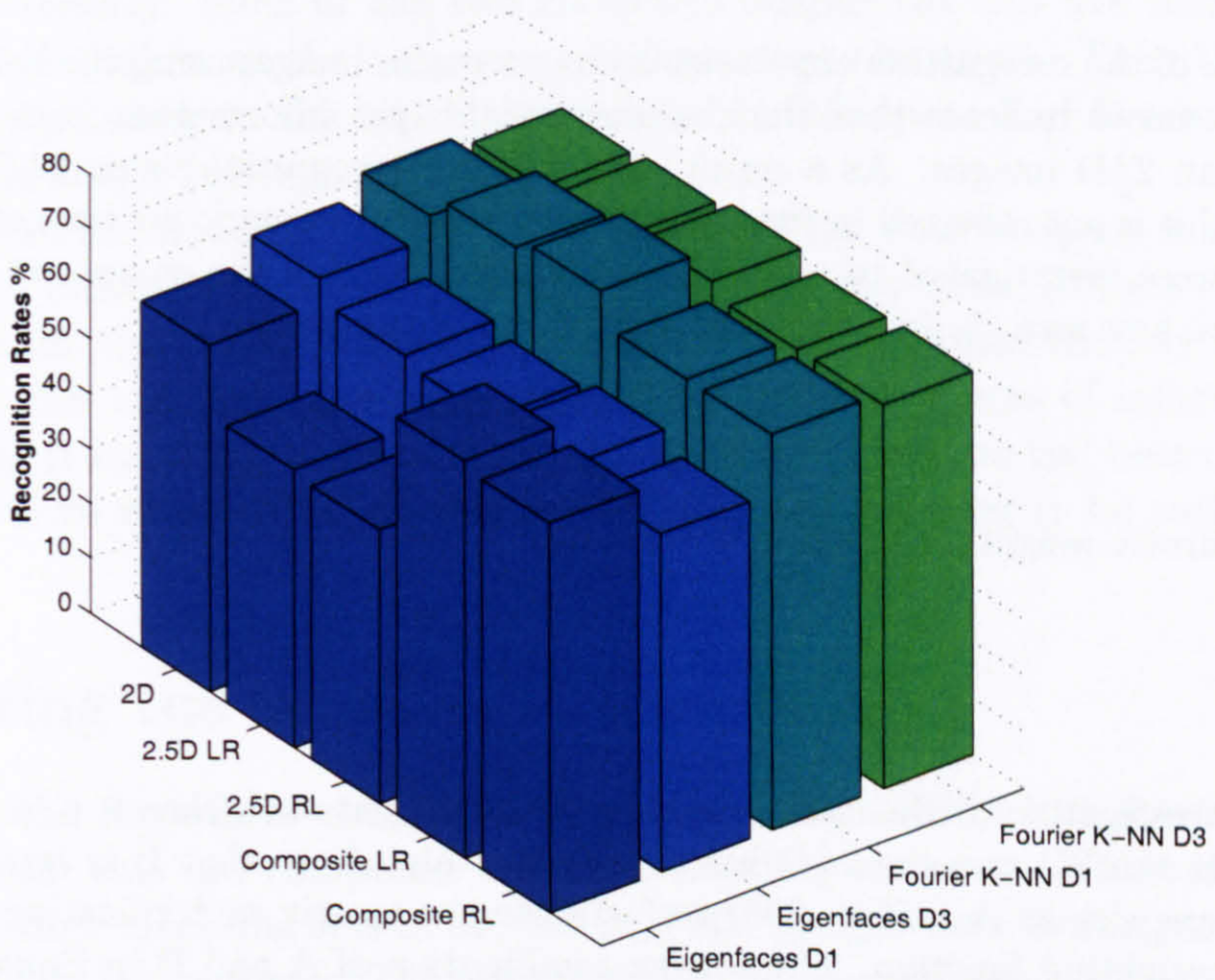


Figure 10.13: LOO cross-validation results for D_1 and D_3 using Eigenfaces in conjunction with the Mahalanobis distance measure and Fourier 1-NN.

with the highest recognition rate of 71.5%. This is the best this classifier has performed so far.

Fourier K-NN classifier is also tested on the dataset D_3. Unlike the Eigenfaces, this classifier's performance in the 2D and the $2\frac{1}{2}$D spaces improves by an average of 4.5% as a result of the dataset being modified. Although the performance in the composite space is better for dataset D_3 than it is for D_1, the improvement is very small - about 2%. This is lower than the 5.5% improvement observed in the $2\frac{1}{2}$D space. As noted in Section 10.4.1, the recognition rate for the composite spaces appears to be bound by the recognition rate for the 2D space. This is peculiar and may be a consequence of the way the 2D and the $2\frac{1}{2}$D components in the composite image are weighted by the classifier.

10.7 Composite images and the Fourier space based classifier

The results of the recognition experiments on composite images using the Fourier space based classifier seem to indicate that the classifier weights the information from 2D images more heavily than $2\frac{1}{2}$D images. As a result, even though recognition using $2\frac{1}{2}$D images is more accurate, this is not reflected in the recognition rates of the composite images. Two weighting functions were investigated to see if the recognition rates of the composite images could be improved in any way. A linear weighting function of the form

$$\begin{pmatrix} 1-\alpha \\ \alpha \end{pmatrix} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \tag{10.8}$$

and a quadratic weighting function of the form

$$\begin{pmatrix} (1-\alpha)^2 \\ 2\alpha(1-\alpha) \\ \alpha^2 \end{pmatrix} \begin{bmatrix} \mathbf{A}^2 \\ \mathbf{AB} \\ \mathbf{B}^2 \end{bmatrix} \tag{10.9}$$

were investigated. α denotes the value of the weight such that $0 \leq \alpha \leq 1$ and $\mathbf{A}$ and $\mathbf{B}$ represent the 2D and the $2\frac{1}{2}$D images respectively. Note that B is interpolated so that it is the same size as A (256×256) to facilitate the matrix multiplication required for the quadratic weighting function. The convex combination of $\mathbf{A}$ and $\mathbf{B}$ in Equation 10.9 results in the cross-coupling term $2\alpha(1-\alpha)\mathbf{AB}$. This has the effect of amplifying the underlying relation between $\mathbf{A}$ and $\mathbf{B}$, if there exists one. Both the weighting functions were tested on a dataset of 150 images. Five images were chosen at random from each of the 30 classes. The weighting functions were tested using LOO cross validation. No underlying relationship came to light despite investigating more than 30 different values of α between 0 and 1. The performance of the classifier on the composite images could not be improved beyond the bounds of the 2D recognition rates using Fourier K-NN classifier.

The linear weighting function produced better results in that for all values of α, the recognition rate of the composite images was the same as the recognition rate of the 2D images. The quadratic weighting function reported 30% lower results than the linear weighting function for all values of α except $\alpha = 0$, for which the recognition rate was the same as that achieved with the linear weighting function. This rate was in turn the same as that achieved without any weighting function.

Failure of the weighting function to improve the performance of the Fourier K-NN classifier in the composite space led to the exploration of an alternative way of combining the results of

the 2D and the $2\frac{1}{2}$D spaces. Majority voting is the simplest way of collating information from multiple classifiers. In (Lam & Suen 1997) it has been reported to give very good performance in a variety of applications. In this classifier combination strategy, the combined decision is obtained by a majority vote of the individual classifiers. When n classifiers are combined, the test sample is assigned the class for which there is a consensus, or when at least k of the classifiers are agreed on the identity, where (Lam & Suen 1997)

$$k = \begin{cases} \frac{n}{2} + 1 & \text{if } n \text{ is even} \\ \frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases}$$

In this work, the two classifiers are Fourier K-NN in the 2D space and Fourier K-NN in the $2\frac{1}{2}$D space. Majority voting is impossible to apply with only two classifiers. Hence the strategy is adapted slightly. Each of the two classifiers output the top five matches, Top5-NN. These matches are taken as the outputs of independent classifiers. Thus there are ten classifiers: five corresponding to the top 5 matches in the 2D space and five corresponding to the top 5 matches in the $2\frac{1}{2}$D space.

Unfortunately, this did not improve the performance of the Fourier space based classifier either. The average recognition rate was circa 42%. This was most likely to be a consequence of the fact that the results that were combined were not from entirely independent classifiers. Further, this strategy does not combine the different pieces of information from the different images. It implicitly assumes that the classifiers used are the best classifiers for the recognition problem at hand. This assumption may not necessarily be satisfied in this case.

10.8 Capturing the Systematic Variability

In this experiment, the Eigenfaces classifier was trained using 3 different subsets of D_3. The recognition rates from this are compared with the classifier accuracy achieved with leave-one-out cross-validation. Training sets of 30, 150 and 300 images were investigated, with 1, 5 and 10 images per class respectively. The remainder of the images were used for testing. The training images were chosen randomly, and the experiment was repeated 10 times to give mean recognition rates for the different classifiers, along with the standard deviation, which acts as an error measure. Recognition was performed in 2D,$2\frac{1}{2}$D and the composite spaces. The same images were used for training and testing the classier in each of the spaces. The same number of Eigenfaces were used for recognition in the 2D and the composite spaces: $\{30, 100, 175, 250\}$ for training sets of size $\{30, 150, 300, 539\}$. The corresponding Eigenfaces for the $2\frac{1}{2}$D images are $\{30, 50, 75, 100\}$. These were chosen to such that the RMS error was $< 5\%$

The mean recognition rates along with the error margins are presented in Table 10.6.

As before, the Mahalanobis distance measure outperforms the Euclidean by a significant margin. And as expected, increasing the number of training images per class increases the classifier accuracy, with the best results obtained for leave-one-out cross validation.

The results from this experiment clearly indicate that the composite image representation incorporating both 2D texture information and $2\frac{1}{2}$D depth information is a significant improvement on either representation by itself. As in the previous experiments, the images of individuals with and without glasses are treated as belonging to different classes. If they are treated as images of the same individual, then the highest recognition rates obtained for the

	Mean Recognition Rate %											
	2D				$2\frac{1}{2}$D				Composite			
imgs/class	1	5	10	LOO	1	5	10	LOO	1	5	10	LOO
Euc.	29.06	39.46	43.04	**48.15**	22.55	28.87	28.96	31.85	32.06	**46.74**	50.17	53.15
M'bis	33.27	53.13	58.58	**62.59**	30.16	43.95	46.46	50.00	37.96	**61.79**	67.71	71.48
Error ±	3.15	2.11	2.68		2.72	2.77	2.85		3.82	1.39	2.58	

Note that the error margins for both Euclidean & Mahalanobis distance measures are equal

Table 10.6: Mean recognition rates and error margins after 10 recognition experiments with the Eigenfaces classifier trained on datasets containing 1, 5, 10 and ≈17 images per class, chosen randomly from D_3. Note the composite classifier is consistently better than the 2D classifier. Further, at only 5 training images per class, it achieves comparable performance to LOO classifier which has ≈ 17 training images per class (recognition scores in bold font).

composite image classifier in LOO cross-validation increases from 71.48% to 76.85%. This higher recognition rate is comparable with the classifier accuracies reported in the literature on relatively less challenging datasets.

In order to establish that the improvement in the classifier accuracy is indeed a consequence of using the disparity information and not simply using additional information, the recogniser was tested by simply concatenating the left and right 2D stereo images, and performing leave-one-out cross validation. Recognition rates were found to be 66.67%, which is comparable with the results achieved using 10 composite images. The accuracy is higher than that obtained using single or 5 composite images. However, this is offset by the significant increase in the computational complexity and classification time. These results also verify that the accuracy of the composite image based classifier is a consequence of using the disparity information.

As explained before, different expressions and poses of the face with respect to the camera result in variation in the image space which is systematic. This systematic variability is modelled as a random variability by most 2D and $2\frac{1}{2}$D recognition algorithms. The classifier is trained on a large number of images of each subject displaying a variety of poses and expressions. This obviously increases the computational overheads and decreases efficiency.

The composite image representation presented in this work addresses this issue of systematic variability. The Eigenfaces classifier achieves greatest accuracy in the 2D space (62.59%) when the maximum number of training images is used. Similar accuracy (61.79% ± 1.39) is obtained in the composite space using only a fraction of the images to train the classifier. This clearly indicates the usefulness of depth information used in conjunction with texture information in face recognition.

10.8.1 Disparity Maps vs. Depth Maps

In literature, $2\frac{1}{2}$D images refer to 2D images where the intensity values have been replaced by actual depth values obtained from range images, laser scans or through structured light projections. These are often referred to as 3D images, depending on the authors' preferences. The existing 2D image processing and face recognition techniques are then applied to these

images and good recognition accuracy has been reported (Heseltine et al. 2004*d*,*b*,*c*,*a*).

Very few existing $2\frac{1}{2}$D face recognition systems obtain their depth values from stereo images. The main reason for this is that the depth values obtained in this way are extremely noisy as errors are introduced in the depth computation at image matching, camera calibration and triangulation stages.

In this work, depth maps (see Figure 10.14 (a) and (b)) for images in D_3 were generated using the camera matrices from Chapter 4 and the triangulation algorithm described in Chapter 3.

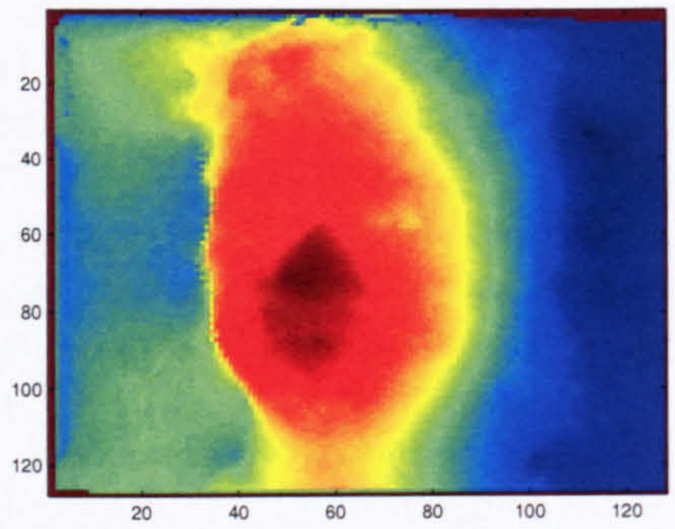
Figure 10.14 depicts the two depth maps and the corresponding disparity map and the 2D intensity image. It can be seen that the depth maps do not appear significantly different to the disparity maps. However, the relief information contained in depth maps is corrupted by noise and the errors accumulated in the reconstruction process. This is underscored by the results in Table 10.7. The Eigenfaces classifier was trained using 150 images from D_3, with 5 images per class. 10 runs of the recognition experiment were executed. Training images in each run are the same as those used in the last experiment (Section 10.8). The composite images are constructed by appending the depth values to the 2D image vector instead of the disparity values. No distinction is made between the LR and RL depth images - only the highest accuracy is reported. 50 Eigenfaces are used for classifying the depth maps and 100 for the composite images. Again, the RMS error was kept below 5%.

Mean Recognition Rate %		
	Depth Maps	**Composite**
imgs/class	5	
Euc.	14.28	36.69
M'bis	28.18	52.36
Error ±	1.89	1.45

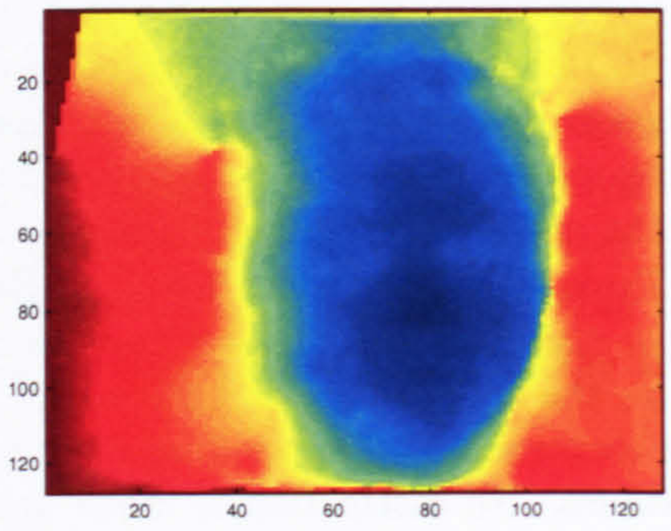
Table 10.7: Mean recognition rates and error margins after 10 recognition experiments with the Eigenfaces classifier trained using 5 images per class, chosen randomly from D_3. Comparing with the results in Table 10.6 shows that in noisy environments, using disparity values gives better classifier accuracy than the depth values.

Note that the error margins for both Euclidean and Mahalanobis distance measures are equal.

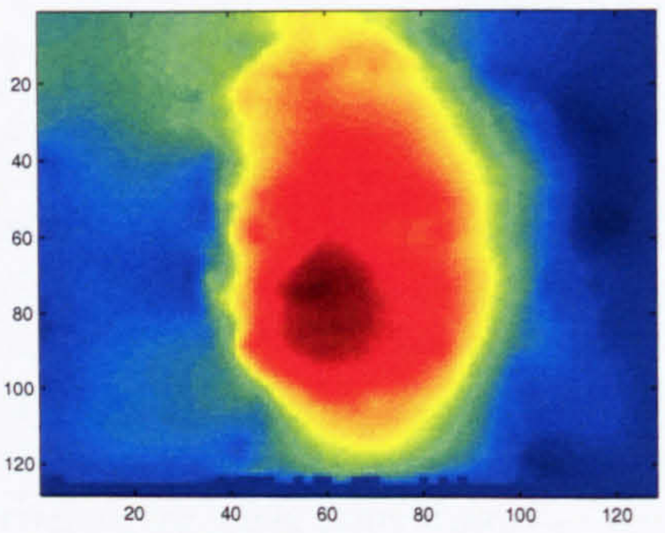
These results clearly indicate the superiority of the disparity information compared with the reconstructed depth information in noisy environments. Being able to use depth information obtained from ordinary cameras in a stereo configuration, without the need for error-prone camera calibration and reconstruction processes vastly increases the scope of face recognition applications. The work can also, in theory, be extended to dynamic images such as those from CCTV cameras. Very little information is available in the literature on completely automatic dynamic face recognition systems since much of this research is carried out commercially.



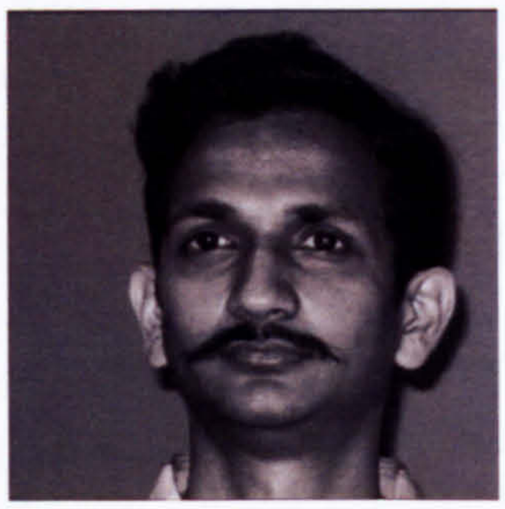
(a) LR Depth Map



(b) RL Depth Map



(c) $2\frac{1}{2}$D Disparity Map



(d) 2D intensity image

Figure 10.14: LR and RL depth maps generated using the camera matrices and the triangulation algorithm described in Chapters 4 and 3. Figures (c) and (d) show the corresponding disparity map and the intensity image.

10.9 ROC Curves

Receiver Operator Characteristics (ROC) graphs are a means of visualising, organising and selecting classifiers based on their performance (Fawcett 2004). Face recognition systems are usually required to perform either identification or verification (or both) depending on their application. Identification deals with the task of assigning an identity label to a given image and is a 1:many task. Verification on the other hand is a 1:1 task and deals with the task of checking or verifying whether identity label assigned to a particular image is correct. This is done by measuring the distance of the image from the class and comparing it against a threshold value. If the distance is less than or equal to the threshold, the claimed identity is accepted as being true, otherwise false.

Consider a two class classification problem (such as verification). Each *instance* I of a dataset has mapping to one element of the set $\{\mathbf{p}, \mathbf{n}\}$ of positive and negative class labels (Fawcett 2004). A classifier is a mapping from the instances I to predicted classes with labels $\{\mathbf{Y}, \mathbf{N}\}$. Given a classifier and an instance, there are four possible outcomes (Fawcett 2004): true positive, false positive, true negative and false negative.

In case of face recognition:

- **true positive:** When an individual has assumed his/her correct identity, and is rightly accepted by the classifier.
- **true negative:** When an individual has assumed a false identity and the classifier wrongly accepts him/her.
- **false negative:** When an individual has assumed his/her correct identity but the classifier still rejects him/her.
- **false positive:** When an individual has assumed a false identity and he/she is correctly rejected by the classifier.

For the verification problem, the classifier is given as input an image and class label. The distance of the image is computed from the class, and if it is below the threshold θ the claim is accepted, otherwise not. The assignment of true/false positive/negative labels is shown schematically in Figure 10.15.

ROC graphs plotted with $T+\ rate$ on the y-axis and the $F+\ rate$ on the x-axis. These are defined as:

$$T+\ rate \approx \frac{\text{No. of positives correctly classified}}{\text{Total no. of positives}}.$$

and

$$F+\ rate \approx \frac{\text{No. of negatives incorrectly classified}}{\text{Total no. of negatives}}.$$

The $T+\ rate$ and the $F+\ rate$ values range between 0 and 1 and hence the ROC space has unit area. ROC curves depict the relative trade-offs between the benefits (T+'s) and the costs (F+'s) of using a given classifier. The curves have the special property that they are insensitive to changes in class distribution. That is, if the proportion of positive to negative instances changes in a test set, the ROC curves will not change (Fawcett 2004).

The perfect classifier, in the ROC space is represented by the point $(0, 1)$. The diagonal line $y = x$ represents the random classifier, which can be expected to correctly classify

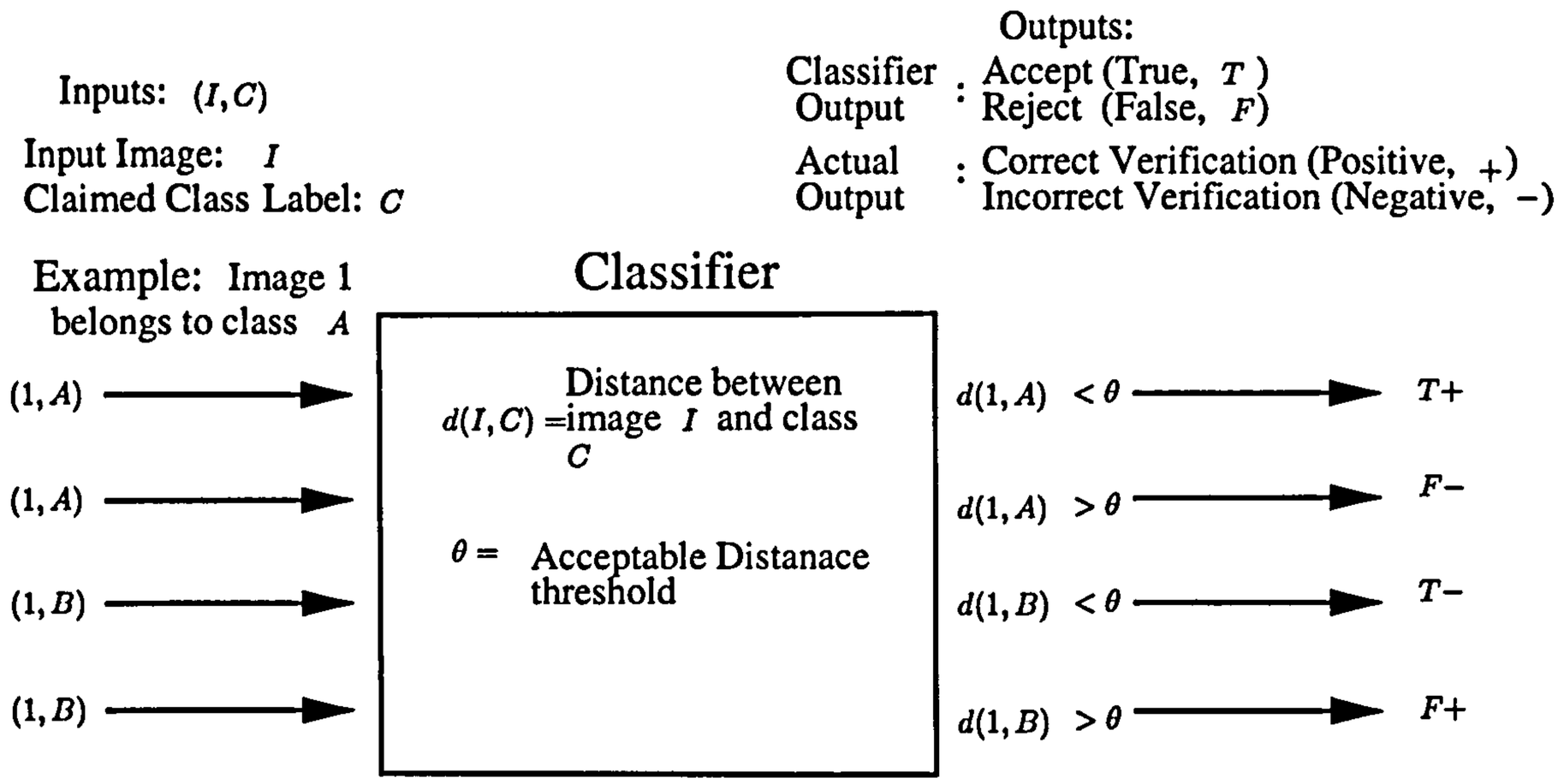


Figure 10.15: The assignment of true/false positive/negative labels for the face verification problem. For a given image 1, belonging to class A, if $d(1, A) < \theta$ then the classifier correctly accepts the individual and his/her claimed identity $(T+)$. If $d(1, A) > \theta$, the individual's classifier rejects the individual and his/her claimed identity even though the individual belongs to the database and has the correct class label $(F-)$. $(1, B)$ represents an individual who is claiming false identity. The individual may or may not belong to the database. If $d(1, B) < \theta$, the classifier incorrectly accepts that individual $(T-)$ and if $d(1, B) > \theta$ the individual is rejected $(F+)$

individuals approximately 50% of the times. For a classifier to be better than the random classifier, it needs to lie above the $y = x$ line. A classifier lying below the $y = x$ line performs worse than the random classifier. However, since this decision space is symmetrical about the $y = x$ line, if the decision of a classifier lying in the below this line is negated, then its ROC curve lies above the diagonal line of symmetry (Fawcett 2004).

The ROC performance can be reduced to a single number to allow quantitative comparison of classifiers.

A common method is to calculate the area under the ROC curve (AUC) (Hanley & McNeil 1982, Bradley 1997). The area of the entire ROC space is 1 (unit square), so the value of the AUC will always lie between 0 and 1. The diagonal line $y = x$ represents the random classifier and since it bisects the unit square, its AUC is 0.5. Hence, no realistic classifier should have an AUC of less than 0.5. The AUC of a classifier has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett 2004). By this token, a higher value of AUC for a given classifier indicates a better classifier.

Figures 10.16 and 10.17 show the ROC curves in the 2D, $2\frac{1}{2}$D and the composite spaces for the Eigenfaces classifier used in conjunction with the Euclidean distance measure and the Mahalanobis distance measure respectively, using dataset D_3. The ROC curves for the Fourier K-NN classifier using dataset D_3 in the 2D, $2\frac{1}{2}$D and the composite spaces is shown in Figure 10.18. The curves are generated by computing the true positive and false positive rates for all the individuals in the dataset D_3 at 101 threshold values between 0 and 1 inclusive. The identity of each of the 540 images in D_3 is checked against the 30 classes in the dataset for the 101 different threshold values. In total, 1,636,200 ($540 \times 30 \times 101$) verification operations are conducted for each classifier.

The corresponding values for the AUC are given in Table 10.8.

	2D	$2\frac{1}{2}$D (Imaginary LR)	Composite (Imaginary LR)
Eigenfaces (Euclidean)	0.8922	0.8880	0.8749
Eigenfaces (Mahalanobis)	0.9221	0.9186	0.9130
Fourier K-NN	0.9554	0.9808	0.9553

Table 10.8: The AUC values for the ROC curves in figures 10.16, 10.17, 10.18, corresponding to the performance of the Eigenfaces classifier (using both Euclidean and Mahalanobis distances) and the Fourier K-NN classifier in the 2D, $2\frac{1}{2}$D and the composite spaces. Dataset D_3 is used for producing the curves and computing the AUC values.

10.10 Miscellaneous

This section contains a brief analysis of the results to see which individuals in the dataset were the best and the worst recognised. Also studied are the effects of glasses and head-scarf.

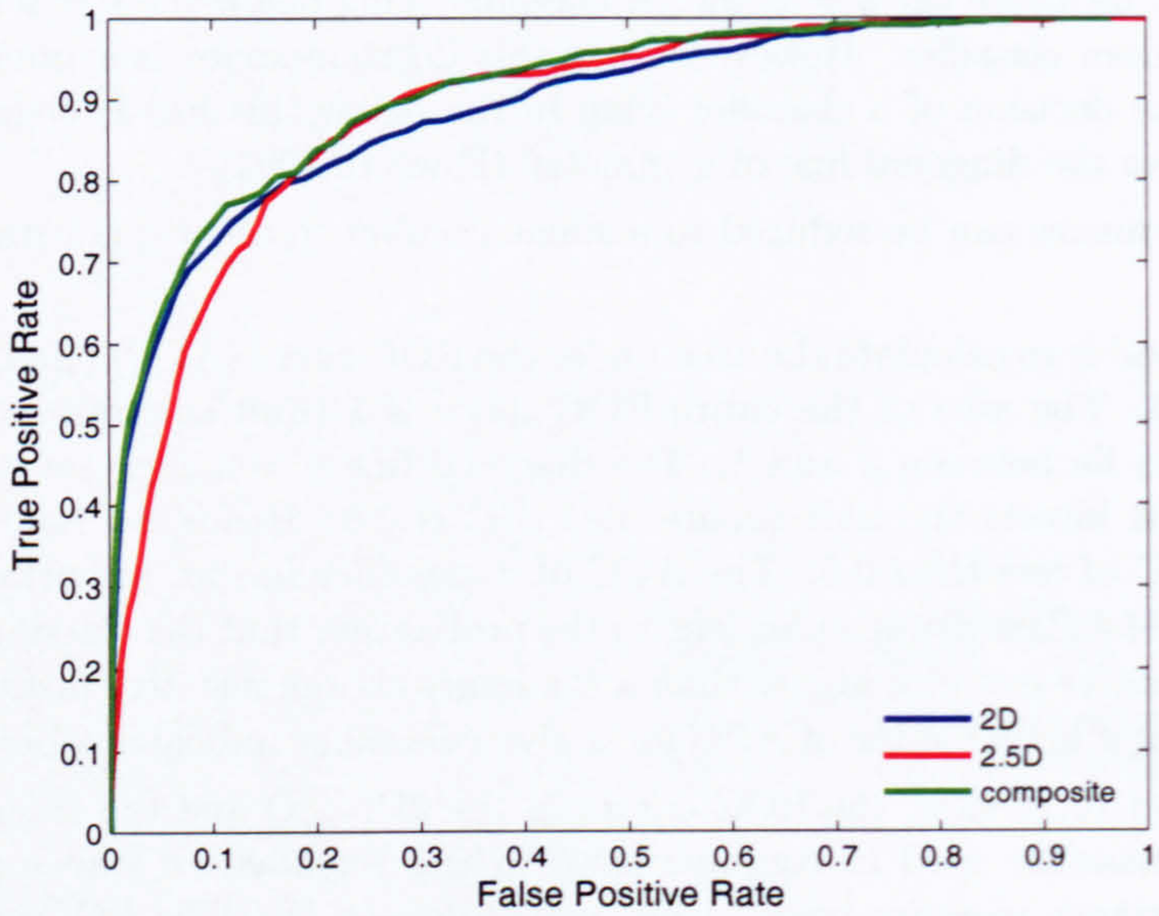


Figure 10.16: ROC curves for the Eigenfaces classifier (Euclidean distance) in the 2D, $2\frac{1}{2}$D and the composite spaces, using dataset D_3

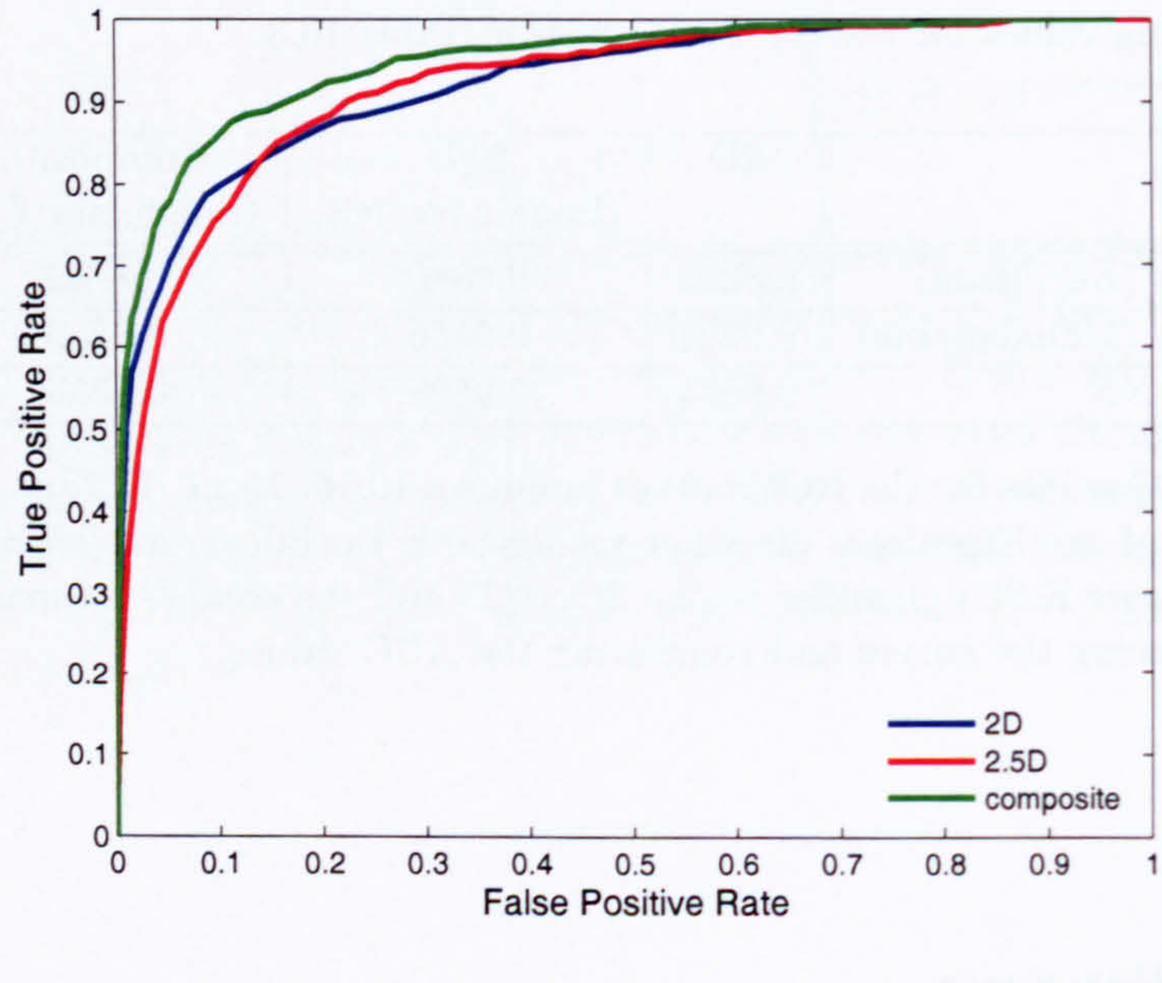


Figure 10.17: ROC curves for the Eigenfaces classifier (Mahalanobis distance) in the 2D, $2\frac{1}{2}$D and the composite spaces, using dataset D_3

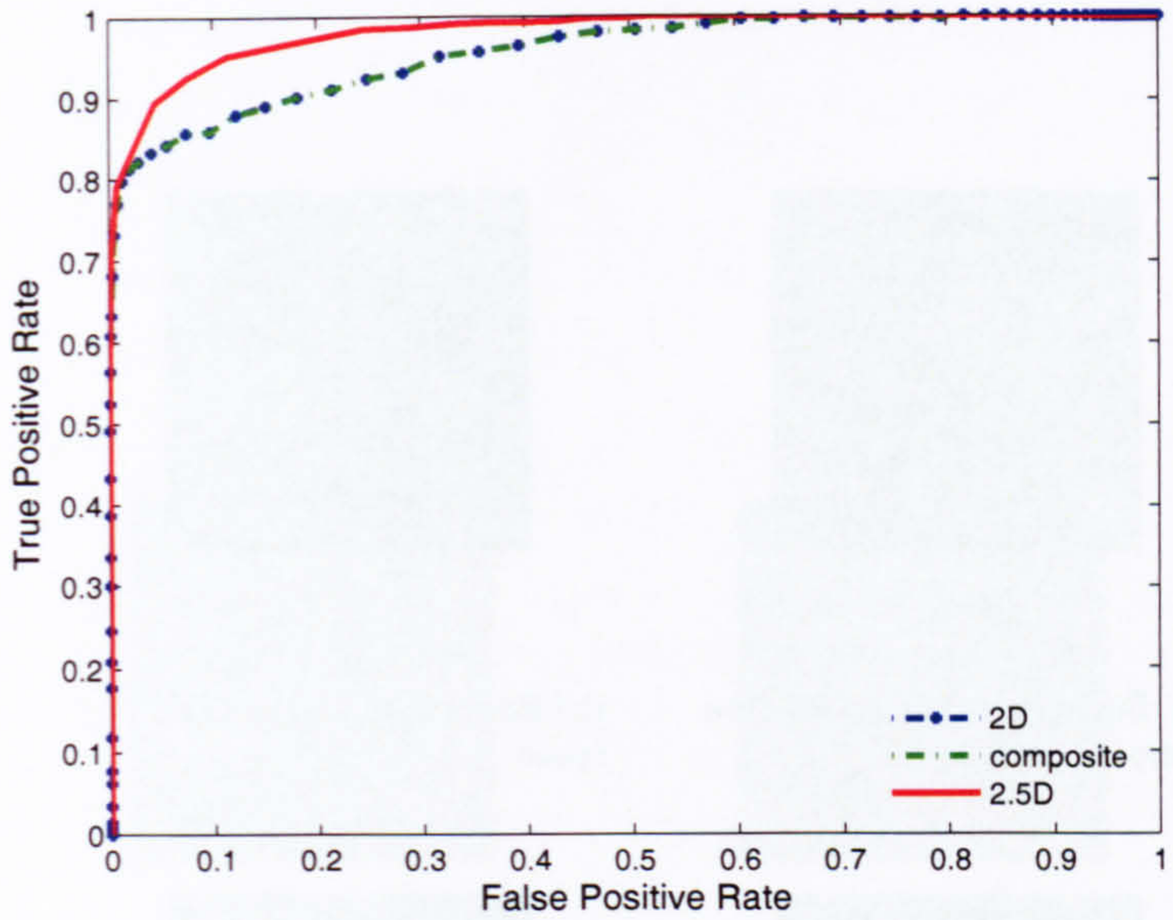


Figure 10.18: ROC curves for the Fourier K-NN in the 2D, $2\frac{1}{2}$D and the composite spaces, using dataset D_3

10.10.1 Best and Worst Recognised Individuals

The best and the worst recognised individuals for Eigenfaces classifier and the Fourier K-NN classifier are shown in Figures 10.19 and 10.20 respectively.

The fact that subjects 6 and 16 were both consistently classified correctly by both the classifiers could be a consequence of the dataset containing only two individuals of Afro-Carribean origin. The images of these two individuals are significantly different from the other subjects at least in the 2D and the composite spaces, and are therefore always correctly classified. Investigations using a larger dataset with greater variation in the subject ethnic groups is required to investigate the effects of ethnic origin on classifier performance.

10.10.2 Effects of Head-Scarf

Recall that in the Sheffield Dataset, two of the females are imaged with a head-scarf. Of these, one is also imaged without the head-scarf. Refer to Figure 10.21 to see the subject numbers that correspond to the various images of the two individuals.

The recognition results of the classifiers were checked for two things. Firstly, is the classifier able to distinguish between the images of subjects $\{24, 25\}$ and subjects $\{17, 18\}$, and secondly, does the classifier realise that subject 22-25 are in fact the same individual.

The Eigenfaces classifier in 2D space struggled the most with the first criteria. On average, about 25% of the times it confused subjects $\{24, 25\}$ and $\{17, 18\}$. The misclassification rate fell to 18% when using the Mahalanobis distance measure. The classifier performed much better in the $2\frac{1}{2}$D and the composite spaces. In the composite space, the misclassification

(a) Best recognised in 2D and composite spaces

(b) Best recognised in the $2\frac{1}{2}$D space

(c) Worst recognised in 2D and composite spaces

(d) Worst recognised in the $2\frac{1}{2}$D space

(e)

(f)

(g)

Figure 10.19: In addition to subjects (c) and (d), subjects (e), (f) and (g) are frequently misclassified in 2D, $2\frac{1}{2}$D and composite spaces using the Eigenfaces classifier.

(a) Best recognised in 2D and composite spaces, Worst recognised in the $2\frac{1}{2}$D space

(b) Best recognised in the $2\frac{1}{2}$D space

(c) Worst recognised in 2D and composite spaces

(d) Worst recognised in the $2\frac{1}{2}$D space

Figure 10.20: The best and the worst recognised individuals when using the Fourier K-NN classifier.

(a) subject 22

(b) subject 23

(c) subject 24

(d) subject 25

(e) subject 17

(f) subject 18

Figure 10.21: The two subjects imaged with head-scarf

rate dropped to circa 14% for the Euclidean distance measure, and to circa 5% for the Mahalanobis distance measure. The classifier performed the best in the $2\frac{1}{2}$D space, with 0% misclassifications, regardless of the distance measure used. These results reinforce the fact that Mahalanobis distance measure has better discriminatory power and that classifiers make effective use of the depth information (in $2\frac{1}{2}$D images) to correctly classify those images that are harder to classify in the 2D space.

On the same criteria, the Fourier K-NN classifier performed much better. In the 2D and the composite spaces, the misclassification rate was as low as 3%. Similarly to the Eigenfaces classifier, in the $2\frac{1}{2}$D space, the classifier is not confused between the images of subjects 17, 18, 24 and 25.

On the second criteria, both the classifiers performed poorly in 2D and the composite spaces. The classifiers are unable to recognise the similarities between subjects 22-25 and are unable to realise that these subjects are in fact the same individual. In the $2\frac{1}{2}$D space, the classifiers perform better. They "realise" that subjects 22-25 are the same individual circa 4% and 15% of the times when using the Fourier K-NN and the Eigenfaces classifiers respectively. These results echo the findings in Section 9.3.3.

It should be noted here that these results are only computed out of interest. To draw any reliable conclusions about the effects of head-scarf, a much larger dataset is required.

10.10.3 Effect of Glasses

As mentioned before, 6 of the individuals in the database are imaged both with and without glasses. In all the results presented in this chapter, images of an individual with glasses and without glasses are treated as two separate individuals.

The classifier accuracy obviously increases if the images of individuals with and without glasses are treated as the same individual. In LOO cross-validation on D_3 using Eigenfaces and Mahalanobis distance measure, the recognition rates increase by 10%, 5% and 5% for the 2D, $2\frac{1}{2}$D and the composite spaces respectively. The new recognition rates are 71.85%, 54.63% and 76.65%.

The improvement in the performance is slightly higher when the Fourier K-NN classifier is used. Increases of approximately 13.5%, 6.4% and 13.5% are noted, with the new recognition rates of 84.26%, 83.52% and 84.44% in the 2D, $2\frac{1}{2}$D and the composite spaces respectively. Notice that the recognition rates for the 2D and the $2\frac{1}{2}$D spaces are almost similar. In this case, depending on the application and the training dataset size, it may be better to use $2\frac{1}{2}$D images rather than 2D images since the processing time is much faster for the smaller $2\frac{1}{2}$D images.

10.11 Summary

The concepts of disparity based $2\frac{1}{2}$D images and composite images are introduced in this chapter. $2\frac{1}{2}$D images are 2D images that encode the depth information of the scene in the form of disparity values. Five different representations of $2\frac{1}{2}$D images are presented. Although the results of the feasibility study revealed marked improvement in the classifier performance using $2\frac{1}{2}$D images, the results of experiments on the larger dataset D_1 were extremely disappointing. $2\frac{1}{2}$D images performed worse than the 2D images in these experiments.

Composite images are a simple and effective way of combining 2D and $2\frac{1}{2}$D images, such that the strengths from both these dimensions are utilised in an advantageous way. This image representation also performed worse than the 2D images when tested on the Sheffield Dataset D_1.

Wavelets-based pre-processing is investigated to see if the performance of the benchmark classifiers in the $2\frac{1}{2}$D space can be improved. Again, the feasibility study showed promising results, but the experiments using the larger dataset were only marginally better than before. The only advantage of using the pre-processing is that the resulting wavelets-decomposed images are much smaller, and so the storage requirements and the processing power requirements are greatly reduced.

The recognition results of dataset D_1 were analysed to see which image groups performed well and poorly in each of the 2D and the $2\frac{1}{2}$D spaces. The groups resulting in the lowest recognition rates are Rotation2, RotationXYLookUp and RotationXYLookDown. As a result, further experiments are conducted only using dataset D_3, which does not contain these image groups. LOO cross-validation on this dataset yielded the expected results: the composite representation performed better than both 2D or $2\frac{1}{2}$D representations.

Finally, the performance of the Eigenfaces classifier is tested in all 3 spaces using training sets of different sizes. As the training set gets larger, the classifier accuracy gets better. The composite classifier consistently performs better than the 2D classifier. In addition, it achieves comparable performance to the 2D LOO classifier using only a fraction of training images. This is a clear indication that the composite representation is an effective way of combining the 2D texture information and the $2\frac{1}{2}$D depth information. It also shows that this representation treats the systematic variability in the face images as such rather than as random variability. The $2\frac{1}{2}$D disparity representation is also compared with the standard

depth representation, and it is shown that the disparity representation is superior when the depth in computed through the noisy stereoscopic process. This also greatly widens the scope for using multiple images from arbitrarily placed cameras since no knowledge of the camera matrices is required.

CHAPTER 11

Conclusion and Future Directions

The objective for this work was to investigate the usefulness of depth information in improving face recognition accuracy in the presence of noise.

A stereo-based system, inspired by the human binocular vision, was devised using a pair of manually calibrated digital off-the-shelf cameras in a stereo setup to compute depth information. Depth values extracted from 2D intensity images using stereoscopy are extremely noisy, and as a result this approach for face recognition is rare. Recognition experiments are performed on the Sheffield Dataset, consisting 692 images of 22 individuals with varying pose, illumination and expressions. Stereo image pairs from this dataset were matched and sub-pixel disparity information for each pair in the dataset was computed. Psychology literature elucidated that although depth information is crucial to the way humans recognise faces, this information is perceived through *disparity* values rather than actual depth values in the form of familiar structure and texture information seen in 2D images. Hence, disparity information can be thought of as neural proxy for depth. Computationally too, depth and disparity are proportional up to camera parameters. Initially, disparity values and camera parameters are used to compute depth values, which are subsequently used to construct $2\frac{1}{2}$D images (depth maps) and 3D wire-frame models of the facial surfaces. Turk and Pentland's (Turk & Pentland 1991*a*,*b*) Eigenfaces algorithm and Spies and Ricketts' (Spies & Ricketts 2000) Fourier space based nearest-neighbours algorithm (Fourier K-NN) are used as benchmark algorithms throughout this work.

Shaded wire-frame (polygonal mesh) models are used to test a signature-based 3D surface matching algorithm for face recognition. Johnson's (Johnson 1997) spin-image based representation and recognition algorithm is an object-centred representation, in which a 2D intensity image (a spin-image) is generated for each point on the polygonal mesh. Two objects are matched by comparing a large number of randomly chosen spin-images. Although this technique has been tested in the presence of noise in the form of clutter and occlusions, it was unable to cope with the noisy models used in this work. The noise is introduced in these models during the camera calibration, triangulation and mesh pre-processing stages. Real-time processing was hindered by the size of the models (number of mesh faces) and the nearest-neighbours type approach to recognition. Extremely low recognition accuracies were obtained during leave-one-out cross validation on the models in dataset D_1 and consequently,

the approach was abandoned.

Also proposed are two new image representations: $2\frac{1}{2}$D disparity images and composite images. $2\frac{1}{2}$D disparity images, unlike the $2\frac{1}{2}$D depth images commonly found in literature, use disparity values as a representation of depth information. This representation has many advantages including the avoidance of error-prone camera calibration and triangulation procedures, the potential to be used with arbitrary camera placements and higher recognition accuracies compared with the equivalent depth images generated using stereoscopic methods. Composite images combine 2D texture and 3D shape information was investigated in an attempt to utilise the most discriminant information from both these modalities. It is obtained by concatenating the two images results in the highest recognition accuracies for the Eigenfaces classifier, when used in conjunction with the Mahalanobis distance.

Finally, the Eigenfaces classifier's performance was also tested using wavelets-decomposed 2D and $2\frac{1}{2}$D images. This led to some improvement in the recognition accuracy and the processing time was reduced significantly.

11.1 Conclusions

Some of the important conclusions that can be drawn from the experiments conducted in this work are listed below:

- Depth information in the form of disparity values plays a crucial role in machine recognition of faces. Disparity information significantly enhances the performance of a face recognition system when the stereo images are captured in a partially controlled environment and display varying pose, illumination, expressions.

- In the presence of noisy or inaccurate 3D data, using 2D intensity images and established 2D face recognition algorithms results in more accurate recognition. Reconstructing depth from disparity led to a loss of discriminatory information in noisy environments. This also explained why this approach is rarely seen in 3D face recognition literature. Despite marked improvement in recognition accuracies when compared with $2\frac{1}{2}$D depth images and 3D images (using spin-image representation), the classifiers still performed worse than the 2D images.

- A multi-modal Eigenfaces-based system incorporating both 2D texture and 3D shape information in the form of composite images proposed in this work, results in better classifier performance than either modality by itself. This representation also captures the systematic variability arising from representing the 3D shape of a face by a 2D illumination intensity matrix. This variability is treated as random in most existing systems by collecting numerous samples of the face in different pose and under varying imaging conditions.

- Fourier K-NN is a powerful algorithm and typically performs better than Eigenfaces for small datasets. As the size of the dataset increases, so does the processing time since the test image needs to be compared with all the existing images in the dataset. This approach does not handle the systematic variability very well since large amounts of training data are required. $2\frac{1}{2}$D complex disparity images result in the optimum classifier performance with this algorithm rather than the 2D or the composite images.

In the composite space, the Fourier classifier's performance was constrained by its performance in the 2D space, although the classifier was more accurate in the $2\frac{1}{2}$D space. Further investigations were unable to highlight the reasons for this peculiar phenomenon.

- The novel representations used in this work, particularly the $2\frac{1}{2}$D disparity images, are powerful enough to discern between the images of the same individual with and without glasses. However, the classifiers fail to establish the similarities in the faces in the presence of more gross changes such as addition or removal of head-scarf. Hence, it may be better to pre-process the images in a similar manner to the FERET datasets.

- The use of wavelets based pre-processing using Magarey and Kingsbury's MKC-4 wavelets (Magarey & Kingsbury 1995, 1996, Magarey 1997, Magarey & Dick 1998, Magarey & Kingsbury 1998*b*) marginally improves the recognition accuracy and the storage and the processing time requirements are reduced significantly.

- For reconstructing surfaces that do not contain sharp features (e.g. faces), image matching techniques that result in dense disparity maps are better than feature based approaches.

11.2 Contributions

The major contributions of this work to face recognition using depth information are listed below:

- **Disparity based $2\frac{1}{2}$D image representation**
  This work is the first to use disparity values to represent 3D shape. Disparity information is used to compute the actual depth values in existing systems. Experiments showed that the transformation from depth to disparity caused a loss of discriminatory information due to the presence of noise in the camera parameters and triangulation. Hence, these error-prone processes are by-passed and pixel intensities are replaced with disparity values in $2\frac{1}{2}$D images. This representation has a distinct advantage over depth based systems when deployed in noisy, uncontrolled environments and has the potential to be used with arbitrary camera placements, thus widening the scope of face recognition applications.

- **Composite Images**
  In general, multi-modal systems incorporating both shape and texture information process the data from both these modalities separately, and then fuse the recognition scores to assign an identity to the input image. In this work, the texture and disparity information are combined in a simple yet effective manner by concatenating the two matrices to form composite images. Thus the information from both these modalities can be processed simultaneously. Such an approach is more in line with how humans process information from multiple cues (Todd 2002) and results in greater classifier accuracy than either modality by itself. This is also confirmed by Bowyer's findings (Bowyer et al. 2004).

- **Capturing the systematic variability**
  In typical optical image based face recognition systems, the systematic variability arising from representing the 3D shape of a face by a 2D illumination intensity matrix is treated as random, by collecting multiple images of the face with varying pose and expression. The composite image based classifier, trained using only 5 images per class, is able to achieve the leave-one-out cross validation accuracy of a 2D intensity image based classifier (approximately 18 images per class). Thus, the composite representation captures the systematic variability in the appearance of the face and clearly highlights the importance of using both shape and texture information in face recognition.

- **Quantitative analysis of image matching algorithms**
  Pan's and Magarey's complex wavelets based image matching algorithms are both evaluated qualitatively and quantitatively. Qualitative analysis is done by inspection, as is common in literature. Both algorithms are quantitatively analysed using Lin and Barron's (Lin & Barron 1994) Backward Image Reconstruction. To the author's knowledge, this the only work to evaluate these algorithms quantitatively using real (rather than synthetic) face images. Magarey's algorithm, used with Magarey and Kingsbury's complex wavelets (MKC-4) gave superior results to Pan's algorithm which was tested using MKC-4 and Symmetric Complex Daubechies wavelets SCD-4 and SCD-6.

- **3D face recognition using spin images**
  Spin-image representation and recognition algorithm of Johnson (Johnson 1997) are used in this work to represent and recognise 3D face images. This approach has so far been used to classify rigid objects such as toys and plumbing equipment and for terrain images. The results obtained with the 3D face models from the Sheffield Database were disappointing and it was concluded that this approach is not suitable for 3D face recognition in a noisy environment.

- **$2\frac{1}{2}$D face recognition in the Fourier space using complex images** Horizontal and vertical disparity values, obtained during the image matching process are represented as the real and imaginary components of a complex number (Magarey 1997). Spies and Ricketts' Fourier space based nearest neighbours algorithm (Fourier K-NN) and Turk and Pentland's Eigenfaces algorithm are both applied to this complex disparity image matrix. The results for the Eigenfaces algorithm were disappointing. However, the Fourier K-NN algorithm's performance is optimal with this complex representation as a wider range of frequencies are used to represent this shape information. This is the only work to apply this algorithm to complex disparity images for face recognition.

- **Wavelets based pre-processing of 2D and $2\frac{1}{2}$D images**
  Although wavelet decomposition is frequently used in image processing and vision tasks, it is not often used for pre-processing face images for recognition. Magarey and Kingsbury's MKC-4 complex wavelets (Magarey & Kingsbury 1995, 1996, Magarey 1997, Magarey & Dick 1998, Magarey & Kingsbury 1998*b*) are used in this work and compared against the benchmark Haar (or the Daubechies-1) wavelet, which has been used to this end with some promising results. Experiments showed a marginal improvement in the performance of classifiers using MKC-4 processed images. These images are smaller than the original and this leads to a reduction in the storage requirements and processing time.

11.3 Future Directions

This work presented a stereo-based face recognition system using a pair of digital off-the-shelf cameras. Although promising results have been obtained with the Sheffield Database, much research still needs to be conducted before an identification system based on this work can be deployed in a non-co-operative scenario. Keeping this long-term goal in mind, the future research should be administered in the following areas:

- **Image matching**
  During the image matching process, correspondences are sought in the reference image (say, right) for every every pixel in the current image (say, left). The reference and the current images are swapped and the process is repeated to obtain left-right (LR) and right-left (RL) disparities. Future work should investigate ways of combining the output of the two disparity images to obtain more robust estimates. Normalisation and pre-processing techniques should also be investigated with the aim of reducing incorrect matches.

- **Arbitrary camera placements**
  In order to use the disparity maps from arbitrarily placed stereo cameras, the algorithm needs to be robust in the presence of asymmetric illumination between the two (or more) stereo images. This scenario is not investigated in this work owing to time constraints. Future work using the $2\frac{1}{2}$D and composite images should investigate the algorithm's invariance to changes in illumination. Further investigation is also required in combining disparity information from multiple cameras before such a system can be deployed in a real-world setting.

- **Disparity maps of varying resolutions**
  A disparity map is generated at each decomposition level for each of the image pairs. The non-interpolated disparity map from the lowest decomposition level (128×128) is used in this work to perform recognition. Reduction in resolution does not result in poorer recognition accuracy and the matching process is much faster due the reduced image size. The accuracy of the classifiers should be tested with the lower resolution disparity maps from higher decomposition levels to ascertain the decomposition level at which an acceptable balance between recognition accuracy and speed is reached.

- **Depth information**
  Experiments in this work showed that using disparity instead of depth information leads to increased classifier accuracy. However, the depth information used in this work is generated using noisy camera calibration parameters and a simple triangulation algorithm. Comparisons between disparity values and depth values obtained with more sophisticated algorithms should be compared.

- **Recognition algorithms**
  The Eigenfaces algorithm used in this work is the original proposed by Turk and Pentland in (Turk & Pentland 1991*a*,*b*) and is used since it is a standard benchmark in face recognition. However, since 1991, many variants of this algorithm have emerged, with different levels of success. Bayesian approach to PCA based face recognition, proposed by Moghaddam et. al at MIT (Moghaddam & Pentland 1997, Moghaddam et al. 1998,

1999) was one of the most successful algorithms at the recent FERET trials. Investigations using some of the more accurate variants of the Eigenfaces algorithm should be conducted to see if lower error rates can be achieved. Variants of the Fourier K-NN algorithm and ways of combining it with the Eigenfaces approach should be investigated to see if the classifier performance can be boosted further.

- **Canonical faces**
  Experiments showed that the benchmark algorithms, when presented with the images of an individual with and without head-scarf, were oblivious to the similarities in the face. Therefore, there is a definite case for the use of pre-processing to obtain canonical faces so that the face images only contain the oval of the face. It is expected that this will increase classifier accuracy.

- **Bigger dataset**
  The Sheffield Dataset, although a good starting point, is still very small compared to other publicly available databases such as the FERET. It can be improved by adding images:

  - of more subjects
  - taken outdoors
  - with minor/major changes in cosmetics
  - of subjects with sunglasses, head-scarves, balaclavas, caps, hats, "hoodies", etc.

- **Dynamic inputs**
  The long term goal of this work was to extend the face recognition system to using dynamic inputs such as those obtained from CCTV cameras. This would involve detecting and tracking faces correctly in two or more cameras, pre-processing them and computing the disparity information prior to the identification stage.

- **Multi-modal identification system**
  In a non-co-operative environment, this would involve using a face recognition system in conjunction with gait recognition, for example. Gait recognition involves recognising an individual from the way in which they walk. Again, like face recognition, this has the advantage of being non-invasive, non-intrusive and covert but is also subject to change. Finger-print or iris recognition are the ideal candidates for use in a co-operative environment.

APPENDIX A

Wavelets

A.1 Introduction

In the last two decades, wavelets have evolved from a popular field of study to a well established branch of mathematical analysis (Kobayashi 2000). A wavelet transform decomposes a signal into different frequency components which can then be analysed at a resolution that is appropriate to its scale. Wavelets theory is well established for one and two dimensional data and it is possible to extend it to three (and higher) dimensions. In two dimensions, the wavelet transform decomposes a digital image into a set of subimages with different characteristic orientations and scales.

Wavelets are closely linked with signal processing and the knowledge and understanding of filters and filter-banks is essential in understanding and designing wavelets with desirable properties. The subject of filters and filter-banks will not be reviewed here. The interested reader is directed to (Strang & Nguyen 1997) for an excellent exposition of the subject, leading to the development of the wavelet theory.

Up until recently, all the wavelets that were being designed were based on real-valued filter coefficients. However, these have many shortcomings when they are used in the two-dimensional space to analyse images. Consequently, new wavelets, based on complex-valued filter coefficients are now emerging.

A.2 Wavelets and Fourier Transforms: An Introduction

Wavelets are a way of representing and transforming a signal, very much like the classical Fourier Transforms. It is important to stress at this stage that wavelets are an alternative to the Fourier transforms rather than a replacement. The choice of the transform depends ultimately on the application, the signal itself and its bandwidth (Strang & Nguyen 1997).

Fourier Analysis decomposes a given signal into sinusoids of varying frequencies - it is a way of transforming raw data from the time domain into the frequency domain. However, it has a major drawback - Fourier transforms only have frequency resolution. They lack in time resolution, so, although it is possible to determine all the frequencies present in a signal,

it is not possible to establish the exact time when they occur (Valens 1999, Winkler n.d., Polikar 1999). Fourier techniques analyse the *total* frequency content of a signal using infinite exponential waves (Castellano 1999). Consequently, they only give satisfactory results for stationary signals. The results are extremely poor for transient signals and signals with drifts, trends and abrupt changes (Winkler n.d.). In short, these techniques are not suitable for detecting changes in signals.

The short-time Fourier Transform (STFT) is a joint time-frequency representation that cuts the signal of interest into several parts and then analyses the parts separately. It provides information about when an event in a signal occurred, and the frequency content of the event by dividing the signal into small segments and treating each of these segments as stationary. This clearly gives more information, but it also raises the issue of how to cut up the signal (Valens 1999). A window of constant length (duration) is used for the entire signal, both low and high frequency regions, which makes it very inefficient. If the chosen window is too narrow, it is no longer possible to know the exact frequency components that exist in the signal. Only the band of frequencies that exist is distinguishable. Conversely, if the window is too wide then the stationarity assumption is violated and the time resolution is lost as it is in the Fourier transform (Polikar 1999). The Gabor transform is an example of the STFT, which uses a Gaussian kernel as the window function. Gabor proved that with this window, the STFT achieved the best joint time-frequency localisation (Castellano 1999).

The shortcomings of the STFT led to the development of the Wavelet Transform - a *localised, scale-independent* technique, which analyses a signal at different time locations with kernels of varying sizes (Castellano 1999). These kernels are formed by translations and dilations of a prototype function called the *mother wavelet*.

A.3 Wavelet Theory

In wavelet analysis, the use of a fully scalable modulated window solves the dilemma of how to cut the signal. The window is shifted along the signal and for every position the spectrum is calculated. This process is repeated many times with a slightly shorter (or longer) window for every new cycle. The end result is a collection of time-frequency representations of the signal, all with different resolutions, hence the idea of *Multiresolution* being associated with wavelet analysis. Wavelet analysis does not have time-frequency representation but rather a time-scale representation (Valens 1999).

Figure A.1: Wavelet Transform: An efficient representation in both time and frequency domains

In wavelet decomposition, an input signal $f(t)$ is projected onto a family of functions which are the dilations and translations of a unique function $\psi(t)$ (Pan 1996*b*). This function is called *mother wavelet* and it is a complex-valued window function. $\psi(t)$ is scale dependent. It is made scale-independent by considering *all possible scalings of* $\psi(t)$. Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet (Graps 1995).

$$CWT_f(\tau,\alpha) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{+\infty} f(t)\psi\Big(\frac{t-\tau}{\alpha}\Big)dt \tag{A.1}$$

The Continuous Wavelet Transform (CWT), CWT_f is a time-scale representation of f and α is the *scale* parameter (Magarey 1997). Large values of α are indicative of long basis functions and hence larger scale features of f. τ is the translation factor and helps to achieve time-localisation. This adaptivity to scale α and translation τ leads to its good locality in both frequency and spatial domains - a property desirable in image matching algorithms (Pan 1996*b*). The $\frac{1}{\sqrt{\alpha}}$ ensures energy normalisation at various scales.

It is easy to see that there is a large degree of redundancy in the CWT representation of the signal. This can be addressed by constraining τ and α to take on discrete values only, which leads to the definition of the Discrete Wavelet Transform (DWT).

A.4 Discrete Wavelet Transform

Discrete wavelet transform discretises both α and τ such that

$$\alpha = 2^j \quad \text{and} \quad \tau = k, \qquad j,k \in \mathbb{Z} \tag{A.2}$$

where $\mathbb{Z}$ is the set of integers. This dyadic sampling makes the DWT more efficient and just as accurate as the CWT. The DWT of a signal can now be written as a set of integer-indexed coefficients:

$$c_f(j,k) = \int_{-\infty}^{+\infty} f(t)\psi_k^{(j)}(t)dt \tag{A.3}$$

$$\psi_k^{(j)}(t) = \frac{1}{\sqrt{2^j}}\psi\Big(\frac{t-2^jk}{2^j}\Big) \tag{A.4}$$

A.5 Wavelets and Filter Banks

Writing equations A.3 and A.4 as

$$c_f(j,k) = \Big(x * \psi^{(j)}\Big)(t)\Big|_{t=2^jk} \tag{A.5}$$

$$\text{where} \qquad \psi^{(j)}(t) = 2^{-\frac{j}{2}}\psi\Big(-\frac{t}{2^j}\Big), \quad j \in \mathbb{Z} \tag{A.6}$$

demonstrates that the wavelet coefficients $c_f(j,k)$ can be interpreted as the sampled outputs of a bank of filters whose impulse responses are scaled versions of a mother wavelet.

The Discrete Time Wavelet Transform (DTWT) corresponds to a filter bank iterated a finite number of times along the low pass channel (Cooklev et al. 2000). For discrete signals

$x(n)$, it is defined as

$$y^{(j)}(n) = \sum_k x(k)\psi^{(j)}(2^j n - k) \tag{A.7}$$

$$\text{or } y^{(j)}(n) = \left(x * \psi^{(j)}\right)(n) \downarrow_{2^j} \quad j = 1, 2, \ldots \tag{A.8}$$

where $\downarrow_n$ refers to the downsampling-by-n operation.

The DTWT is an efficient representation in the sense that it has no redundancy. This is true if the filter bank that implements it is a maximally decimated multirate filter bank, i.e. it represents the signal $x(n)$ using the same number of coefficients $y^{(j)}(n)$ as the original signal has samples (Magarey 1997).

A.6 The subband decomposition tree

The dyadic DTWT provides information about the detail contained in a signal at many different scales. In the frequency domain, this is same as analysing the signal into octave frequency bands. In such a subband filtering scheme, the input signal $x(n)$ is analysed by a pair of halfband filters, one highpass and one lowpass. The highpass filter h_1 provides the first level of detail, i.e. the coefficients $y^{(1)}(n)$ (after downsampling by 2), while the downsampled lowpass output $\hat{y}^{(1)}$ becomes a coarse approximation to $x(n)$, with half the resolution, but double the scale (Magarey 1997). The basic 2-band structure is shown in figure A.2. The approximated signal is again analysed by the two filters, resulting in the second level of detail coefficients, and another approximation signal. This process is repeated iteratively till the desired maximum level of decomposition m_{max} is reached. The resulting structure (figure A.2) is known as the the *subband decomposition tree.*

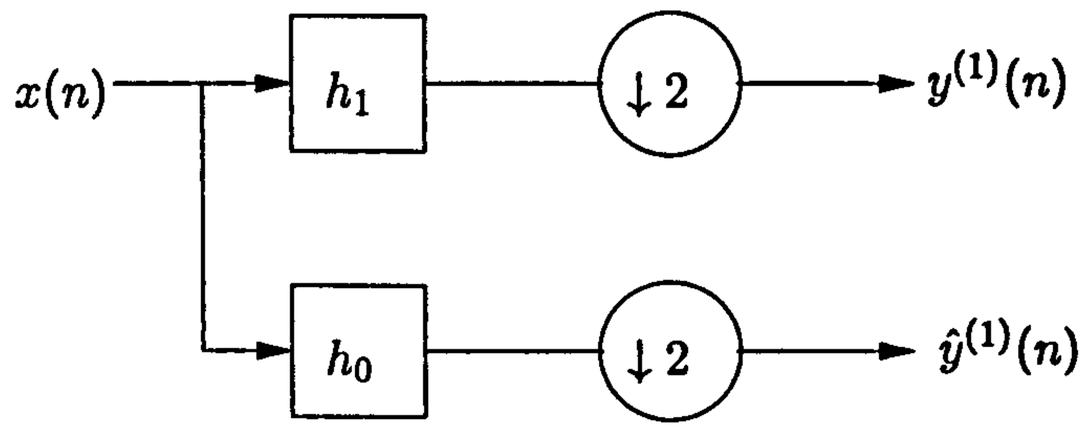


Figure A.2: Two-band building block for dyadic DWT

At the maximum level of decomposition m, there are m set of detail signals $\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$ and a remainder signal $\hat{y}^{(m)}$, which is a coarse, low-resolution approximation to the original signal x. If the lowpass and the highpass filters h_0 and h_1 are ideal halfband filters, no information is lost and the input signal can be reconstructed exactly. See (Strang & Nguyen 1997) for a detailed exposition on the links between wavelets and filter-banks.

The subband decomposition tree also links in with the idea of *Multiresolution Analysis* (MRA), one of the most desirable and versatile properties of the wavelet analysis (Misiti et al. 2000). The concept of MRA was introduced by Mallat in (Mallat 1989*b*). He showed the suitability of wavelet bases to represent the difference in information between approximation of a signal at different resolutions.

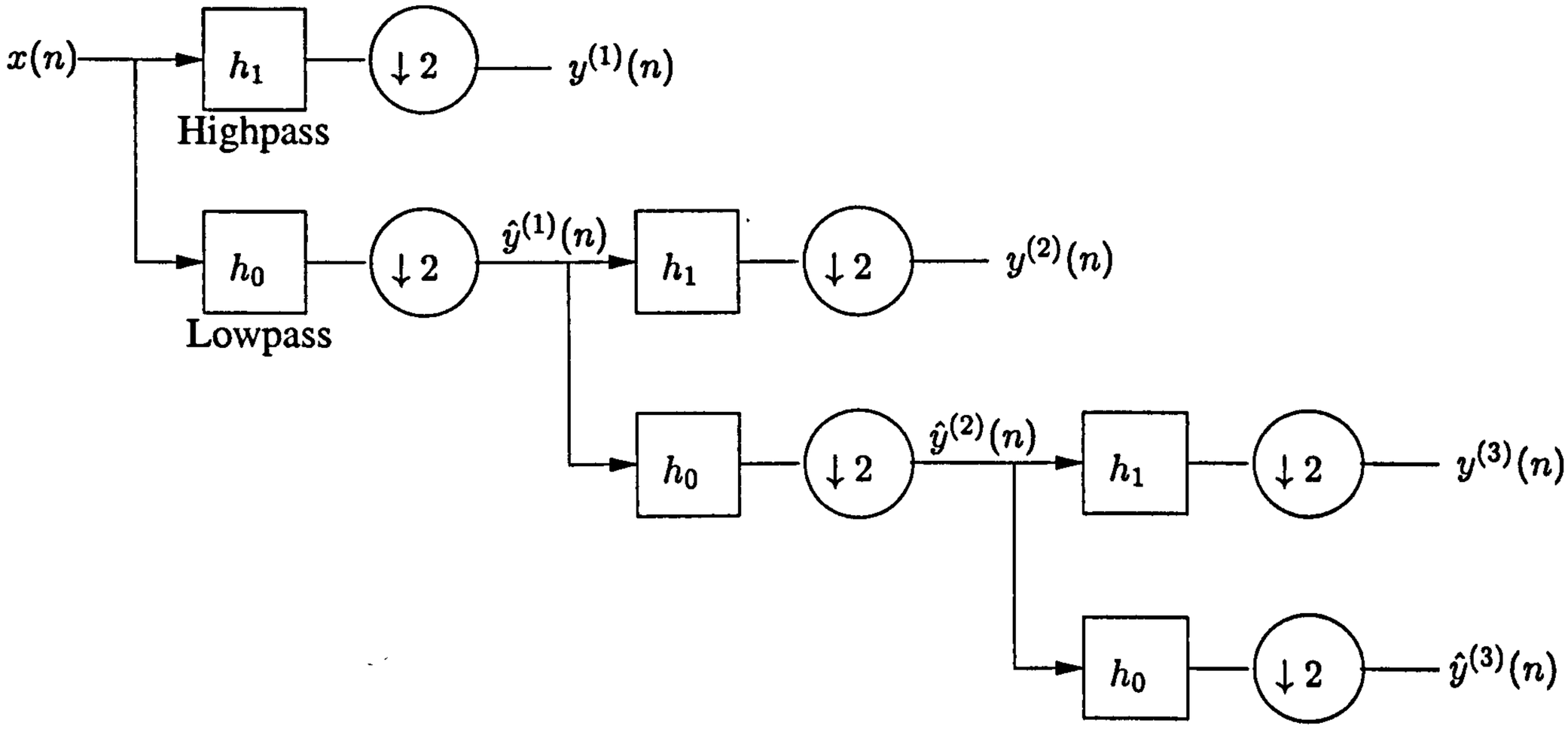


Figure A.3: Subband Decomposition Tree (3 levels). The output $\hat{y}^{(m)}$ of the lowpass channel is decomposed further at each level to generate the coefficients at the next coarser level $m+1$.

A.7 Two-Dimensional Wavelets

In 2D, the DWT is implemented most efficiently using a 2D *separable* filter. Separable image transforms are implemented using 1D convolutions and by applying the 1D building block (see figure A.2) first down the columns and then along the rows of the two resulting half-size images. The 2D building block is depicted in figure A.4

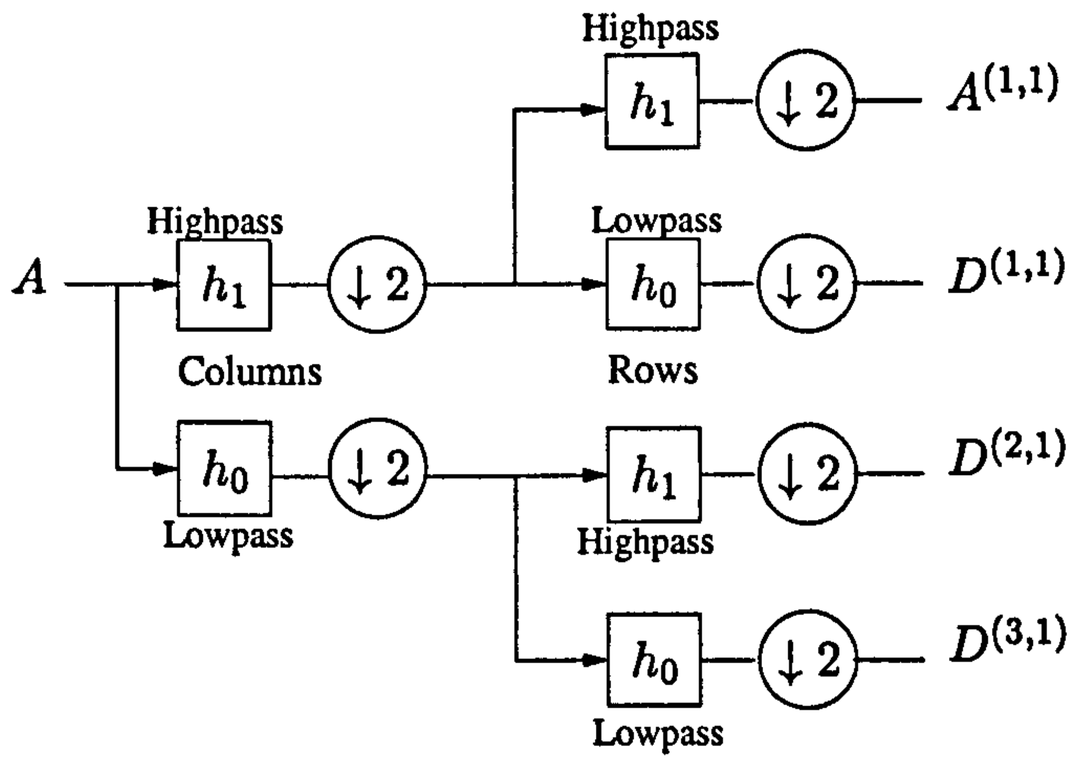


Figure A.4: Building block for separable DWT on a 2D input image A

Similarly to the 1D equations in A.8, the level m approximation and detail coefficients

can be written as:

$$D^{(n,m)}(\mathbf{n}) = \sum_{\mathbf{k}} A(\mathbf{k})\psi^{(n,m)}(2^m\mathbf{n} - \mathbf{k}) \tag{A.9}$$

$$A^{(m)}(\mathbf{n}) = \sum_{\mathbf{k}} A(\mathbf{k})\phi^{(m)}(2^m\mathbf{n} - \mathbf{k}) \tag{A.10}$$

where $\psi^{(n,m)}$ is the wavelet filter associated with subband (n, m) and $\phi^{(m)}$ is the level m scaling filter. Because of the separable nature of the transform, the 2D wavelets and scaling filters can be implemented as tensor products of 1D functions:

$$\psi^{(1,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\phi^{(m)}(n_2) \tag{A.11}$$

$$\psi^{(2,m)}(\mathbf{n}) = \phi^{(m)}(n_1)\psi^{(m)}(n_2) \tag{A.12}$$

$$\psi^{(3,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\psi^{(m)}(n_2) \tag{A.13}$$

$$\phi^{(m)}(\mathbf{n}) = \phi^{(m)}(n_1)\phi^{(m)}(n_2) \tag{A.14}$$

These filters partition the unit frequency cell as shown in figure A.5. Their orientational emphasis corresponds to their position in the frequency cell. For example, $\psi^{(1,m)}$, being lowpass in the horizontal direction but highpass in the vertical, emphasises horizontal edges. $\psi^{(1,m)}$, $\psi^{(2,m)}$ and $\psi^{(3,m)}$ capture the vertical, horizontal and diagonal (corners) features of the image. The coefficients $D^{(n,m)}(\mathbf{n})$ contain information from heavily overlapping blocks of input pixels, each spatially filtered to emphasise a specific orientation and scale (Magarey 1997).

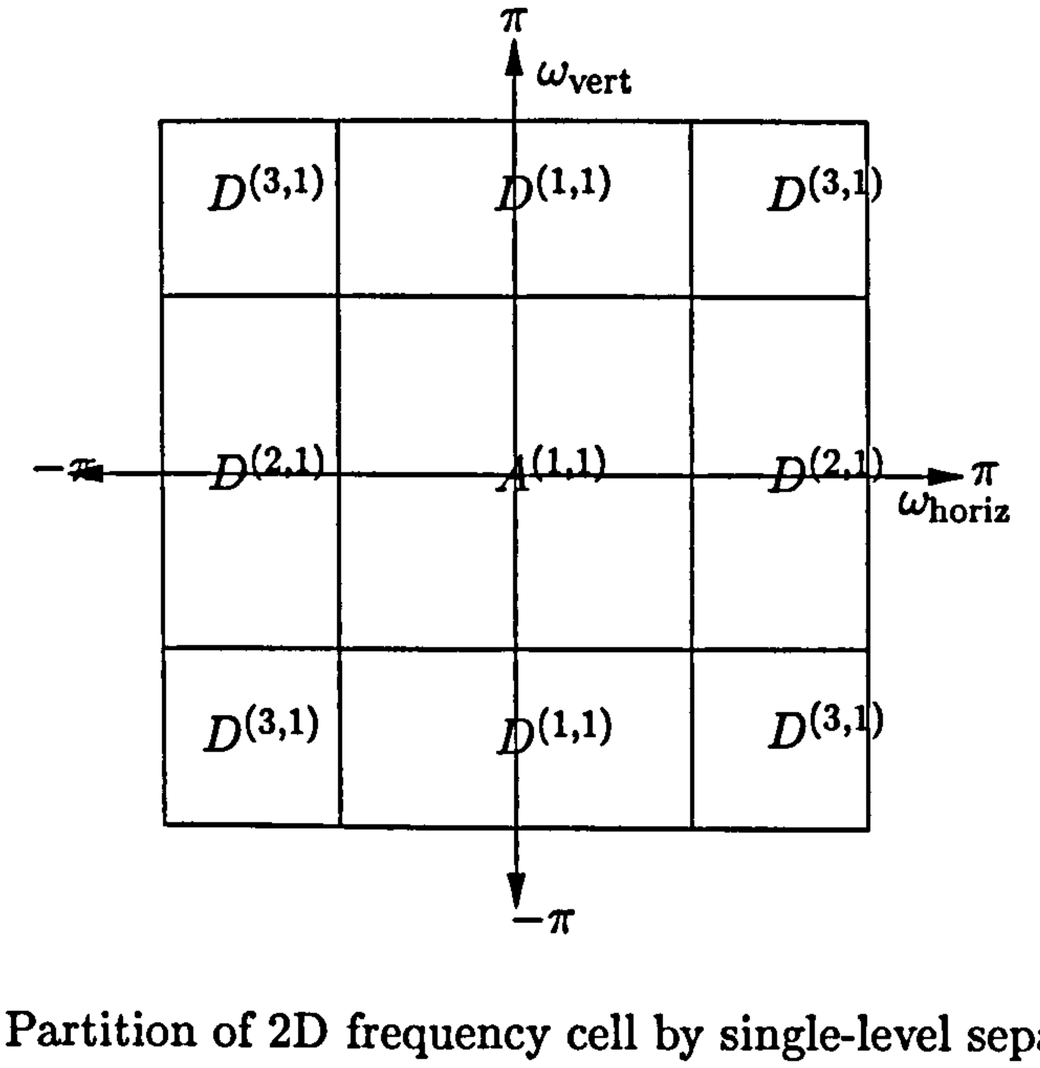


Figure A.5: Partition of 2D frequency cell by single-level separable DWT

Figure A.6 depicts the implementation of a 2D wavelet transform. First, each row of the image undergoes decomposition into its high and low-pass components. The resulting images' horizontal resolution is reduced by a factor of 2 and their scale is doubled. Then, both, the high and low-pass subimages are each separately filtered column-wise to obtain four row-column filtered subimages.

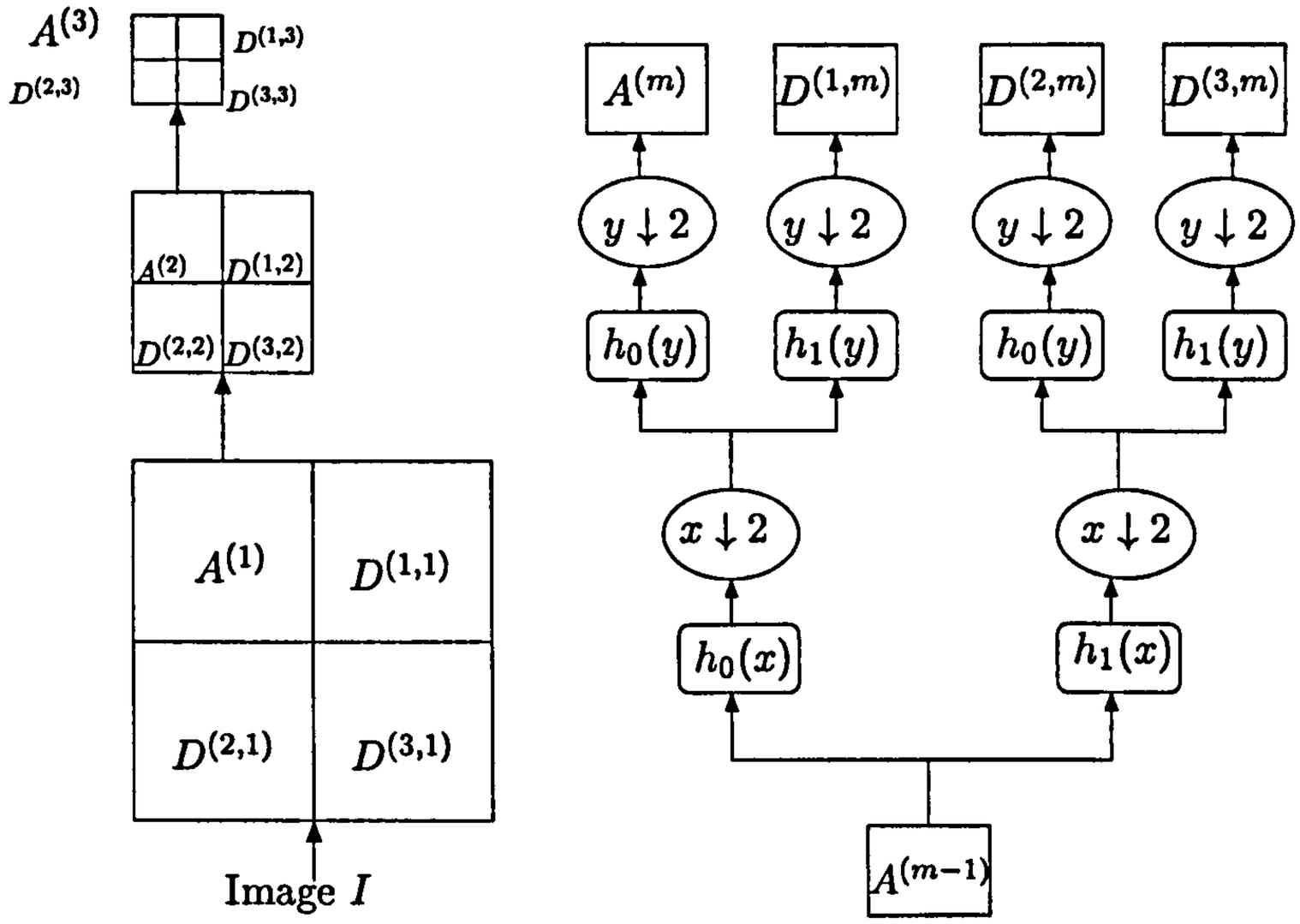


Figure A.6: Wavelet image pyramid obtained by 2D DWT (left) and Flow-chart of the 2D DWT, from level $m-1$ to level m, implemented using the high-pass filter h_1 and the low-pass filter h_0 (right).

A.8 Complex Discrete Wavelet Transform

The complex discrete wavelet transform (CDWT) is implemented using a pair of complex valued filter pair $\{h_0, h_1\}$ and was designed to address some of the short-comings of the real DWT such as lack of shift invariance and poor directional selectivity.

Details of the complex wavelets used in this work are presented along with the algorithms that use them. Symmetric Daubechies Complex Wavelets (SCD-4 and SCD-6) and Magarey & Kingsbury's complex wavelets (MKC-4) are described briefly in Appendix C and a detailed mathematical exposition of the MKC-4 wavelets is presented in Appendix D.

Note: To be consistent with Magarey's (Magarey 1997) and Castellano's (Castellano 1999) notation, spatial co-ordinates $\mathbf{n} = (n_1, n_2)^\mathsf{T}$) are used, with vertical listed first. Positive directions are down and to the right.

A.9 Summary

The basics of 1D and 2D wavelet transforms have been presented in this appendix. Wavelets, like Fourier transforms, decompose a signal into different frequency components. But unlike Fourier transforms, wavelets have both frequency and scale resolution (analogous to time resolution). Signals are decomposed into shifted and scaled versions of the mother wavelet. This results in a collection of time-frequency representations of the signal, all with different resolutions. This affords many advantages in terms of signal analysis, the main one being that of multiresolution analysis. They allow the use of long time intervals where more precise

low-frequency information is required, and shorter regions where high frequency information is required. The other major advantage of wavelet analysis is the facility to perform *local analysis* - that is, to analyse a localised area of a larger signal. Wavelet analysis is capable of revealing aspects of data such as trends, breakdown points, discontinuities in higher derivatives, and self similarity, which are often not highlighted by other techniques.

Wavelet analysis is implemented using pairs of high and low-pass filters in a filterbank configuration. At each level of decomposition, the input signal is decomposed into a low-pass Approximation of the signal and a high-pass Detail signal. The Approximation signal is again analysed using the filters to obtain the decomposition at the next coarser level. Wavelet techniques are economical in their representation because there is no redundancy or overlapping in the representation across the scales. This is because the wavelet decomposition is with respect to an orthonormal basis - the approximation and the detail sequences are both uncorrelated and half the size of the original sequence, which implies that the size of the decomposed representation is equal to the size of the original sequence (Prasad & Iyengar 1997).

Wavelet analysis in 2D is executed in a similar way using a separable filter pair. These are implemented using 1D convolutions by applying the filters first to the columns of the 2D signal (image) and then to the rows. This results in an Approximation image and three Detail images that capture the horizontal, vertical and diagonal features of the images.

APPENDIX B

Magarey & Kingsbury's Wavelet - MKC-4

B.1 Introduction

The complex wavelets designed by Magarey and Kingsbury (Magarey & Kingsbury 1995, 1996, Magarey 1997, Magarey & Kingsbury 1998*a*,*b*), referred to as MKC-4 wavelets, have been used extensively in this work. Both the image matching algorithms crucial to this work employ these wavelets with varying results. This Appendix describes MKC-4 wavelets in some detail.

B.2 1-Dimensional CDWT

The Complex Discrete Wavelet Transform (CDWT) is based on a pair of even-length FIR filters $\{h_0, h_1\}$ which may be modelled as Gabor filters

$$h_0(n) \approx a_0 e^{-\frac{(n-n_0)^2}{2\sigma_0^2}} e^{i\omega_0(n-n_0)} \tag{B.1}$$

$$h_1(n) \approx a_1 e^{-\frac{(n-n_0)^2}{2\sigma_1^2}} e^{i\omega_1(n-n_0)} \tag{B.2}$$

$$\text{for } n = \quad -D, \ldots, D-1$$

with n_0 set to $-\frac{1}{2}$ in order to position the Gaussian window symmetrically in the interval $[-D, D-1]$. ω_0 and ω_1 are the centre frequencies.

The 1D CDWT is implemented using these filters in the standard subband decomposition tree (figure A.3) (Magarey & Kingsbury 1998*b*). The bandpass coefficients $y^{(m)}$ at level m can be thought of as the downsampled output of a convolution with an equivalent wavelet filter (equation B.5), while the lowpass coefficients $\hat{y}^{(m)}$ are obtained using a scaling filter

(equation B.6:

$$y^{(m)}(n) = \sum_k x(k)\psi^{(m)}(2^m n - k) \tag{B.3}$$

$$\hat{y}^{(m)}(n) = \sum_k x(k)\phi^{(m)}(2^m n - k) \tag{B.4}$$

For a particular choice of the parameters $a_0, a_1, \sigma_0, \sigma_1, \omega_0, \omega_1$, the equivalent wavelet and scaling filters can be approximated as Gabor filters (Magarey 1997):

$$\psi^{(m)}(n) \approx a_m e^{\frac{(n-n_m)}{2\sigma_m^2}} e^{j\omega_m(n-n_m)} \tag{B.5}$$

$$\phi^{(m)}(n) \approx \hat{a}_m e^{\frac{(n-n_m)}{2\sigma_m^2}} e^{j\hat{\omega}_m(n-n_m)} \tag{B.6}$$

$$\text{for } n = -(2^m-1)D, \ldots, (2^j-1)(D-1) \tag{B.7}$$

The parameters of $\psi^{(m)}$ and $\phi^{(m)}$ can be calculated from the parameters of h_0 and h_1. If $\sigma_0, \sigma_1, \omega_0, \omega_1$ are correctly chosen, the set of filters $\{\phi^{(m)}, \psi^{(j)}, j = 1, \ldots, m\}$ will adequately cover the range $[0, \pi]$.

Of particular interest is the behaviour of $2^m\omega_m$ and $2^m\hat{\omega}_m$ as m varies. As m gets large, they converge to constants, but for the first few levels, they exhibit significant variation. As a result, a prefilter f is applied to the input *before* the first level of the DWT tree. The prefilter is defined by:

$$(h_0 * f)(2n) = \lambda f(n) \tag{B.8}$$

$$\text{where } \lambda = \frac{1}{2}|H_0(\omega)|_{\omega=0} \tag{B.9}$$

$H_0(\omega)$ is the Fourier transform of the filter h_0. The prefilter f is carefully chosen to simulate the lowpass branch of an infinitely deep DWT tree, so that after the first 2-band split, the equivalent wavelet and scaling filters have converged to their final behaviour (Magarey 1997, Magarey & Kingsbury 1998*b*). Without the prefilter, the complex wavelet filters given by equations B.5 and B.6 have the same characteristics as the real-valued wavelet filters in terms of *not* being perfectly scaled versions of one another (Castellano 1999). However, perfect scaling is a desirable property for the image matching application, and this is achieved through the use of the prefilter. The decomposition is now a *perfectly scaled wavelet decomposition.* It is implemented as the standard DWT, except that the filters for the first level are $h_{0f} = h_0 * f$ and $h_{1f} = h_1 * f$, instead of h_0 and h_1. The subsequent levels of the tree are computed by applying h_0 and h_1 as usual.

The actual filters used are 4-tap filters with rational complex-valued coefficients:

$$h_0 = \Big[1-j \quad 4-j \quad 4+j \quad 1+j\Big]/10 \tag{B.10}$$

$$h_1 = \Big[-1-2j \quad 5+2j \quad -5+2j \quad 1-2j\Big]/14 \tag{B.11}$$

Their Gabor parameters are (Magarey & Kingsbury 1998*b*):

$$\begin{aligned} \omega_0 &= \pi/6 & \omega_1 &= 0.76\pi \\ \sigma_0 &= 0.97 & \sigma_1 &= 1.07 \\ a_0 &= 0.47 & a_1 &= 0.43j \end{aligned}$$

B.3 2-Dimensional CDWT

In the 2D space, the CDWT is implemented separably as the real-valued DWT. The 1D CDWT wavelet filters have significant magnitude response only in the range $[0, \pi]$. Therefore, if the 2D CDWT is implemented in the same manner, the equivalent wavelet filters will cover only the first quadrant of the unit frequency cell (see figure A.5). However, real-valued images contain significant information in the first and second quadrants of the unit frequency cell (the third and fourth quadrants are conjugated versions of the first and second). Hence, the 2D separable implementation needs to be modified to capture this information.

In (Magarey 1997) Magarey supplements the complex filters h_0 and h_1 with the conjugates of these filters since this reflects their magnitude frequency responses about $\omega = 0$. Hence, the conjugated filters cover the frequency range $[\pi, 0]$ and there is no loss of information from the images. The same separable 2D CDWT implementation is used for the conjugate filters - the conjugate filters are applied to the columns first and then the rows of the resulting images. The resulting band-pass images are denoted by $\{D^{(4,1)}, D^{(5,1)}, D^{(6,1)}\}$ In addition, there is a second low-pass image at each level of decomposition, which is labelled $A^{(2,m)}$ (since the first one is labelled $A^{(1,m)}$). The 2D CDWT is shown in figure B.1

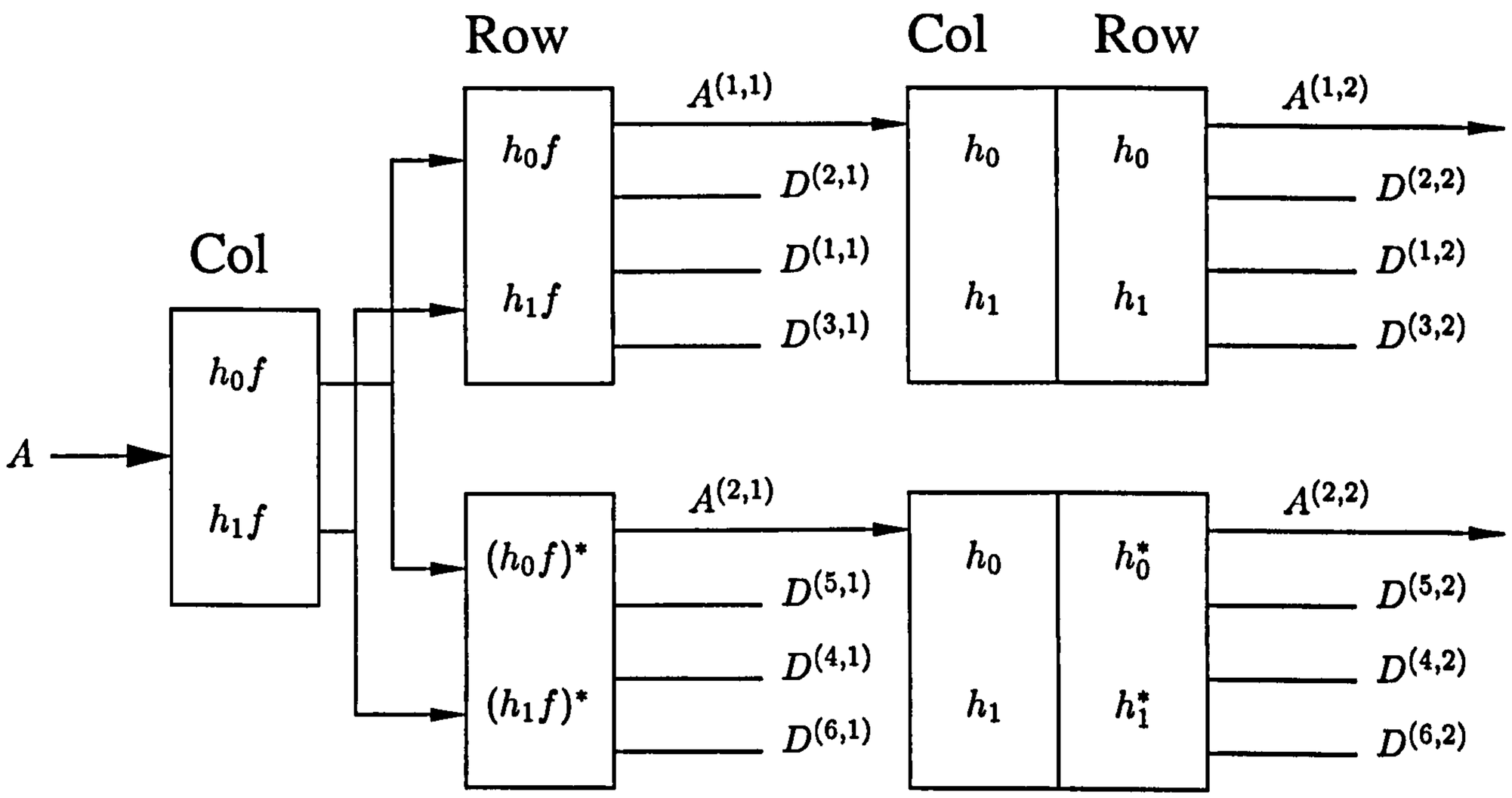


Figure B.1: 2D CDWT. Note the use of f-modified filters in the first level. (Each convolution is followed by a downsampling)

Eight complex subimages are produced, each a quarter of the size of the original and hence there is a 4 : 1 redundancy. Subsequent stages do not increase the redundancy, because each stage takes two complex images and produces eight subimages of one quarter the size. The over all redundancy of the 2D CDWT is therefore 4 : 1, regardless fo the depth of the tree. This redundancy is vital for the *interpolability* of the transform and is a crucial feature of the image matching algorithm (see section D.2).

Note: From here onwards all the equivalent wavelet and scaling filters considered are

f-modified.

Each subimage $D^{(n,m)}$ has a corresponding wavelet filter $\psi^{(n,m)}$. Equations B.12 to ?? illustrate the relationship between the first quadrant and second quadrant filters respectively and the 1D wavelet and scaling filters $\psi^{(m)}$ and $\phi^{(m)}$. The second quadrant filters are obtained by applying h_0^* and h_1^*, and can be written:

$$\psi^{(1,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\phi^{(m)}(n_2) \quad \text{(B.12)}$$
$$\psi^{(2,m)}(\mathbf{n}) = \phi^{(m)}(n_1)\psi^{(m)}(n_2) \quad \text{(B.13)}$$
$$\psi^{(3,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\psi^{(m)}(n_2) \quad \text{(B.14)}$$
$$\phi^{(1,m)}(\mathbf{n}) = \phi^{(m)}(n_1)\phi^{(m)}(n_2) \quad \text{(B.15)}$$
$$\psi^{(4,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\phi^{*(m)}(n_2) \quad \text{(B.16)}$$
$$\psi^{(5,m)}(\mathbf{n}) = \phi^{(m)}(n_1)\psi^{*(m)}(n_2) \quad \text{(B.17)}$$
$$\psi^{(6,m)}(\mathbf{n}) = \psi^{(m)}(n_1)\psi^{*(m)}(n_2) \quad \text{(B.18)}$$
$$\phi^{(2,m)}(\mathbf{n}) = \phi^{(m)}(n_1)\phi^{*(m)}(n_2) \quad \text{(B.19)}$$

The 2D wavelet filters are also Gabor like (since they are the products of 1D Gabor filters) and can be approximated as:

$$\psi^{(n,m)}(\mathbf{n}) \approx a^{(n,m)} N(\mathbf{n}|\mathbf{n_m}, \Lambda_{n,m}) e^{j\Omega^{(n,m)}.(\mathbf{n}-\mathbf{n_m})} \quad \text{(B.20)}$$

where $N(\mathbf{n}|\mathbf{m}, \Lambda)$ is an un-normalised bivariate Gaussian in $\mathbf{n}$ with mean $\mathbf{n}_m$ and covariance Λ:

$$N(\mathbf{n}|\mathbf{m}, \Lambda) = exp\Big(-\frac{1}{2}(\mathbf{n}-\mathbf{m})^\top \Lambda^{-1}(\mathbf{n}-\mathbf{m})\Big) \quad \text{(B.21)}$$

The parameters can be derived from those of h_0 and h_1. The centre frequencies specifying the orientation of the wavelet filters is given by:

$$\mathbf{\Omega}^{(1,m)} = (\omega_m, \hat{\omega}_m)^\top \quad \text{(B.22)}$$
$$\mathbf{\Omega}^{(2,m)} = (\hat{\omega}_m, \omega_m)^\top \quad \text{(B.23)}$$
$$\mathbf{\Omega}^{(3,m)} = (\omega_m, \omega_m)^\top \quad \text{(B.24)}$$
$$\mathbf{\Omega}^{(4,m)} = (\omega_m, -\hat{\omega}_m)^\top \quad \text{(B.25)}$$
$$\mathbf{\Omega}^{(5,m)} = (\hat{\omega}_m, -\omega_m)^\top \quad \text{(B.26)}$$
$$\mathbf{\Omega}^{(6,m)} = (\omega_m, -\omega_m)^\top \quad \text{(B.27)}$$

Since the CDWT is to be used to estimate motion, it is paramount that no information is lost when transforming to the CDWT domain. To facilitate this, the chosen set of 2D wavelet filters should tile the upper half of the unit frequency cell as evenly and as completely as possible. Because of the iterative structure of the CDWT, this can be ensured by placing the six filters $\psi^{(n,m)}$ evenly around the edge of the rectangle $[-2\pi/2^m, 2\pi/2^m] \times [0, 2\pi/2^m]$. Figure B.2 shows that most even tiling is achieved by setting $\omega_m/\hat{\omega}_m$ equal to 3. Since the tree in consideration is a perfectly scaled tree, this can be achieved all levels of decomposition m by simply setting ω_{1f}/ω_{0f} to 3, and this in turn can be achieved by appropriate choice of the parameters $\{\omega_0, \omega_1, \sigma_0, \sigma_1\}$.

It is preferable if a near uniform spacing of orientations between 0 and π can be achieved, i.e. a spacing of $\pi/6$. Magarey found in (Magarey 1997) that this can be achieved by setting

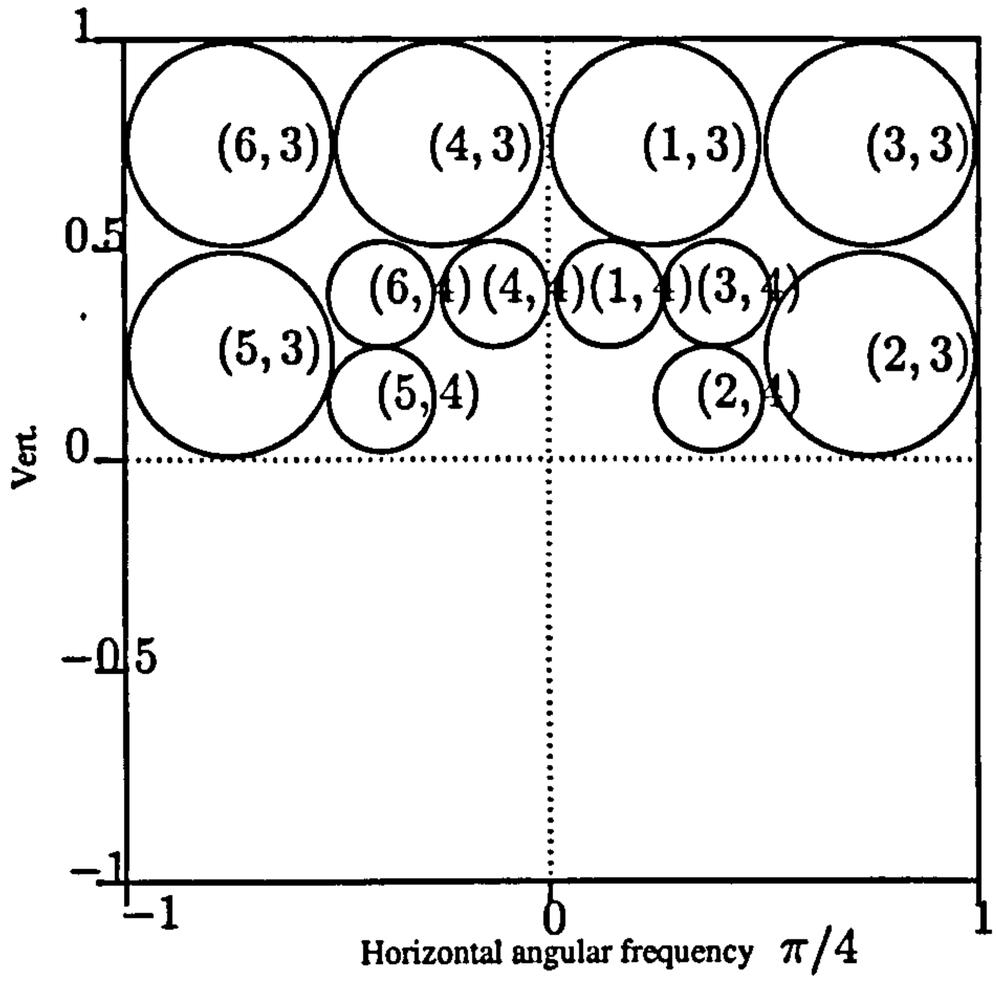


Figure B.2: Circular contours of the magnitude responses of the 2D CDWT wavelet filters at levels 3 and 4 derived from the Gabor pairh_0 and h_1 with parameters $a_0 = 0.39$, $a_1 = 0.39j$, $\omega_0 = \pi/6$ and $\omega_1 = 5\pi/6$, $\sigma_0 = \sigma_1 = 1.27$ and $D = 4$

$\omega_m/\hat{\omega}_m$ equal to 3.73. However, this conflicts with the previous requirement for an even tiling, and a value somewhere between these must be chosen.

The contours of the Fourier transform $|\Psi^{(n,m)}(\Omega)|$ of $\psi^{(n,m)}$ are elliptical, centred on $\Omega^{(n,m)}$, with shape determined by $\Lambda_{n,m}$ (Magarey 1997, Magarey & Dick 1998). Each coefficient $D^{(n,m)}$ of the Detail or the bandpass images is derived from the contributing region of pixels in the original image. The size of the contributing region depends on the spatial extent of the corresponding filter $\psi^{(n,m)}$. Because of the Gabor-like nature of the wavelet filters $\psi^{(n,m)}$, the region contributing to $\mathbf{n}$ can be thought of as an elliptical area centred on pixel $2^m\mathbf{n} - \mathbf{n_m}$, with axes specified by $\Lambda_{n,m}$. This property is crucial to the matching criterion used in Magarey's matching algorithm, described in detail in Appendix D.

B.4 Summary

Magarey and Kingsbury's complex-valued 2-channel wavelets have been described in this Appendix. First, the 1D CDWT using Gabor-like filters is described. The theory is then extended to 2D real-valued signals (images). Properties of this wavelet particularly useful the image matching application and Magarey's algorithm are also highlighted.

APPENDIX C

Pan's Uniform Full-Information Image Matching Algorithm

In (Pan 1996*b*,*a*), Pan presents *uniform full-information image matching* technique. This technique is described briefly and is presented in pseudo-code form in Chapter 6. The results of its application to matching face images are also presented in the same chapter. This chapter presents the mathematical details for the algorithm and follows a similar structure to Pan's papers (Pan 1996*b*,*a*).

C.1 Uniform Full-Information Image Matching

A digital image is a function $f(x, y)$ in the 2D space. For the purposes of image matching, a new representation of $f(x, y)$ should be chosen such that the constructs in this new representation harness the salient information contained in the original image $f(x, y)$.

Assume that $f(x, y)$ is to be represented by a vector of projections of $f(x, y)$ onto n basis function $\psi_j(x, y)$

$$f(x, y) \longrightarrow (a_1, a_2, \ldots, a_n) \tag{C.1}$$

$$a_j = \langle f(x, y), \psi_j(x, y) \rangle, \quad j = 1, 2, \ldots, n \tag{C.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two functions, and a_j's are called representation coefficients. Appropriate choice of the basis functions ψ_j's results in the salient information of $f(x, y)$ to be encoded in the representation coefficients, a_j's. If the coefficients a_j's computed via equation C.2 result in

$$f(x, y) = F(a_1, a_2, \ldots, a_n; \psi_1, \psi_2, \ldots, \psi_n) \tag{C.3}$$

being true, then equation C.1 is said to be a *full information* representation. Note that $F()$ is a computable function. An example would be when ψ_j's constitute an orthonormal basis. In this case, the reconstruction can be achieved using

$$f(x, y) = \sum_{j=1}^{n} a_j \psi_j(x, y)$$

A representation of the form in equations C.1 and C.3 is said to be *uniform* because each representation coefficient a_j is defined and computed with exactly the same mathematical form of equation C.2 (Pan 1996*b*).

Pan states that for image matching, it is desirable if the properties of equation C.3 include good dimensional orthogonality, discriminative uniqueness, space-frequency locality, multiresolution adaptivity, and computational efficiency and robustness (Pan 1996*b*).

C.2 Similarity Distance and Continuous Matching

Pan defines a similarity distance measure for any pair of points in the reference image (e.g. left) and the matched image (e.g. right) in terms of the complex conjugate wavelet analysis.

C.2.1 Notation

- **Standard Similarity Distance:** $SB_j((x,y),(x',y'))$
- **Generalised Similarity Distance:** $S_j((x,y),(x',y'))$
- **Reference Image:** $f(x,y)$
- **Matched Image:** $f'(x',y')$
- **Approximation sub-matrices at j^{th} level of decomposition:** $A^{(1,j)}$ and $A^{(2,j)}$. Note that the use of two channel complex wavelets such as MKC-4 results in twice as many coefficients: 2 Approximation coefficient matrices and 6 Detail coefficient matrices. If single channel wavelets are used then there is only 1 Approximation and 6 Detail images.
- **Detail sub-matrices at j^{th} level of decomposition:** $D^{(1,j)}$, $D^{(2,j)}$, $D^{(3,j)}$, $D^{(4,j)}$, $D^{(5,j)}$ and $D^{(6,j)}$.

C.3 Implicit Feature Vectors

It is assumed for the remainder of this Appendix that 2-channel complex wavelet such as the MKC-4 is used, and the theory is presented here with that in mind. If this is not the case, then only the available coefficients are used.

There are 8 complex wavelet analysis coefficient matrices for each level of the wavelet pyramid: $A^{(1,j)}$, $A^{(2,j)}$, $D^{(1,j)}$, $D^{(2,j)}$, $D^{(3,j)}$, $D^{(4,j)}$, $D^{(5,j)}$ and $D^{(6,j)}$. The 2 approximation components, $A^{(1,j)}$ and $A^{(2,j)}$, are further decomposed at each level to obtain the coefficients for the next level of the wavelet pyramid.

The standard case of similarity distance measure for matching the two images uses the detail coefficients only. The *implicit feature vector* $\mathbf{B}_j(x,y)$ for each position (x,y) is given

as

$$\mathbf{B}_j(x,y) = \left(\frac{D^{(1,j)}(x,y)}{|A^{(1,j)}|}, \frac{D^{(2,j)}(x,y)}{|A^{(1,j)}|}, \frac{D^{(3,j)}(x,y)}{|A^{(1,j)}|}, \frac{D^{(4,j)}(x,y)}{|A^{(2,j)}|}, \frac{D^{(5,j)}(x,y)}{|A^{(2,j)}|}, \frac{D^{(6,j)}(x,y)}{|A^{(2,j)}|} \right)$$

with

$$\mathbf{B}_{p,j} = \frac{D^{(p,j)}(x,y)}{|A^{(q,j)}(x,y)|}, \quad p = 1,\ldots,6 \quad \text{and} \quad q = \begin{cases} 1 & \text{if } p = 1,2,3, \\ 2 & \text{if } p = 4,5,6. \end{cases} \tag{C.4}$$

where $|\cdot|$ denotes the L_2-norm. This feature vector also incorporates normalisation in order to insure the matching process against local image intensity variations.

C.4 Standard Similarity Distance Measure

If the relative rotation angle γ is small enough, then the similarity distance measure $SB_j((x,y),(x',y'))$ can be defined as

$$SB_j((x,y),(x',y')) = \sum_{p,j}^{6} SB_{p,j}((x,y),(x',y')) \tag{C.5}$$

where $SB_{p,j}$'s are the subband similarity distances, defined by

$$SB_{p,j}((x,y),(x',y')) = |B_{p,j}(x,y) - B'_{p,j}(x',y')|^2 \tag{C.6}$$

For an image point $(x,y) = (k,l)$ on the reference image, its precise correspondence (x',y') lies somewhere around the integer positions (m',n').

$$f(k,l) \mapsto f'(m'+u, n'+v) \tag{C.7}$$

where $(u,v) \in \mathbf{R}^2$ and denotes the differences:

$$u = x' - m', \quad v = y' - n' \tag{C.8}$$

The similarity distance measure is now reformed to:

$$SB_{p,j}((k,l),(m'+u,n'+v)) = |B_{p,j}(k,l) - B'_{p,j}(m'+u,n'+v)|^2 \tag{C.9}$$

The best matching point is the one that minimises this distance measure.

$$\min_{u,v} SB_{p,j}((k,l),(m'+u,n'+v)) \tag{C.10}$$

Note that in the above formulations, the position (x,y) refers to the continuous 2D space. However, $B_{p,j}(x,y)$'s are available only at discrete positions $(x = k, y = l | (k,l) \in \mathbf{Z}^2)$. Therefore, the wavelet subbands are continuously interpolated in order to minimise $SB_{p,j}$ of C.6. Detailed exposition of the continuous interpolation process can be found in (Pan 1996*b*).

C.5 Local Parallax Continuity and Generic Pattern Matching

For the robustness of image matching, Pan assumes local parallax continuity, i.e. the parallax field (or the disparity field) in a neighbourhood of two neighbouring points will be nearly the same. To maintain a compromise between fine locality of matching and robustness of matching, a generalised similarity distance measure, in the generic pattern matching sense is defined. Let $\mathcal{N}$ denote a minimum neighbourhood containing the central position and its four closest diagonal positions:

$$\mathcal{N} = \{(0,0),(-0.5,-0.5),(-0.5,0.5),(0.5,-0.5),(0.5,0.5)\} \tag{C.11}$$

Let $\mathbf{PA}_j(k,l)$ and $\mathbf{PD}_j(k,l)$ denote the approximation and the detail subband pattern vector on an integer-indexed position (k,l):

$$\mathbf{PA}_{i,j}(k,l) = [A^{(i,j)}(k+r,l+c) \quad | \quad i=1,2; \quad (r,c) \in \mathcal{N}] \tag{C.12}$$

$$\mathbf{PD}_j(k,l) = [B_{p,j}(k+r,l+c) \quad | \quad p=1,\ldots,6; \quad (r,c) \in \mathcal{N}] \tag{C.13}$$

Note that:

- $\mathbf{PA}_{1,j}$ and $\mathbf{PA}_{2,j}$ are complex conjugates.
- $(k+r,l+c)$ with $\{(r,c) \in \mathcal{N} - (0,0)\}$ corresponds to diagonal positions which can also be computed with rigorous bottom-up wavelet transform.
- An integer-indexed position (k,l) may be matched with another integer-indexed position (m',n') or a diagonal position $(m'+r,n'+c)$.
- In order to use the generalised similarity distance measure $S_j((k,l),(m',n'))$, two wavelet pyramids need to be prepared: one on integer-indexed positions (k,l) and another one on diagonal positions $(k+r,l+c)$. This eliminates the need for continuous interpolation of wavelet subbands.

The generalised similarity distance measure $S_j = ((k,l),(m',n'))$ is defined as

$$S_j((k,l),(m',n')) = SA_j((k,l),(m',n')) * SD_j((k,l),(m',n')) \tag{C.14}$$

where $SA_j((k,l),(m',n'))$ and $SD_j((k,l),(m',n'))$ are the similarity distances in terms of approximation and detail subbands respectively,

$$SA_j((k,l),(m',n')) = 1 - \frac{|PA_{1,j}(k,l) * PA_{2,j}(m',n')|}{|PA_{1,j}(k,l)||PA_{2,j}(m',n')|} \tag{C.15}$$

$$SD_j((k,l),(m',n')) = |\mathbf{PD}_j(k,l) - \mathbf{PD}'_j(m'n')| \tag{C.16}$$

Note that Pan uses full information here - approximation subbands and normalised detail subbands are used for defining the similarity distance. Pan also suggests an alternative set of approximation and detail subbands similarity measures. However, he ascertained through experimental work that this alternative set of measures reduces the efficiency but not the gross errors. Consequently, these measures are not detailed here. The interested reader is directed to (Pan 1996*a*) for more information.

C.6 Matching Two Images

For any image point (k,l) on the reference image, its approximate correspondence (k'_0,l'_0) on the matched image is obtained through spiral and hierarchical parallax propagation strategies. The identification of the precise correspondence (k',l') is achieved via the discrete search in a small neighbourhood of (k'_0,l'_0). This neighbourhood is defined by a distance threshold T_r.

$$\min_{(k',l')} \quad S_j((k,l),(k',l')) \quad \forall (k',l'): \quad |(k',l')-(k'_0,l'_0)| \leq T_r \tag{C.17}$$

The distance threshold T_r should be defined in such a way that the allowed errors in the parallax propagated from the last higher level can be corrected. In general $1 \leq T_r \leq 2$. Note that the search takes place not only on integer positions, but also on diagonal positions.

C.7 Spiral and Hierarchical Parallax Propagation

Without loss of generality, only stereo pairs of square images with size $2^n \times 2^n$ are considered. Let M_j denote the parallax vector field on the j^{th} level of the image pyramid for image $f(x,y)$,

$$M_j = (\mathrm{M}_j(k,l)), \quad k,l = -m,-m+1,\ldots,-1,+1,\ldots,m-1,m, \quad \text{with } m = 2^{n-j-1} \tag{C.18}$$

where each element $\mathrm{M}_j(k,l)$ contains parallax vector for a pair of homologous image points $(x=k,y=l)$ and (x',y') on the j^{th} level

$$\mathrm{M}_j(k,l) = (x'-k, y'-l) \tag{C.19}$$

In case there is no correspondence for $(x=k,y=l)$, $\mathrm{M}_j(k,l) = \emptyset$. With a minimum overlap of 60%, the central area on the coarsest reference image (i.e. at the highest level of decomposition) is guaranteed to have a correspondence on the coarsest matched image. The search for potential matches for the central area starts with an exhaustive search at the coarsest level (the smallest decomposed image). The similarity distance defined by equation C.14 yields an approximate parallax vector $\mathrm{M}_j(0,0)$ for the central area.

Using the assumption of local parallax continuity, the parallax vector for each integer-indexed position of this central area is initiated as

$$\mathrm{M}_j(k,l) \approx \mathrm{M}_j(0,0), \quad \{(k,l) \in \mathcal{N} - (0,0)\} \tag{C.20}$$

This can also be fine-tuned using equation C.14. Spiral propagation is used to initialise the unknown parallax field. Known parallax vectors are propagated, from the central area to the outer rings, ring by ring, until the boundary of partial correspondence is reached. Gross errors of the resultant parallax field can be detected and corrected automatically using the local continuity constraint

$$|\mathrm{M}_j(k,l) - \overline{\mathrm{M}}_j(k,l)| \leq T_M \tag{C.21}$$

where $\overline{\mathrm{M}}_j(k,l)$ denotes the mean (or median for robustness) parallax field vector on the smallest (e.g. 4-connected) neighbourhood centred on the position (k,l) of level j. T_M denotes the maximal allowed parallax difference usually $1 \leq T_M \leq 2$.

After image matching on a higher $(j+1)^{\text{th}}$ level, the parallax field is then propagated to the next lower (finer) j^{th} level (hierarchical propagation). The initial parallax field on the

current j^{th} level can be obtained by interpolating the parallax field at the $(j+1)^{\text{th}}$ level. The inverse of the similarity distance of equation C.14 for each position on the higher level may be taken as the weighting factor for linear or nonlinear interpolation. The matched parallax field $\mathbf{M}_j$ on each j^{th} level yields pairs of matched homologous image points. These, along with the relative imaging geometry (obtained through camera calibration), are used to reconstruct the object surfaces via triangulation.

C.8 Wavelets

Pan describes the properties that are desirable in a wavelet family to achieve optimality in the context of the image matching task in (Pan 1996*a*). The three main properties, orthonormality, locality and symmetry are described briefly. Since the 2D wavelets are implemented as separable filters, Pan states that it suffices to study the 1D scaling and wavelet functions ϕ and ψ, where

$$\phi(x) = \sqrt{2}\sum_k h_k \phi(2x-k) \tag{C.22}$$

and

$$\psi(x) = \sqrt{2}\sum_k g_k \phi(2x-k) \tag{C.23}$$

- **Orthonormality:** Orthogonality of the scaling and the wavelet functions ϕ and ψ guarantees that there is no loss of information and no redundancy in the representation of a particular input signal. In particular, the similarity distance measure $S_j((k,l),(m',n'))$ (equations C.6,C.14) is defined as a Euclidean distance, which requires the orthogonality of implicit feature space. The normality condition ensures uniform comparison in terms of similarity distance.

- **Locality:** Salient information is often contained in the local variations or the transient phenomena. This imposes two requirements on the chosen representation:

  1. The representation should be able to detect local changes in the input signal $f(x,y)$, and
  2. Abrupt changes in the input signal $f(x,y)$ over a small spatial span should result in changes in only a few of the wavelet coefficients.

  In the context of image matching, locality requires the scaling and the wavelet functions to have *compact support.* This means that only a limited number of h_k's are non-zero or non-vanishing:

  $$h_k \neq 0 \quad \forall k : -N \geq k \leq N+1$$

  Such scaling function has support length $L = 2N+2$. Locality in the frequency domain is reflected in the *smoothness* of the scaling and wavelet functions. Mathematically, a maximum number of *vanishing moments* for the scaling and wavelet functions is desirable. For a wavelet family with compact support of $2N+2$, the maximum number of vanishing moments is N.

- **Symmetry:** In signal processing, symmetry is necessary for linear phase of the filters. Linear phase is desirable in order to avoid phase compensation in pyramidal

filter structure and for reducing border effects by symmetric extension. The symmetry of the scaling and wavelet functions ϕ and ψ is equivalent to

$$h_k = h_{1-k}, \quad \text{for } k = 1, 2, \ldots N+1,$$

where N is an odd number. In image matching, this property is also a necessity in order to achieve a minimum basis of the rotation-invariance of the filters and in matching the homologous image neighbourhoods. It is important in particular in case the relative orientation of the two images is unknown prior to the image matching.

C.8.1 Real Wavelets

Pan goes on to analyse three of the real wavelet families with the above properties in mind. The Haar and the Daubechies-4 (Db4) Wavelets are both orthogonal and have compact support. The Haar has the most compact support of 2. The Db4 wavelets have the most compact support of 4 among all orthonormal wavelets. The scaling filters for the two are given by

$$h_0(\text{Haar}) = \left(\frac{1}{2}, \frac{1}{2}\right)$$

and

$$\text{h(Db4)} = \left(\frac{1+\sqrt{3}}{8}, \frac{3+\sqrt{3}}{8}, \frac{3-\sqrt{3}}{8}, \frac{1-\sqrt{3}}{8}\right)$$

Unfortunately, the Haar is too compactly supported so that not enough information can be extracted for similarity analysis of any given pair of homologous points on any given level and the Db4 wavelets are not symmetric. Also analysed is Symmlet-L family of wavelets. This family of wavelets satisfy the orthonormality property and have the maximum vanishing moments. They are known to be the "least asymmetric" wavelets with the compact support of L.

No real wavelet families satisfy all of these conditions for optimality. Lawton (Lawton 1993) showed that only complex valued scaling and wavelet filters and functions exist under the four stringent conditions of orthonormality, compact support, maximum vanishing moments and symmetry. He presents a method for constructing complex valued linear filters and associated wavelet bases. Lawton also derived a length-6 complex valued linear phase filter associated with real-valued Daubechies-6 wavelet bases.

Lina and Mayrand (Lina & Mayrand 1993) investigated the general complex solution of the four conditions using a particular parametrisation of the multiresolution analysis. They show that realness and symmetry are incompatible. Hence, all symmetric Daubechies wavelets are complex valued, and Lawton's complex valued Daubechies-6 wavelet is a particular example of the more general set of complex wavelets found by Lina and Mayrand.

C.8.2 Complex Wavelets

Pan (Pan 1996*a*) refers to the symmetric complex wavelet family associated with real Daubechies-L wavelet family the "Symmetric Complex Daubechies Wavelets" (SCD-L). He investigates two of these for use with his algorithm - SCD-4 and SCD-6. SCD-6 is the shortest wavelet that satisfies the conditions of orthonormality, vanishing moments and symmetry. However, Pan believes that SCD-6 is not compact enough for stereo image matching. SCD-4, although

symmetric, does not have vanishing moments. Because of its highly non-smooth nature, this particular wavelet has not received much attention in signal analysis and reconstruction applications. However, Pan felt it was worthwhile investigating whether it was useful and effective for the image matching application. The scaling filters for SCD-6 and SCD-4 are given by equations C.24 and C.25.

$$h_0(SCD\text{-}6) = \frac{1}{64}\left(-3-i\sqrt{15}, 5-i\sqrt{15}, 30+2i\sqrt{15}, 30+2i\sqrt{15}, 5-i\sqrt{15}, -3-i\sqrt{15}\right) \quad \text{(C.24)}$$

$$h_0(SCD-4) = \frac{1}{4}\left(1+i, 1-i, 1-i1+i\right) \quad \text{(C.25)}$$

In addition to the Symmetric Complex Daubechies Wavelets, Pan also investigates the complex wavelets designed by Magarey and Kingsbury (Magarey & Kingsbury 1995). These wavelets do not satisfy the orthogonality condition and are described in detail in Appendix B. These wavelets have a compact support of length 4, with $N = 2$. Pan uses an approximate version of these wavelets, with the lowpass (h_0) and the highpass (h_1) filters given by:

$$h_0 = \left[1-j \quad 4-j \quad 4+j \quad 1+j\right]/10 \quad \text{(C.26)}$$

$$h_1 = \left[-1-2j \quad 5+2j \quad -5+2j \quad 1-2j\right]/14 \quad \text{(C.27)}$$

APPENDIX D

Magarey's Motion Estimation using Complex Wavelets

D.1 Introduction

The mathematical details relating to Magarey's complex wavelets based motion estimation algorithm, described in Chapter 6 are provided in this Appendix.

Motion estimation starts at the coarsest level $m_{\max}$, producing a motion vector for each sub-pixel at this resolution. A coarse-to-fine hierarchical matching strategy is adopted. At each level of decomposition, the corresponding transform coefficients plus the estimates of the previous coarser level are utilised to produce a more dense and accurate motion field. This field is then smoothed and regularised using Anandan's approach (Anandan 1989). This process is repeated till the finest resolution is achieved at level m_{min}. At this level of decomposition the motion field has a density of $2^{-2m_{min}}$, i.e. one motion estimate for every $2^{m_{min}} \times 2^{m_{min}}$ block of input pixels. In order to obtain a motion vector for every pixel in the original image (full resolution motion field), the motion field at level m_{min} is upsampled and interpolated m_{min} times (Castellano 1999).

D.2 Coarse Level Estimation

The matching criterion is defined at subpixel $\mathbf{n}$ and sub-band (n, m), where n refers to the bandpass image number, and m is the decomposition level, as follows:

$$SD^{(n,m)}(\mathbf{n}, \mathbf{f}) = \frac{\left| D_1^{(n,m)}(\mathbf{n}, \mathbf{f}) - D_2^{(n,m)}(\mathbf{n}) \right|^2}{\sigma P^{(n,m)}} \tag{D.1}$$

$D_1^{(n,m)}(\mathbf{n})$ and $D_2^{(n,m)}(\mathbf{n})$ are the bandpass or the detail coefficients from the reference and the current image frames A_1 and A_2. $\mathbf{f}$ is the vector offset or the displacement at the subpixel $\mathbf{n}$ at scale m (or equivalently, a displacement of $2^m \times \mathbf{f}$ at the original resolution of the input images). $P^{(n,m)}$ eliminates the filter dependency when the different subbands

are combined in the hierarchical estimation. σ avoids the scale dependency of $P^{(n,m)}$ when combining the matching criteria over different scales. It has the effect of assigning the finer level information a progressively greater weight in the final motion estimate.[1]

$P^{(n,m)}$ is a weighting factor corresponding to the energy of the wavelet filter at subband n and scale m.

$$P^{(n,m)} = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \left| \Psi^{(n,m)}(\mathbf{\Omega}) \right|^2 d\mathbf{\Omega} \tag{D.2}$$

where $\Psi^{(n,m)}$ represents the Fourier transform of $\psi^{(n,m)}$.

Hierarchical matching algorithms work by accumulating disparity information at all levels of decomposition. This allows the more confident coarser-scale motion components to be passed down the hierarchy, unchanged into the aperture-affected regions. Hence, the aperture problem (see Section D.4) is addressed to an extent.

At the coarsest level of decomposition, there is no prior information available about the motion field, and the search region is centred on zero displacement:

$$\mathbf{f} \in [-f_{max}, f_{max}] \times [-f_{max}, f_{max}] \tag{D.3}$$

Estimating motion at vector offsets as well as integer valued coefficients gives sub-pixel accuracy at the coarsest level. Scaling and interpolating these sub-pixel values at finer resolutions results in a dense and accurate disparity field.

Summing over the six oriented subbands forms the *subband squared difference*, SSD:

$$SD^{(m)}(\mathbf{n}, \mathbf{f}) = \sum_{n=1} SD^{(n,m)}(\mathbf{n}, \mathbf{f}) \tag{D.4}$$

The motion estimate or the disparity at sub-pixel $\mathbf{n}$ is taken to be the location of $\mathbf{f}_0$, the minimum of $SD^{(m)}$. $\mathbf{f}_0$ is converted to the original pixel resolution by scaling it by 2^m. This is to allow for the downsampling that has taken place at level m. The maximum detectable motion is therefore $f_{max} \times 2^{m_{max}}$ pixels in each direction.

A distinctive feature of this matching criterion is that it is defined as a *surface* over a real 2D interval, $\mathbf{f}$, rather than as a set of values on an integer or half-integer interpolated grid. This is made possible by the *interpolability* of the CDWT, the ability to estimate non-integer-indexed coefficients $D_1^{(n,m)}(\mathbf{n}, \mathbf{f})$ from the known integer-indexed coefficients in the same subband. This property also makes it possible to estimate the location of the minimum of the SSD analytically in terms if the coefficients $\{D_1^{(n,m)}(\mathbf{n}), D_2^{(n,m)}(\mathbf{n}), n = 1, \ldots, 6\}$ with high precision, without having to conduct a computationally-intensive interpolated search (see Section B.3). The interpolability of the CDWT also links phase-correlation, phase matching and gradient-based estimation.

A further important property of these surfaces is that the steeper the surface is at the minimum point (large curvature parameters), the more precise (high confidence) the corresponding motion estimate is. Inversely, if the surface is flat in a particular direction (small curvature parameters), it highlights the unreliability (low confidence) in the component of the motion in that direction (Castellano 1999).

---

[1] Note that in his thesis (Magarey 1997), Magarey finds that the best value for $\sigma^{(m)}$ is $4\lambda^4$, with λ given by equation B.9.

D.3 Subband Coefficient Interpolation

The interpolability property (described in preceding Section) of linear transforms follows from another property, termed *shiftability* by Simoncelli et. al (Simoncelli et al. 1992). A transform is defined to be *shiftable* if the energy distribution amongst subbands is independent of shifts in the input signal. Real-valued critically-sampled transforms cannot be shiftable in this sense. However, the implicit 2 : 1 redundancy of the 1D CDWT provided by the complex basis filters enables this without the need for explicit oversampling. Since the CDWT approximately satisfies this criterion (Castellano 1999), $D^{(n,m)}(\mathbf{nf})$ can be written as

$$D^{(n,m)}(\mathbf{n},\mathbf{f}) \approx \sum_{\mathbf{k}} W_{\mathbf{f}}^{(n,m)}(\mathbf{k}) D^{(n,m)}(\mathbf{n}+\mathbf{k}) \tag{D.5}$$

$$W_{\mathbf{f}}^{(n,m)}(\mathbf{k}) = H_{\mathbf{f}}(-\mathbf{k}) e^{j2^m (\mathbf{\Omega}^{(n,m)})^{\mathsf{T}}(\mathbf{f}-\mathbf{k})} \tag{D.6}$$

This is based on the work by Fleet and Jepson (Fleet & Jepson 1990), who showed that it is possible to interpolate the downsampled outputs of complex bandpass filters by modulating a low-pass interpolation kernel $W_{\mathbf{f}}^{(n,m)}$ to the centre frequency of the equivalent wavelet filter, and convolving with the modulated kernel (equation D.5).

As the 2D CDWT is constructed separably, it follows that 2D interpolation can also be implemented separably. That is, the 2D interpolating low-pass kernel can be constructed as the product of two 1D kernels:

$$H_{\mathbf{f}}(\mathbf{k}) = hf_1(k_1) hf_2(k_2) \tag{D.7}$$

The range of $\mathbf{k}$ depends on the range of $\mathbf{f}$ values required. In practise, only a unit range (in each direction) needs to be considered, as the CDWT coefficients can be integer-shifted to cope with values that fall outside that range.

The simplest low-pass kernel is a delta function, corresponding to a "staircase" interpolation.

$$h_f(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for} f \in [-0.5, 0.5]. \tag{D.8}$$

This results in the simplest possible interpolation formula

$$D^{(n,m)}(\mathbf{n}+\mathbf{f}) \approx D^{(n,m)}(\mathbf{n}) e^{j\theta(\mathbf{f})} \tag{D.9}$$

$$\text{where } \theta(\mathbf{f}) = 2^m \left(\mathbf{\Omega}^{(n,m)}\right)^{\mathsf{T}} \mathbf{f} \tag{D.10}$$

$$\text{for } \mathbf{f} \in [-0.5, 0.5] \times [-0.5, 0.5]$$

(The factor 2^m in equation D.10 accounts for the downsampling in each direction). Equation D.9 gives the approximate relationship between small input translations and linear phase changes. This relationship may be modelled as a plane, whose gradient is determined by the centre frequency of the associated wavelet filter, scaled up by 2^m. It is assumed here that for small $\mathbf{f}$ shifts, the magnitude of the CDWT coefficients is approximately constant. Magarey finds in his thesis (Magarey 1997) that the model works well over the unit interval as long as the input image has no "strong spectral components in the passband of the filter, but is

reasonably spectrally flat". Hence, it is assumed that the input spectrum has no sharp peaks in the support region of the associated Gabor-like wavelet filter. Since the filter bandwidth decreases as m increases, this property is more readily satisfied at the higher pyramid levels, where accuracy is most important.

However, Magarey and Kingsbury's investigations, in common with those of Fleet and Jepson's (Fleet et al. 1991), found that the actual phase gradient, or *local frequency*, can vary considerably around the expected value. As a result, the algorithm accuracy is improved by explicitly estimating the phase gradient in equation D.10 at each subpixel n of each subband. The phase gradient can be estimated using the formula (Fleet & Jepson 1990) and equation D.5:

$$\nabla\phi^{(n,m)}(\mathbf{n}+\mathbf{f}) = \frac{\Im m\left\{D^{(n,m)*}(\mathbf{n}+\mathbf{f})\nabla D^{(n,m)}(\mathbf{n}+\mathbf{f})\right\}}{|D^{(n,m)}(\mathbf{n}+\mathbf{f})|^2} \tag{D.11}$$

$$\text{where } \nabla D^{(n,m)}(\mathbf{n}+\mathbf{f}) = \sum_{\mathbf{k}} \nabla W_{\mathbf{f}}^{(n,m)}(\mathbf{k}) D^{(n,m)}(\mathbf{n}+\mathbf{f}) \tag{D.12}$$

The accuracy of the algorithm is further increased by replacing the staircase interpolation kernel of equation D.8 with a 4-tap windowed-sinc kernel:

$$h_f(k) = \left(\frac{\cos\frac{\pi}{2}(f+k)}{1-(f+k)^2}\right)\left(\frac{\sin\pi(f+k)}{\pi(f+k)}\right), \quad k=-2,\ldots,1 \tag{D.13}$$

D.4 Quadratic SSD surfaces

Magarey shows in (Magarey 1997) that the disparity estimation using only one sub-pixel (see equation D.1) is sufficient in the CDWT domain, and averaging over a region of sub-pixels is not necessary. This is due to the fact that small translations in the input image result phase rotation in the CDWT domain, which is both predictable and signal independent, and its values can be derived from the CDWT coefficients. However, at least two distinct orientations are required to produce a unique estimate and all six are required to ensure robustness of the estimates. This is because each subband, rather than defining the actual displacement, defines only that component of the displacement that is *normal* to the orientation of the filter associated with the particular subband. This a manifestation of the *aperture problem* in the subband-phase-space. Aperture problem is also one of the contributing factors to the motion estimation problem being ill-posed (Magarey 1997).

The aperture problem in detecting the motion of a feature that varies along one spatial dimension only (either horizontally or vertically) arises because the vector component of motion parallel to the orientation of the feature has no effect on the image. If the true motion vector is decomposed into two components, parallel and perpendicular to the feature orientation, then only the orthogonal component can be detected (Hildreth 1983). More specifically, the component of the motion along the direction of an edge cannot be determined unless the size of the aperture of the analysing device is larger than the length of the edge (Castellano 1999). In the case of this algorithm, the size of the aperture is determined by the region of support of the wavelet filter (Castellano 1999).

A match between two sub-pixels is indicated by the minimum of the $SD^m(\mathbf{n},\mathbf{f})$ (equation D.4). The minimum of the SSD surface is located using the CDWT interpolation formulae.

First, expand the equation D.1:

$$SD^{(n,m)}(\mathbf{n},\mathbf{f}) = \left|D_1^{(n,m)}(\mathbf{n},\mathbf{f})\right|^2 + \left|D_2^{(n,m)}(\mathbf{n})\right|^2 - 2\Re e\left\{D_2^{*(n,m)}(\mathbf{n})D_1^{(n,m)}(\mathbf{n},\mathbf{f})\right\} \quad \text{(D.14)}$$

This implies that minimising the SSD is equivalent to maximising the cross-correlation of the two sub-images across all six orientation bands, assuming that $\left|D_1^{(n,m)}(\mathbf{n}+\mathbf{f})\right|^2$ remains relatively constant as $\mathbf{f}$ varies. This assumption is further justified by equation D.9 over a small range of $\mathbf{f}$ values about $(0,0)$. Further, this allows $SD^{(n,m)}(\mathbf{n},\mathbf{f})$ to be approximated as

$$SD^{(n,m)}(\mathbf{n},\mathbf{f}) \approx \left|D_1^{(n,m)}(\mathbf{n})\right|^2 + \left|D_2^{(n,m)}(\mathbf{n})\right|^2 - 2\left|D_1^{(n,m)}(\mathbf{n})D_2^{(n,m)}\right| \cos\left(\phi_1^{(n,m)}(\mathbf{n},\mathbf{f}) - \phi_2^{(n,m)}\right) \quad \text{for } \mathbf{f} \in [-0.5,0.5]\times[-0.5,0.5] \quad \text{(D.15)}$$

where $\phi_1^{(n,m)}(\mathbf{n},\mathbf{f})$ and $\phi_2^{(n,m)}(\mathbf{n})$ are the phases of $D_1^{(n,m)}(\mathbf{n},\mathbf{f})$ and $D_2^{(n,m)}(\mathbf{n})$ respectively. Hence, minimising equation D.8 is much like maximising the weighted phase correlation between the two sub-pixels. The weighting factor is given by the *activity* in each subband at sub-pixel $\mathbf{n}$:

$$E^{(n,m)}(\mathbf{n}) = \frac{\left|D_1^{(n,m)}(\mathbf{n})D_2^{(n,m)}(\mathbf{n})\right|}{P^{(n,m)}} \quad \text{(D.16)}$$

The effective f_{max} for this approximation of $SD^{(n,m)}$ is 0.5. Hence this method assumes that the motion is less than half the separation between level m sub-pixels in each direction. That is, an m-level algorithm can estimate motion up to 0.5×2^m pixels in each direction (consequence of using the staircase approximation). This range can be extended and details of this can be found in (Magarey 1997). In (Magarey & Kingsbury 1995), Magarey and Kingsbury report that in practise this measurement is more like ≤ 0.35 due to the deterioration in the staircase approximation as $\mathbf{f}$ increases. Consequently, the m_{max} chosen must be large enough to allow this measurement range to encompass the largest expected displacement.

The minimum of the quadratic surface $SD^{(n,m)}$ as approximated in equation D.15 is specified by the *equiphase equation*

$$\phi^{(n,m)}(\mathbf{n},\mathbf{f}) = \phi_2^{(n,m)}(\mathbf{n}) \quad \text{(D.17)}$$

which needs to be solved for $\mathbf{f}$ over $n = 1,\ldots,6$. From the planar model of interpolated phase (equations D.9 and D.10):

$$2^m\left(\mathbf{\Omega}^{(n,m)}\right)^\top \mathbf{f} = \theta^{(n,m)}(\mathbf{n}) \quad \text{(D.18)}$$

$$\text{where } \theta^{(n,m)}(\mathbf{n}) = \angle\left[\frac{D^{(n,m)_2}}{D_1^{(n,m)}}\right] \quad \text{(D.19)}$$

The minimum of $SD^{(n,m)}$ therefore lies along a line perpendicular to the centre frequency vector $\mathbf{\Omega}^{(n,m)}$ of the associated wavelet filter. The minimum line is therefore parallel to the preferred direction of the filter (see Figure D.1).

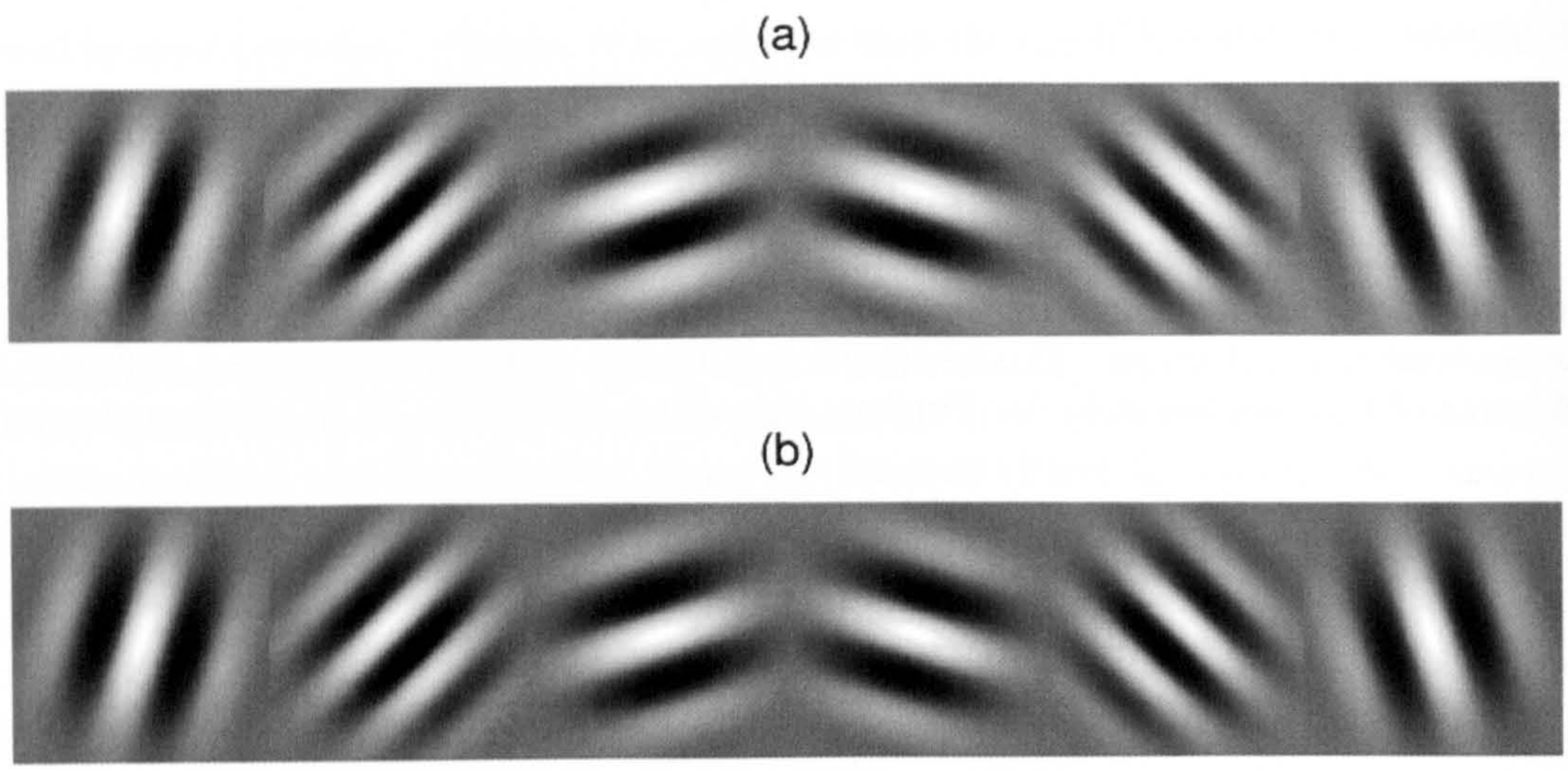


Figure D.1: Greyscale (grey=0) plot of the impulse responses of the 6 wavelet filters $\psi^{(n,4)}$ at level 4 of the 2D CDWT based on the pair of Gabor filters in equations B.1 and B.2($a_0 = 0.39, a_1 = 0.39j, \omega_0 = \pi/6, \omega_1 = 5\pi/6, \sigma_0 = \sigma_1 = 1.27$). (a) Real part. (b) Imaginary part. From left to right, the order of orientations is $n = 2, 3, 1, 4, 6, 5$.

The quadratic surface $SD^{(n,m)}$ can be characterised around its minimum line by using the approximation $\cos x \approx 1 - \frac{x^2}{2}$ and the planar phase model (equation D.9):

$$SD^{(n,m)}(\mathbf{f}) \approx \left(\left| D_1^{(n,m)}(\mathbf{n}) \right| - \left| D_2^{(n,m)}(\mathbf{n}) \right| \right)^2 + \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| \left(2^m (\mathbf{\Omega}^{(n,m)})^\top \mathbf{f} - \theta^{(n,m)}(\mathbf{n}) \right)^2 \quad \text{(D.20)}$$

If the minimum lines of the six $SD^{(n,m)}$ surfaces do not lie too far apart, then the sum $SD^{(m)}$ can also be approximated as a quadratic surface.

$$SD^{(m)}(\mathbf{n}, \mathbf{f}) \approx A f_1^2 + B F_2^2 + C f_1 f_2 + D f_1 + E f_2 + G \quad \text{(D.21)}$$

Closed form expressions for the coefficients $\{A, B, C, D, E, F, G\}$ in terms of coefficients $D_1^{(n,m)}(\mathbf{n})$ and $D_2^{(n,m)}(\mathbf{n})$ and the centre frequencies $\mathbf{\Omega}^{(n,m)}$ may be obtained from equations

D.20 and D.4:

$$A = \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| \left(\Omega_1^{(n,m)} \right)^2 \tag{D.22}$$

$$B = \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| \left(\Omega_2^{(n,m)} \right)^2 \tag{D.23}$$

$$C = \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| 2\Omega_1^{(n,m)} \Omega_2^{(n,m)} \tag{D.24}$$

$$D = \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| (-2)\Omega_1^{(n,m)} \theta^{(n,m)}(\mathbf{n}) \tag{D.25}$$

$$E = \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| (-2)\Omega_2^{(n,m)} \theta^{(n,m)}(\mathbf{n}) \tag{D.26}$$

$$G = \sum_{n=1}^{6} \left(\left| D_1^{(n,m)}(\mathbf{n}) \right| - \left| D_2^{(n,m)}(\mathbf{m}) \right| \right)^2 \tag{D.27}$$

$$+ \sum_{n=1}^{6} \left| D_1^{(n,m)}(\mathbf{n}) D_2^{(n,m)}(\mathbf{n}) \right| \left(\theta^{(n,m)}(\mathbf{n}) \right)^2$$

where

$$2^m \boldsymbol{\Omega}^{(n,m)} = \left(\Omega_1^{(n)}, \Omega_2^{(n)} \right)^\top \tag{D.28}$$

By completing the square, $SD^{(m)}$ can be transformed from the representation in D.21 to

$$SD^{(m)}(\mathbf{n}, \mathbf{f}) \approx \alpha(f_1 - f_{10})^2 + \beta(f_2 - f_{20})^2 + \gamma(f_1 - f_{10})(f_2 - f_{20}) + \delta \tag{D.29}$$

where $\mathbf{f}_0 = [f_{10} f_{20}]^\top$ are the co-ordinates of the surface minimum. Also, it can be seen that

$$\mathbf{f}_0 = \frac{\left[2BD - CE, 2AE - CD \right]^\top}{C^2 - 4AB} \tag{D.30}$$

The curvature parameters $\alpha, \beta, \gamma = A, B, C$ together define the curvature matrix of the surface at its minimum point such that

$$K = \begin{bmatrix} 2\alpha & \gamma \\ \gamma & 2\beta \end{bmatrix} \tag{D.31}$$

δ in equation D.29 represents the surface minimum value and $\delta = G - Af_{10}^2 - Bf_{20}^2 - Cf_{10}f_{20}$. The value of δ is indicative of the closeness of the match between $D_2^{(n,m)}(\mathbf{n})$ and $D_1^{(n,m)}(\mathbf{n}+\mathbf{f}_0)$ over $n = 1, \ldots, 6$.

Hence the location $\mathbf{f}_0(\mathbf{n})$ of the minimum, scaled by 2^m, is the coarse motion estimate for the $2^m \times 2^m$ block of pixels in the reference frame centred on $2^m\mathbf{n}$. It is composed of information from each subband, weighted by the energy in that subband.

D.5 Hierarchical Estimation

D.5.1 SSD Parameter Field Interpolation

The coarse level estimates are used as initial guesses and these are progressively refined by including the finer scale information. At the coarsest level m_{max}, a field of SD parameters, in the form of six real-valued parameter matrices (either $\{A, B, C, D, E, G\}$ or $\{\mathrm{f}_0, \alpha, \beta, \gamma, \delta\}$) is computed. The dimensions of the level m parameter matrices are doubled (interpolated and scaled as $\mathbf{f} \longmapsto 2\mathbf{f}$ to give a one-to-one correspondence since at level $m-1$ there are four times as many sub-pixels).

The interpolated field of level m surfaces is denoted by $SD'^{(m)}(\mathbf{n}, \mathbf{f})$, with parameters $\{A', B', C', D', E', G'\}$ or $\{\mathrm{f}'_0, \alpha', \beta', \gamma', \delta'\}$.

Two interpolation schemes are investigated by Magarey in (Magarey 1997): staircase interpolation and bilinear interpolation. The bilinear interpolation kernel is found to give better results (Magarey 1997). The columns and the rows are interpolated separately, first by upsampling and then by filtering with the bilinear kernel [1 3 3 1]/4 (Castellano 1999).

The set $\{A, B, C, D, E, G\}$ is interpolated, and then equation D.30 is used to find f'_0. The motion field is effectively weighted by curvature (*curvature weighted*) and the parameters increase in rough proportion to the activity at that sub-pixel. As a result, estimates from high activity regions tend to propagate into regions of low activity in curvature weighted interpolation.

Since the separation between the adjacent pixels is halved at the next finer level, the variables are scaled:

$$\alpha', A' \quad \mapsto \quad \alpha'/4, A'/4 \tag{D.32}$$
$$\beta', B' \quad \mapsto \quad \beta'/4, B'/4 \tag{D.33}$$
$$\gamma', C' \quad \mapsto \quad \gamma'/4, C'/4 \tag{D.34}$$
$$D', E' \quad \mapsto \quad D'/2, E'/2 \tag{D.35}$$
$$\mathrm{f}'_0 \quad \mapsto \quad 2\mathrm{f}'_0 \tag{D.36}$$

D.5.2 Cumulative SD Surfaces

Once the disparity field $SD'^{(m)}$ is obtained, the disparity field at the next finer level $m-1$ is estimated using the procedure outlined in Section D.2. The resulting parameters $SD^{(m-1)}$ are added to the estimates $SD'^{(m)}$ from the previous level to give the *cumulative squared difference*, CSD, at level m. This process of interpolation, computing the new disparity field estimates and adding the two is repeated iteratively until the desired level m_{min} is reached, i.e.

$$CSD^{(m)}(\mathbf{n}, \mathbf{f}) = \begin{cases} CSD'^{(m+1)}(\mathbf{n}, \mathbf{f}) + SD^{(m)}(\mathbf{n}, \mathbf{f}) & m_{min} \leq m < m_{max} \\ SD^{(m)}(\mathbf{n}, \mathbf{f}) & m = m_{max} \end{cases} \tag{D.37}$$

Thus, information from all resolutions and all subbands is utilised to form one single quadratic surface, which gives the final motion vector. In areas and directions where there is limited fine detail, such as along edges, the curvature parameters will indicate a shallow surface (low confidence) in that direction. Such a surface will contribute very little when combined with a coarser level surface, which is steep in that direction. Such a strategy in known as a *refining strategy*.

D.5.3 Coarse-to-fine Strategies

The refining strategy uses the coarser level estimates to warp the CDWT coefficients of the reference image at the next finer level using equations D.5 to D.9 to produce new coefficients $D^{(m,n)}(\mathbf{n}+\mathbf{f}_0'), n = 1,\ldots,6$. The $SD^{(m)}(\mathbf{n},\mathbf{f}-\mathbf{f}_0)$ is formed by using the interpolated (windowed-sinc kernel) coefficients instead of the integer-indexed ones. This is then used to find the parameters of valid $SD^{(m)}(\mathbf{n},\mathbf{f})$ and thereby translate the origin back to $(0,0)$. This valid $SD^{(m)}$ is added to the scaled $CSD'^{(m+1)}$ from the coarse level to form $CSD^{(m)}$ using equation D.37 and the parameters $\{\mathbf{f},\alpha,\beta,\ldots\}$ are formulated. This approach has the advantage of detecting uniform motion of large scale features with excellent accuracy. However, motion of the smaller scale features that is independent of the motion of the larger scale features is not detectable.

D.5.4 Confidence Measure

Confidence measures partition the motion estimates into reliable and unreliable estimates, based on some threshold. The unreliable estimates can be eliminated so that subsequent estimates are not corrupted. This process is repeated for each of the estimates at each level of decomposition.

The confidence measure adopted by Magarey et al. is based on the residual of the weighted least-squares solution, represented by $\mathbf{f}_0$. It is computed using:

$$C^{(m)}(\mathbf{n}) = 1 - \frac{\delta - \Delta^{(m)}(\mathbf{n})}{E^{(m)}(\mathbf{n})} \tag{D.38}$$

$$\text{where } E^{(m)}(\mathbf{n}) = \sum_{n=1}^{6} E^{(n,m)}(\mathbf{n}) \tag{D.39}$$

$$\text{and } \Delta^{(m)}(\mathbf{n}) = \sum_{n=1}^{6} \frac{\left(\left|D_1^{(n,m)}(\mathbf{n})\right| - \left|D_2^{(n,m)}(\mathbf{n})\right)^2}{P^{(n,m)}} \tag{D.40}$$

The $\Delta^{(m)}(\mathbf{n})$ is the *discrepancy*. $C^{(m)}$ is equal to the normalised cross-correlation of $D_1^{(n,m)}(\mathbf{n}+\mathbf{f})$ with $D_2^{(n,m)}(\mathbf{n})$ (Magarey & Kingsbury 1998*b*). δ is the surface minimum, and $E^{(m)}(\mathbf{n})$ and $P^{(n,m)}$ can be computed from equations D.16 and D.2 respectively. Because $\delta \geq \Delta$, then $C \leq 1$. All values of C that lie below a certain threshold value are nullified and their effects are not observed in subsequent computations. Magarey found that a value of 0.95 for the threshold gives the best balance between preserving sufficient field density and increasing accuracy (Magarey & Kingsbury 1998*b*).

D.5.5 Curvature Correction

Recall that flat surfaces are indicative of low confidence in the motion estimate, while steep surfaces indicate higher confidence. In addition, the coarser levels estimates tend to be more aperture-free than the finer level estimates due to their larger region of support (Castellano 1999). The aperture problem is addressed by propagating the coarser levels estimates to the aperture affected finer levels so that the parallel component of motion may be determined.

A curvature correction measure is introduced to improve the accuracy of these estimates. The *rotation invariance* property of the filter pair $\{h_0, h_1\}$ enables this curvature correction. This in turn stems from the fact that the filters h_0 and h_1 were designed to produce ellipses whose (finite) eccentricity is nearly independent of edge orientation.

The curvature correction process "corrects" the curvatures of the strongly aperture-affected surfaces by assigning them very large eccentricity while leaving surfaces at all other subpixels relatively unchanged. This involves subtracting from *all* $SD^{(m)}$ surfaces (prior to combination with previous level surfaces $CSD'^{(m+1)}$) a circular bowl-shaped surface with the same minimum location $\mathbf{f}_0$ as $SD^{(m)}$. This is done so that the minimum location and the height of $SD^{(m)}$ are not affected.

$$SD^{(m)}_{corr}(\mathbf{n}, \mathbf{f}) = SD^{(m)}(\mathbf{n}, \mathbf{f}) - \rho(f_1 - f_{10}(\mathbf{n}))^2 - \rho(f_2 - f_{20}(\mathbf{n}))^2 \tag{D.41}$$

The curvature of this bowl is calculated such that surfaces corresponding to the aperture-affected estimates become nearly flat, thus reducing their contribution to the cumulative squared difference, while surfaces corresponding to aperture-free estimates are left nearly unchanged.

The radius ρ is chosen so that all surfaces with eccentricity $e > e_t$, for some threshold e_t, have very large eccentricity after correction:

$$\rho = \min\left\{\left(\frac{\alpha+\beta}{e_t^2+1}\right), 0.98\lambda_2\right\} \tag{D.42}$$

where λ_2 is the smaller eigenvalue of curvature matrix K (equation D.31).

D.5.6 Disparity Field Regularisation

Anandan's regularisation procedure, proposed in (Anandan 1989), is used to smooth the disparity field at each level of decomposition. This approach preserves the coarse-to-fine framework of the matching algorithm. Furthermore, sub-pixel accuracy achieved by this algorithm can be retained since the approach does not require that the disparity at each pixel refers to the precise location of another pixel (Dick 1997).

Recall that the SSD is defined as an elliptical surface centred on its minimum value. Two *principal axes* are defined for this surface: $\mathbf{e}_{max}$ in the direction of maximum surface curvature (steepest sides of the bowl) and $\mathbf{e}_{min}$ in the direction of the minimum surface curvature. These are analogous to the major and minor axes of a 2D ellipse. Associated with these axes are two curvatures C_{max} and C_{min} in the directions of $\mathbf{e}_{max}$ and $\mathbf{e}_{min}$ respectively. C_{max} and C_{min} are the greater and the lesser eigenvalues respectively of the curvature matrix K (equation D.31) and $\mathbf{e}_{max}$ and $\mathbf{e}_{min}$ are the eigenvectors corresponding to these eigenvalues.

In (Anandan 1989) these values are used to define maximum and minimum confidence measures c_{max} and c_{min}.

$$c_{max} = \frac{C_{max}}{k_1 + k_2 SD^{(m)}(\mathbf{n}, \mathbf{f}_0) + k_3 C_{max}} \tag{D.43}$$

$$c_{min} = \frac{C_{min}}{k_1 + k_2 SD^{m}(\mathbf{n}, \mathbf{f}_0) + k_3 C_{min}} \tag{D.44}$$

k_1, k_2 and k_3 are parameters that influence the behaviour of c_{max} and c_{min}. k_1 is an overall scaling factor and prevents the values of c_{max} and c_{min} from becoming too large

if $SD^{(m)}(\mathbf{n}, \mathbf{f}_0)$ is close to 0. k_2 controls the relative influence of the curvatures and the similarity distance, and k_3 is used to restrict the range of possible confidence values to the interval $(0, 1/k_3)$. Values of c_{max} and c_{min} determine the amount of smoothing to apply in each direction. Directions with high confidence values for a match require less smoothing.

Given the field $\{\mathbf{f}_0\}$ of feature space minimum location offsets, the regularisation procedure obtains an optimal compromise between feature similarity and disparity field smoothness by finding a field $\{\mathbf{u}\}$ such that the function

$$E(\{\mathbf{u}\}) = E_{sm} + \lambda E_{ap}(\{\mathbf{u}\})$$

is minimised (Magarey & Dick 1998). The smoothness term $E_{sm}(\{\mathbf{u}\})$ is some measure of the difference between $\{\mathbf{u}\}$ and a uniform disparity field and defined in terms of the gradients of the components (u_x, u_y) of $\mathbf{u}$ (Magarey & Dick 1998). The global "approximation error" energy $E_{ap}(\{\mathbf{u}\})$ is the normalised measure of difference between $SD^{(m)}(\mathbf{n}+\mathbf{u})$ and $SD^{(m)}(\mathbf{n}+\mathbf{f}_0)$, the feature space minimum. The scaling factor λ controls the influence of each difference on the resultant disparity field.

An exact global minimisation of $E(\{\mathbf{u}\})$ is computationally expensive to compute. Anandan (Anandan 1989) shows that an approximate solution to the set $\{\mathbf{u}\}$ that minimises $E(\mathbf{u})$ may be found using only the local information surrounding each subpixel, i.e. using disparity vectors which satisfy the equation:

$$(\mathbf{u}-\overline{\mathbf{u}}) + \lambda c_{max}(\mathbf{u}\cdot\mathbf{e}_{max} - \mathbf{f}_0\cdot\mathbf{e}_{max})\mathbf{e}_{max} + \lambda c_{min}(\mathbf{u}\cdot\mathbf{e}_{min} - \mathbf{f}_0\cdot\mathbf{e}_{min})\mathbf{e}_{min} = 0 \quad \text{(D.45)}$$

$\overline{\mathbf{u}}$ is the four-neighbour average of $\mathbf{u}$. $(\mathbf{u}-\overline{\mathbf{u}})$ is a measure of local variation in $\{\mathbf{u}\}$, and hence represents $E_{sm}(\{\mathbf{u}\})$ (Dick 1997).

Equation D.45 can be solved for each $\mathbf{u}$ using Gauss-Seidel iteration (Anandan 1989). This involves repeatedly applying the update equation

$$\mathbf{u}^{n+1} = \overline{\mathbf{u}}^n + \frac{\lambda c_{max}}{\lambda c_{max}+1}\Big((\mathbf{f}_0-\overline{\mathbf{u}}^n)\cdot\mathbf{e}_{max}\Big)\mathbf{e}_{max} + \frac{\lambda c_{min}}{\lambda c_{min}+1}\Big((\mathbf{f}_0-\overline{\mathbf{u}}^n)\cdot\mathbf{e}_{min}\Big)\mathbf{e}_{min} \quad \text{(D.46)}$$

until a set $\{\mathbf{u}\}$ is found whose elements $\mathbf{u}$ have all converged to within some neighbourhood of a solution to equation D.45 (Dick 1997). In this work, λ is set to 0.75 with 20 Gauss-Seidel iterations.

The presence of the matching surface curvatures in the regularisation means that only limited smoothing is applied in regions and directions in which the confidence is high. The effect of smoothing is increased in regions and directions of lower confidence. This results in a disparity field which preserves sharp features and smoothes out featureless regions which may be riddled with noise.

The accuracy of the disparity field is further increased by thresholding the disparity field $\{\mathbf{f}_0\}$ to remove low confidence measures prior to smoothing. A threshold on the product of the confidence measures c_{max} and c_{min} is effective in removing the matches between the featureless points in the background, while still retaining matches on the faces.

APPENDIX E

Johnson's 3D Object Recognition Using Spin-Images

The spin-image representation comprises of a set of descriptive images associated with the oriented points (defined using the 3D position of the mesh vertex and its surface normal) on the surface of an object (Ruiz-Correa et al. 2001). They derive their name from the image generation process, which can be visualised as a sheet spinning about the normal of the various points on the polygonal mesh representing the surfaces. Surfaces are matched by comparing spin-images from points on one surface to spin-images from points on another surface. A correlation coefficient is used to compare the points. When two spin-images are highly correlated, a point correspondence is established. Point correspondences are then grouped and outliers are eliminated using geometric consistency. Groups of geometrically consistent correspondences are used to calculate a rigid transformation that aligns the two surfaces. Finally surface matches are verified using a modified closest iterative point algorithm. The following sections detailing the algorithm have been taken almost entirely from Johnson's original thesis (Johnson 1997).

E.1 Assumptions

Surface Representation

It is assumed that surfaces are represented using polygonal meshes. This allows the user to choose the level of detail that the surface is represented at. In addition, if the meshes can be generated directly from sensed 3D data then fitting or approximation is not required, and the introduction of unnecessary errors can be avoided.

Surface Normals

It is also assumed that the surface normals are oriented outside the object surface. If this is not the case then the following heuristic can be used to ensure all the vertex normals are oriented to the outside: first, a vertex is chosen and the orientation of its normal is

propagated outwards to the normals of its adjacent vertices. This process is repeated until all the normals are consistently oriented to the inside or the outside of the object. Next, the orientation of all the normals is determined by calculating the scalar products of the surface normal at each vertex and the vector from the centroid of the object to the vertex. If the majority of the scalar products are positive, the normals have been oriented outside. Otherwise, the normals have been oriented to the inside and they need to be inverted.

Mesh Resolution

In (Johnson 1997), Johnson discuses the concept of *mesh resolution* at length, as it is closely related to the level of detail contained in the mesh. He defines mesh resolution as the median of all edge lengths in a mesh. This relies on the assumption that all the edges in the mesh are of a similar length. When this is the case, the spacing between the vertices will be approximately uniform, and the defined metric is meaningful. Note that according to this definition, mesh resolution is inversely proportional to the number of vertices in the mesh, i.e. the resolution increases when there are fewer vertices in the mesh (edges are longer in a coarser mesh) and it decreases as the number of vertices increases. It can be seen that the metric is poorly defined if there is a great deal of variation in the edge lengths. Hence, if the mesh has an uneven distribution of vertices, then it has to be recreated with uniform spacing. Johnson's mesh simplification algorithm is presented in Section E.2. Details of Johnson's experiments, results and an analysis of how this algorithm compares with other mesh simplification algorithms from computer graphics (e.g. (Guéziec 1995, Heckbert & Garland 1997, Hoppe 1996, Schroeder et al. 1992, Turk 1992)) can be found in Johnson's PhD thesis (Johnson 1997).

Uniform sampling of surfaces is a requirement for the spin-images of two corresponding points on different instances (surface mesh representations) of the same object to be similar. This is a much weaker and a more preferable constraint than requiring the positions of points to be the same for the two instances. If the surfaces are uniformly sampled then on average, each corresponding bin of the spin-images will have the same number of points projected into it, making the spin-images similar, *even though the co-ordinates and surface normals of the vertices are not exactly the same.* If the sampling is not uniform, then the mesh resampling algorithm (section E.2) is used to ensure uniform sampling. If the sampling is uniform, then it is not necessary for the resolution of the meshes being matched to be equal.

Noise

Noise is introduced to the process of spin-image generation in two forms: noise in the 3D vertex position and noise in the surface normal. Since the 3D position of the vertex is bilinearly interpolated during the accumulation process, the effect of this type of noise on spin-image appearance is minimised, as long as its magnitude is less than the distance between mesh vertices. The surface normal noise however, can alter the spin-images significantly. Its effect on the spin-image appearance increases as α and β increase across the spin-image. As a result, small errors in the surface normal manifest as large changes in spin-image appearance. This can lead to incorrect correspondence between oriented points and spin-map co-ordinates. However, these matches are eliminated during the surface matching process using geometric consistency.

Scale-Invariance

Spin-images are invariant to rigid body transformations, but they are not scale-invariant. So, two surfaces of the same shape but different sizes will generate different sets of spin-images. However, the spin-map is a function of Euclidean distances, so the spin-images from scaled versions of the same shape will be the same up to the scale factor between the surfaces. Therefore, in theory, multiple models do not have to be stored to represent surfaces at different scales; only the original model and the necessary scale factors need to be stored. This has not yet been implemented in practise and the claim remains to be investigated.

E.2 Mesh simplification algorithm

Mesh simplification process is essential in order to normalise the edge lengths so that they are roughly the same, or equivalently, so that the vertices are uniformly distributed. The mesh resolution also determines the level of detail in the mesh, and this can also be controlled using this algorithm. A brief outline of the algorithm is given here. A detailed analysis of the algorithm along with the results of its application to various object meshes is presented in Johnson's PhD thesis (Johnson 1997).

First, a priority queue (a dynamically ordered queue) is formed of all the edges in the mesh to be simplified. The position of an edge in the priority queue is given by

$$\text{Edge Length Weight, } W \times \text{Accumulated Shape Change Measure, } C;$$

Edges with small product ($W \times C$) will be towards the top of the queue. The edge length weight, W is generated from a Gaussian of the edge length:

$$W = e^{\left(-\frac{l-L_0}{L_D^2}\right)}$$

where l is the length of the edge, L_0 is the desired mesh resolution and L_D is the acceptable deviation in length from the desired resolution for edges in the the normalised mesh. A small weight is assigned to edges that are much shorter or longer than the desired resolution.

The shape change measure, D, of an edge is defined as the distance between the current mesh and the mesh that results from the simplification. This places a bound on the maximum change in the shape of the mesh during the normalisation process. Since edge operations affect only a local neighbourhood of the edge, the distance between meshes can be measured by comparing only the local mesh neighbourhoods before and after application of the operation. The mesh shape is conveyed by both the vertices and the faces. Therefore, an accurate measure of distance between meshes must take into account the distance between the mesh faces as well. The asymmetric distance between meshes $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined as the maximum Euclidean distance between a vertex v_i of $\mathcal{M}_1$ and its associated closest point, v_{closest} on the *face* f_j of $\mathcal{M}_2$ that is closest to v_i.

$$d(\mathcal{M}_1, \mathcal{M}_2) = \max_{v_i \in \mathcal{M}_1} \left(\min_{f_j \in \mathcal{M}_2} \|v_i - v_{\text{closest}}(v_i, f_j)\| \right)$$

Since this distance metric is not symmetric, it is re-defined as the maximum of $d(\mathcal{M}_1, \mathcal{M}_2)$ and $d(\mathcal{M}_2, \mathcal{M}_1)$.

$$D(\mathcal{M}_1, \mathcal{M}_2) = \max(d(\mathcal{M}_1, \mathcal{M}_2), d(\mathcal{M}_2, \mathcal{M}_1))$$

This is a useful metric in that it allows the simplification algorithm to operate on edges along surface shape discontinuities (e.g. ridges and corners), provided that the distance between the meshes remains small during the operation. This facility is not available in most other simplification algorithms (Johnson 1997). In the simplification process, the *accumulated* shape change (shape change accrued so far) is used instead of the raw shape change. This ensures that the amount of shape change during the normalisation is limited to a certain user-defined maximum allowable change in shape, $C_{\max}$.

Also defined are the upper and the lower bounds on edge lengths in the simplified mesh:

$$L_{\min} = L_0 - \frac{L_D}{2} \qquad L_{\max} = L_0 + \frac{L_D}{2}$$

Note that edges that fall within the desired edge length bounds are not added to the dynamic priority queue. Neither are the edges whose accumulated shape change measure C exceeds the maximum allowable change in shape. This means that if a particular edge has changed significantly from its original length and has reached $C_{\max}$, then it cannot be operated on, even though it may not fall within the bounds imposed by $L_{\min}$ and $L_{\max}$. Hence, not all edges in the final mesh will have lengths inside of the bounds.

Each edge in the queue is operated on as follows: If the length of the edge is greater than $L_{\max}$, the edge is split at its midpoint. This split changes the neighbourhood of the edge by adding an edge, a vertex and two new faces. This operation does not change the shape of the mesh, and it does not alter the accumulated shape change of the edges in the neighbourhood of the edge. If the edge length is less than $L_{\min}$, the edge is collapsed into a point. This alters the neighbourhood by eliminating an edge, a vertex and two faces. This causes the mesh to either shrink or expand. The shape change measure is added to the accumulated shape change of the edges in the new neighbourhood of the edge. Details of how the exact position of the new vertex is determined are given in (Johnson 1997). After an edge split or an edge collapse, the edges in the old neighbourhood of the edge are removed from the priority queue. Edges in the new neighbourhood of the mesh are added to the queue if they meet the following criteria:

- The edge lengths fall outside the bounds for the minimum and maximum edge lengths, $L_{\min}$ and $L_{\max}$.
- Their accumulated shape change is not greater than $C_{\max}$.
- They meet additional checks that prevent changes in topology and shrinkage of the mesh boundary. Details of these checks can be found in (Johnson 1997).

This process is repeated iteratively till there are no more edges in the priority queue.

In (Johnson 1997), Johnson compares this algorithm with other competing mesh simplification algorithms in computer graphics. This technique is reported to be superior of those investigated.

E.3 Spin-Image Parameters in Detail

Three parameters control the spin-image generation.

1. **Bin Size:** This is the geometric size (storage size) of the bins in the spin-images generated. It determines the averaging (from bilinear interpolation) in the spin-images. Averaging reduces the effect of individual point positions (large bin size implies more averaging) and affects the descriptiveness of the spin-images. It is set as a multiple of the surface mesh resolution (defined in section E.1) to avoid dependence on object scale and resolution. This is intuitive as the mesh resolution is closely related to the density of points in the surface points. In (Johnson 1997), Johnson finds that setting the bin size 1-2 times the mesh resolution sufficiently blurs the position of individual points in the spin-images, while still adequately describing global shape.

2. **Image Width:** For simplicity and without loss of any generality, the number of rows in a spin-image is set equal to the number of columns. They can however, be set to equal any arbitrary number. Setting the rows and the columns to equal one another results in square spin-images whose size can be defined by one parameter. Image width controls the global information content of the spin-image. For a fixed bin-size, decreasing image width decreases the descriptiveness of a spin-image because the amount of global shape included in the image will be reduced. However, decreasing the image width also limits the corruption in the spin-images caused by clutter.

3. **Support Angle:** This is the maximum angle between the direction of the oriented point basis of a spin-image and the surface normal of points that are allowed to contribute to the spin-image. Take two oriented points $\mathcal{A}$ and $\mathcal{B}$ and their respective 3D positions and normals $(\mathbf{p}_\mathcal{A}, \mathbf{n}_\mathcal{A})$ and $(\mathbf{p}_\mathcal{B}, \mathbf{n}_\mathcal{B})$. Then, the support angle constraint can be stated as: *$\mathcal{B}$ will be accumulated in the spin-image of $\mathcal{A}$ if*

$$\arccos(\mathbf{n}_\mathcal{A} \cdot \mathbf{n}_\mathcal{B}) < A_s. \tag{E.1}$$

Support angle is used to limit the effect of self-occlusion and clutter during spin-image matching. The size of the support angle is directly proportional to the descriptiveness of the spin-image. However, if the support angle is too large, many scene points that do not belong to the model are spin-mapped into the scene spin-image.

Given the bin size b and the image width W, the spin-image bin associated with particular spin-map co-ordinates can be computed using

$$i = \left\lfloor \frac{\frac{W}{2} - \beta}{b} \right\rfloor \qquad j = \left\lfloor \frac{\alpha}{b} \right\rfloor \tag{E.2}$$

where $\lfloor f \rfloor$ is the floor operator which rounds f to the nearest integer, (α, β) are the spin-map co-ordinates and (i, j) refer to the spin-image bin. Because the distance to the tangent plane of an oriented point can be both positive and negative, the spin-image has $\frac{W}{2}$ rows above $\beta = 0$ and $\frac{W}{2}$ rows below $\beta = 0$. Given the bin size b, the bilinear weights used to increment the bins in the spin-image are calculated using

$$m = \alpha - ib \qquad n = \beta - jb \tag{E.3}$$

where (i, j) refers to the spin-image bin. The *support distance* D_s is defined as

$$D_s = W \times b \tag{E.4}$$

and it determines the amount of space swept out by a spin-image. The support distance is also directly proportional to how localised the images are or equivalently, how descriptive they are.

The choice of the spin-image generation parameters dictates how localised or "globalised" the spin-images are. When the spin-images are not localised enough, test model and the training model spin-images appear very different.

Detailed experiments and analysis of the spin-image generation parameters is given in (Johnson 1997). Johnson describes his clutter model and investigates the optimum values for the parameters in the presence of clutter and occlusion. However, since the models used in this project contain no clutter and only self-occlusion, these details are not presented here.

E.4 Comparing Spin-Images

Spin-images from two instances of the same object, generated with the same spin-image generation parameters will have similar spin-images since the spin-map co-ordinates (α, β), with respect to a particular oriented point basis, are independent of rigid body transformations.

Two spin-images from proximal points on the surface of two different instances of an object are linearly related because the number of points that fall in corresponding bins will be similar, given that the distribution of points over the surface of the objects is the same. The bin values are directly related to the number of points falling into the bins. Hence, the bin values will be similar. Then, the linearly related images can be compared using the normalised linear correlation coefficient. Given two spin-images P and Q with N bins each, the *linear correlation coefficient* $R(P,Q)$ is

$$R(P,Q) = \frac{N\sum p_i q_i - \sum p_i \sum q_i}{\sqrt{\left(N\sum p_i^2 - (\sum p_i)^2\right)\left(N\sum q_i^2 - (\sum q_i)^2\right)}}, \tag{E.5}$$

where p_i and q_i refer to the individual points in the image. R lies between -1 for completely uncorrelated images and 1 for completely correlated images. It measures normalised error using the distance between the data and the best least squares fit line to the data.

The linear correlation coefficient can be expected to be similar across the entire image if the spin-images are generated from complete objects. However, data obtained from most real scenes fail to meet this criteria. Real test images may be corrupted by both clutter (extra data) and occlusion (missing data). Spin-images from different incomplete data sets will not be the same everywhere even though they correspond to the same object.

This is resolved by removing portions of the images from the image comparison process. If a bin in either of the images does not have a value (i.e., no vertex was spin-mapped into it) then that bin is not considered in the calculation of the linear correlation coefficient. Hence, the data used to compute the correlation coefficient is taken only from the region of overlap between two spin-images.

The linear correlation coefficient is a function of the number of pixels used to compute it. Therefore, the amount of overlap between the images affects the value of the coefficient. The number of pixels used to compute the correlation coefficient indicates the amount of confidence in its value. Two images A and B with a small overlap can have a higher correlation coefficient than images A and C with a large overlap, if (A,B) are very similar in the area of overlap and (A,C) are slightly less similar in the area of overlap. In this case, A would

be incorrectly matched with B. The match (A, C) should be ranked higher as there is more confidence in the value of the correlation coefficient.

Johnson (Johnson 1997) finds that matching spin-images based on the magnitude of correlation coefficient as well as the confidence in it results in more accurate oriented point correspondences. He derives a *similarity measure* for spin-image matching that incorporates both, the linear correlation coefficient and a measure of confidence in it (measured by its variance). The similarity measure $C(P, Q)$ (not derived here) is stated as

$$C(P,Q) = (\text{arctanh}(R(P,Q)))^2 - \lambda\left(\frac{1}{n-3}\right). \tag{E.6}$$

$R(P, Q)$ is the linear correlation coefficient (equation E.5), n is the number of overlapping pixels used to compute $R(P, Q)$ and λ weights the variance against the expected value of the correlation coefficient. The hyperbolic arctangent function is a standard statistical technique for change of variables and transforms R into a normal distribution with better statistical properties. Specifically, the change of variables leads to the variance of the transformed correlation coefficient to be $1/(n-3)$ - a simple function of the number of pixels used to compute R. A high value of C indicates that the images are more likely to come from corresponding oriented points.

λ limits the matches between spin-images of low overlap. Its value should be set as close to the expected overlap between the spin-images. The value of λ is based on the spin-images of models in the training set. For each model spin-image, the number of bins containing data are counted and put in a list. The value of λ is set to one half of the median value in the list. This takes into account matches with smaller overlap, caused by occlusion.

E.5 Outlier Detection for Similarity Measure Histogram

For uni-modal distributions, a standard statistical way of detecting outliers is to determine the fourth spread of the histogram (Devore 1987).

$$f_s = \text{upper quartile} - \text{lower quartile}$$

$$f_s = \text{median of largest } \frac{N}{2} \text{ measurements} - \text{median of smallest } \frac{N}{2} \text{ measurements}$$

Outliers typically lie between $1.5f_s$ and $3.5f_s$ units above (below) the upper (lower) quartile. Outliers $3.5f_s$ units above the upper quartile, if they exist, are chosen as potential matches.

E.6 Correspondence Filtering

Incorrect correspondences will in general have low similarity measures and be geometrically inconsistent when compared to the rest of the correspondences.

Let P and Q be the test and the training set models respectively. Let $(\mathbf{p}_1, \mathbf{p}_2)$ and $(\mathbf{q}_1, \mathbf{q}_2)$ be the oriented points on the surface of $P$and Q. Then the geometric consistency of two correspondences $\mathbf{C}_1 = [\mathbf{p}_1, \mathbf{q}_1]$ and $\mathbf{C}_2 = [\mathbf{p}_2, \mathbf{q}_2]$ between the test and the training set models is given by

$$d_{gc}(\mathbf{C}_1, \mathbf{C}_2) = \frac{\|\mathcal{S}_{\mathbf{q}_2}(\mathbf{q}_1) - \mathcal{S}_{\mathbf{p}_2}(\mathbf{p}_1)\|}{(\|\mathcal{S}_{\mathbf{q}_2}(\mathbf{q}_1)\| + \|\mathcal{S}_{\mathbf{p}_2}(\mathbf{p}_1)\|)/2}$$

$$D_{gc} = \max(d_{gc}(\mathbf{C}_1, \mathbf{C}_2), d_{gc}(\mathbf{C}_2, \mathbf{C}_1)) \tag{E.7}$$

$\mathcal{S}_{\mathbf{p}_2}(\mathbf{p}_1)$ is the spin-map that projects the 3D point $\mathbf{p}_1$ to the 2D co-ordinate space corresponding to the oriented point $\mathbf{p}_2$. d_{gc} measures the distance between the correspondences in spin-map co-ordinates, normalised by the average of the spin-map co-ordinates. Spin-map co-ordinates (rather than the Euclidean co-ordinates) are used to measure geometric consistency because they are a compact way to measure consistency in position and normals. Distance between the spin-map co-ordinates is normalised by the average spin-map co-ordinates so that the geometric consistency is not biased toward correspondences that are close to each other. Distance measure d_{gc} is not symmetric, therefore, the maximum of the distances is used to define the geometric consistency distance D_{gc}. A small value of D_{gc} indicates that $\mathbf{C}_1$ and $\mathbf{C}_2$ are geometrically consistent, and that the test and training model points in $\mathbf{C}_1$ are the same distance apart and have the same angle between surface normals as the test and training model points in $\mathbf{C}_2$.

E.7 Grouping Point Matches with Geometric Consistency

Correspondences are grouped together so that rigid body transformations that align the test and the training model can be computed. The *grouping criterion* W_{gc}, is the geometric consistency distance (equation E.8) augmented by a weight that promotes grouping of correspondences that are far apart.

$$w_{gc}(\mathbf{C}_1, \mathbf{C}_2) = \frac{d_{gc}(\mathbf{C}_1, \mathbf{C}_2)}{1 - e^{-\left((\|\mathcal{S}_{\mathbf{q}_2}(\mathbf{q}_1)\| + \|\mathcal{S}_{\mathbf{p}_2}(\mathbf{p}_1)\|)/(2\gamma)\right)}}$$

$$W_{gc}(\mathbf{C}_1, \mathbf{C}_2) = \max(w_{gc}(\mathbf{C}_1, \mathbf{C}_2), w_{gc}(\mathbf{C}_1, \mathbf{C}_2)) \tag{E.8}$$

A small value of W_{gc} indicates that two correspondences are geometrically consistent and far apart. Geometric consistency is an important aspect of this metric as geometrically inconsistent matches produce transformations of higher error.

Geometric consistency weight is normalised by γ to make it scale independent. Johnson (Johnson 1997) sets this value to be four times the mesh resolution, as this encourages grouping between correspondences that are at least four times the mesh resolution distance from each other. The grouping criterion between a correspondence $\mathbf{C}$ and a group of correspondences $\{\mathbf{C}_i, \ldots, \mathbf{C}_n\}$ is

$$W_{gc}(\mathbf{C}, \{\mathbf{C}_i, \ldots, \mathbf{C}_n\} = \max_i(W_{gc}(\mathbf{C}, \mathbf{C}_i)) \tag{E.9}$$

The correspondences in L are grouped using the procedure described in Section 7.2.3. A plausible rigid body transformation T from the training model to the test model is then calculated from each group $\{[\mathbf{p}_i, \mathbf{q}_i]\}$ of correspondences by minimising

$$E_T = \sum \|\mathbf{q}_i - T(\mathbf{p}_i)\|^2. \tag{E.10}$$

Note that $\mathbf{p}_i$ and $\mathbf{q}_i$ refer to the 3D position of the oriented points. These are used instead of the 3D position and the surface normal together as they allow a well defined algorithm (Faugeras & Hebert 1986, Horn 1987) for finding the best rigid transformation that aligns two points to be used.

E.8 Verifying Matches

The verification algorithm is a modified version of the Iterative Closest Point Algorithm (ICP) of Besl and McKay (Besl & McKay 1992) and Zhang (Zhang 1994). The ICP algorithm works well for registration of free form curves and surfaces when the transformation between the surfaces is small. Transformation between two surfaces is iteratively determined by assigning correspondences between closest points, calculating transformation, transforming all of points on one of the surfaces based on the transformation and then repeating the process. The algorithm is robust even where there is a large amount of noise as long as the initial transformation is small. Its main drawback is that it converges to a local minimum in the pose space and as a result cannot be used if the surfaces to be registered are arbitrarily displaced from each other. Further, the generic ICP is unable to register data sets when one is not a subset of the other, i.e. when there is only partial overlap between the surfaces, because it tries to establish correspondences between all the points in one set with some of the points in the other.

Johnson (Johnson 1997) modifies the generic ICP by limiting the closest point distance measurement only to those areas in the two sets that overlap. This is accomplished by growing the closest point correspondences from initial correspondences established by the matching process thus far.

Two surfaces match (overlap) when many points on one surface correspond to many points on the other surface. If the surfaces are oriented, then the matched surfaces should also have consistently oriented surface normals, i.e. the matched points should be close to each other in 3D position and the surface normal. To this end, a 6D distance metric combining 3D position and surface normal is devised. Given two oriented points $(\mathbf{p}_1, \mathbf{n}_1)$ and $(\mathbf{p}_2, \mathbf{n}_2)$, their 6D distance is

$$d_6 = \sqrt{\|\mathbf{p}_1 - \mathbf{p}_2\| + \nu\|\mathbf{n}_1 - \mathbf{n}_2\|} \tag{E.11}$$

where ν weighs the surface normal information against position information. Johnson (Johnson 1997) sets the value of ν at two times the mesh resolution, causing the normals of vertices to have more of an effect on the distance metric than the positions of the vertices.

E.9 Variants of the Spin-Images Algorithm

The spin-image representation is highly redundant. Most of this redundancy stems from the fact that spin-images generated from two oriented point bases that are close to each other on the surface will be highly correlated (Johnson 1997). Spin images of symmetrical objects are also highly correlated - two oriented point bases on equal but opposite sides of a plane of symmetry will be the same. This adds to the processing time and the storage requirements.

Johnson investigates varying degrees of compression to eliminate this redundancy and make the recognition process faster and more efficient. In total, there are four versions of the algorithm:

1. **MA1:** Matching with no compression. This is the algorithm described in Chapter 7.

2. **MA2:** Matching with model compression. Principal Component Analysis (PCA) is used to compress the spin-images, resulting in Eigen-spin-images. However, instead of the eigenvalues, a measure based on the correlation coefficient (equation E.5) is used

to determine the number of Eigen-spin-images to use. This is because the assumption that input images are distributed in a hyper-ellipsoid is violated.

3. **MA4:** Matching with library compression. Spin-image libraries can be also be compressed using PCA. However, the aim with library compression is to increase differentiation between the models rather than between the actual points on the model. It is implicitly assumed that spin-images from different models are clustered together in spin-image space. Note that this algorithm does not use model compression.
4. **MA3:** Matching with model and library compression. PCA is used to compress both, the model spin-images and the spin image library.

MA2, MA3 and MA4 are not detailed here. The interested reader is directed to (Johnson 1997). MA3 and MA4 were deemed unsuitable for face recognition since the assumption that spin-images from different models are in separate clusters in the spin-image space would not strictly be true for face spin-images. In Johnson's (Johnson 1997) toy and plumbing libraries, all the models have distinct shapes. Although there is some similarity between the spin-images of the plumbing library, the majority are unique to each object. This would not be the case with face images. Since all faces have a similar shape, it is highly unlikely that the library spin-images will be distributed in separate clusters depending on the model. Using MA2 is a possibility, however, it was felt that the spin image representation and the recognition algorithm with no compression should be investigated first, and if suitable results are achieved, then spin-image compression can be investigated.

Also presented in (Johnson 1997) is a spin-image based surface registration algorithm. This facilitates complete models to be built by combining many partial views. It also acts as a noise-filtering mechanism. The details of this algorithm are not presented here as surface registration is not investigated in this work. It would however, constitute an interesting area of future research. For details of the algorithm, the interested reader is directed to (Johnson 1997).

Bibliography

Achermann, B. & Bunke, H. (1996), Combination of face classifiers for person identification, *in* '13th IEEE International Conference on Pattern Recognition (ICPR '96)', Vienna, pp. 416 – 420.

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

Anandan, P. (1989), 'A computational framework and an algorithm for the measurement of visual motion', *International journal of Computer Vision* **2**, 283–310.

Antonini, M., Barland, M., Mathieu, P. & Daubechies, I. (1992), 'Image coding and wavelet transform', *IEEE Transactions on Image Processing* **1**(2), 205 – 220.

Assfalg, J., Bimbo, A. D. & Pala, P. (2004), Spin images for retrieval of 3d objects by local and global similarity, *in* '17^{th} International Conference on Pattern Recognition (ICPR '04)', Vol. 3, pp. 906–909.

Barlow, H. B., Blakemore, C. & Pettigrew, J. D. (1967), 'The neural mechanisms of binocular depth discrimination', *Journal of Physiology (London)* **193**, 327–342.

Barrett, W. A. (1998), A survey of face recognition algorithms and testing results, *in* 'Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers', Vol. 1, pp. 301–305.

Barron, J. & Eagleson, R. (1997), Computation of time-varying motion and structure parameters from real image sequences, *in* F. Solina, W. G. Kropatcsh, R. Klette & R. Bajcsy, eds, 'Advances in Computer Vision', SpringerWien, NewYork, pp. 181 – 190. SM.

Barron, J. L., Fleet, D. J. & Beauchemin, S. S. (1994), 'Performance of optical flow techniques', *International Journal of Computer Vision* **12**(1), 43 – 77. SM.

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

Besl, P. & McKay, N. (1992), 'A method of registration of 3-d shapes', *IEEE Transactions on Pattern Analysis and Machine Vision* **2**(2), 239–256.

Beumier, C. & Acheroy, M. (1998), Automatic fae authentication from 3d surface, *in* 'Proceedings of British Machine Vision Conference', BMVA, University of Southampton, UK, pp. 449 – 458. 3DFR.

Beumier, C. & Acheroy, M. (2000), 'Automatic 3d face authentication', *Image and Vision Computing* **18**(4), 315–321.

Bichsel, M. (1995), Human face recognition: From views to models - from models to views, *in* 'Proceedings of the International Workshop on Automatic Face and Gesture Recognition, ICAFGR 96', Zurich, Switzerland, pp. 59 – 64.

Bichsel, M. & Pentland, A. (1994), 'Human face recognition and the face image set's topology', *CVGIP: Image Understanding* **59**(2), 254–261.

Blanz, V., Romdhani, S. & Vetter, T. (2002), Face-identification across different poses and illuminations with a 3d morphable model, *in* 'Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition 2002'.

Blanz, V., Schölkopf, B., Bülthoff, H., Burgess, C., Vapnik, V. & Vetter, T. (1996), Comparison of view-based object recognition algorithms using realistic 3d models, *in* C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen & B. Sendhoff, eds, 'Proceedings of the International Conference on Artificial Neural Networks ICANN '96', number 1112 *in* 'Springer Lecture Notes in Computer Science', Berlin, pp. 251 – 256.

Blanz, V. & Vetter, T. (1999), A morphable model for the synthesis of 3d faces, *in* 'SIGGRAPH 99 Conference3 Proceedings', Los Angeles, California, USA, pp. 187–194.

Blanz, V. & Vetter, T. (2003), 'Face recognition based on fitting a 3d morphable model', *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Bouguet, J.-Y. (1998), 3d transformations & camera calibration, 3D Photography: Lecture Notes, California Institute of Technology.

Bouguet, J.-Y. (1999), Visual Methods for 3D modelling, PhD thesis, Electrical Engineering, California Institute of Technology (Caltech), Pasadena, USA.

Bowyer, K., Chang, K. & Flynn, P. (2004), A survey of 3d and multi-modal 3d+2d face recognition, Technical Report TR 2004-22, University of Notre Dame, Dept. of Computer Science and Engineering.

Bradley, A. P. (1997), 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern Recognition* **30**(7), 1145 – 1159.

Bronstein, A., Bronstein, M. & Kimmel, R. (2003), Expression-invariant 3d face recognition, *in* J. Kittler & M. S. Nixon, eds, 'Audio- and Video-based Biometric Person Authentication (AVBPA 2003)', Springer-Verlag, pp. 62–70.

Bronstein, A., Bronstein, M. & Kimmel, R. (2005), 'Three-dimensional face recognition', *International Journal of Computer Vision* **64**(1), 5–30.

Bronstein, A., Bronstein, M., Kimmel, R. & Spira, A. (2003), 3d face recognition without facial surface reconstruction, Technical Report CIS-2003-05, Technion - Israel Institute of Technology, Israel.

Brunelli, R. & Poggio, T. (1992), Hyperbf networks for gender classification, *in* 'DARPA Image Understanding Workshop', pp. 311 – 314.

Brunelli, R. & Poggio, T. (1993), 'Face recognition: Features vs. templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Brusco, N., Giorgi, A., Andreeto, M. & Cortelazzo, G. M. (2005), 3d registration by textured spin images, *in* 'The 5th International Conference on Digital Imaging and Modelling (3DIM05)', Banff, Alberta, Canada.

Burner, A. W. (1995), Zoom lens calibration for wind tunnel measurements, *in* 'Proceeding of SPIE Conference on Videometrics IV', Philadelphia, pp. 19 – 33.

Canny, J. (1986), 'A computational approach to edge detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6), 679 – 698.

Carlsson, S. & Weinshall, D. (1998), 'Dual computation of projective shape and camera positions from multiple images', *International Journal for Computer Vision* **27**(3), 227–241.

Carmichael, O., Huber, D. & Hebert, M. (1999), Large data sets and confusing scenes in 3-d surface matching and recognition, *in* 'Proceedings of the Second International Conference on 3-D Digital Imaging and Modelling (3DIM'99)', pp. 358 – 367.

Castellano, G. (1999), Investigation and Application of a Complex Wavelet Transform for the Estimation of Optical Flow, PhD thesis, King's College, University of London, London.

Chang, K. I. & Bowyer, K. W. (2005), 'An evaluation of multimodal 2d+3d face biometrics', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(4), 619-624.

Chang, K. I., Bowyer, K. W. & Flynn, P. J. (2003), Face recognition using 2d and 3d facial data, *in* 'Workshop on Multimodal User Authentication (MMUA)', Santa Barbara, California, USA.

Chen, C.-F., Tseng, Y.-S. & Chen, C.-Y. (2003), Combination of pca and wavelet transforms for face recognition on 2.5d images, *in* D. G. Bailey, ed., 'Image and Vision Computing New Zealand 2003', Massey University, North Palmerston, New Zealand, pp. 343–347.

Cheng, Y. C. & In, S. Y. (1985), 'Wave form correlation by tree matching', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **7**(3), 299 – 305.

Chien, J.-T. & Wu, C.-C. (2002), 'Discriminant waveletfaces and nearest feature classifiers for face recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(12), 1644–1649.

Choudhury, T., Clarkson, B., Jebara, T. & Pentland, A. (1999), Multimodal person recognition using unconstrained audio and video, *in* '2nd International Conference on Audio and Video based Biometric Person Authentication', Washington D. C., USA.

Cochran, S. D. & Medioni, G. (1992), '3d surface description from binocular stereo', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 981 – 994.

Cohen, A., Daubechies, I. & Feauveau, J. C. (1992), 'Biorthogonal bases of compactly supported wavelets', *Communications of Pure and Applied Mathematics* **45**, 485 – 560.

Cooklev, T., Berbecel, G. I. & Venetsanopulos, A. N. (2000), 'Wavelets and differential-dilation equations', *IEEE Transactions on Signal Processing* **48**(8), 2258–2268.

Cootes, T., Edwards, G. & Taylor, C. (2001), 'Active appearance models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 681–685.

Cootes, T., Taylor, C., Cooper, D. & Graham, J. (1995), 'Active shape models - their training and application', *Computer Vision and Image Understanding* **61**, 18–23.

Cootes, T., Walker, K. & Taylor, C. (2000), View-based active appearance models, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition'.

Cottrell, G. & Fleming, M. (1990), Face recognition using unsupervised feature extraction, *in* 'International Neural Network Conference'.

Craw, I., Costen, N., Kato, T. & Akamatsu, S. (1999), 'How should we represent faces for automatic recognition?', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(8), 725 – 736.

Daubechies, I., Guskov, I., Schröder, P. & Sweldens, W. (1999), 'Wavelets on irregular point sets', *Philosophical Transactions: Mathematical, Physical and Engineering Sciences (Series A), The Royal Society, London* **357**(1760), 2397–2413.

del Solar, J. R. & Navarrete, P. (2005), 'Eigenspace-based face recognition: A comparative study of different approaches', *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews* **35**(3), 315–325.

Devore, J. (1987), *Probability and Statistics for Engineering and Sciences*, Brooks/Cole, Belmont, CA, USA.

DeVore, R. A., Jawreth, B. & Lucier, B. J. (1992), 'Image compression through wavelet transform coding', *IEEE Transactions on Information Theory* **38**, 719 – 746.

Dick, A. (1997), Multiresolution stereo image matching using wavelets, Master's thesis, University of Adelaide.

Draper, B., Baek, K., Bartlett, M. S. & Beveridge, J. R. (2003), 'Recognizing faces with pca and ica', *Computer Vision and Image Understanding (Special Issue on Face Recognition)* **91**(1-2), 115–137.

Edelman, S., Reisfeld, D. & Yeshurun, Y. (1992), Learning to recognize faces from examples, *in* 'ECCV '92', ECCV, Santa Margherita Ligure, pp. 787 – 791.

Egnal, G., Mintz, M. & Wildes, R. (2002), A stereo confidence metric using single view imagery, *in* 'CIPPRS/IAPR International Conference on Vision Interface 2002'.

Ekenel, H. K. & Sankur, B. (2005), 'Multiresolution face recognition', *Image and Vision Computing* **23**, 469–477.

Etemad, K. & Chellappa, R. (1997), 'Discriminant analysis for recognition of human face images', *Journal of the Optical Society of America A* **14**, 1724 – 1733.

Faugeras, O. (1993), *Three-dimensional computer vision*, MIT Press.

Faugeras, O. D. (1992), What can be seen in three dimensions with an uncalibrated stereo rig, *in* 'Proceedings of the Second European Conference on Computer Vision', Springer-Verlag, pp. 563–578.

Faugeras, O. D., Luong, Q. T. & Maybank, S. J. (1992), Camera self-calibration: theory and experiments, *in* 'Proceedings of the ECCV92, Vol. 588 of LNCS', Vol. 588, Springer Verlag.

Faugeras, O. & Hebert, M. (1986), 'The representation, recognition and locating of 3-d objects', *International Journal of Robotics Research* **5**(3), 27–52.

Fawcett, T. (2004), Roc graphs: Notes and practical considerations for researchers, Technical Report HPL-2003-4, HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **7**, 179 – 188.

Fleck, M. M. (1991), 'A topological stereo matcher', *International Journal of Computer Vision* **6**(3), 197 – 226.

Fleet, D. J. & Jepson, A. D. (1990), 'Computation of component image velocity from local phase information', *International Journal of Computer Vision* **5**(1), 77 – 104.

Fleet, D., Jepson, A. & Jenkin, M. (1991), 'Phase-based disparity measurement', *CVGIP: Image Understanding* **52**(2), 198–210.

Fromherz, T. (1996), Shape from Multiple Cues for 3D-Enhanced Face Recognition: A Contribution to Error Reduction by Employing Inexpensive 3D Reconstruction Methods, PhD thesis, University of Zurich, Zurich.

Fromherz, T. (1998), Face recognition: a summary of 1995 - 1997, Technical Report TR-98-027, International Computer Science Institute, Berkeley, CA.

Gennert, M. A. & Malin, G. A. (1992), 'Stereo vision using gabor receptive fields', *SPIE Intelligent Robots and Computer Vision 11* **1826**, 64 – 75.

George, N., Dolan, R. J., Fink, G. R., Bayliss, G. C., Russell, C. & Driver, J. (1999), 'Contrast polarity and face recognition in the human fusiform gyrus', *nature neuroscience* **2**(6), 574–580.

Gerbrands, J. J. (1981), 'On relationships between svd, klt and pca', *Pattern Recognition* **14**, 375 – 381.

Ghita, O., Mallon, J. & Whelan, P. F. (2001), Epipolar line extraction using feature matching, *in* 'Proceedings of the Irish Machine Vision and Image Processing Conference 2001', Maynooth, pp. 87 – 95.

Golub, G. H. & Van-Loan, C. F. (1996), *Matrix Computations*, 3rd ed. edn, Johns Hopkins, Baltimore, MD.

Gordon, G. (1992), Face recognition based on depth and curvature features, *in* 'IEEE Computer Society Conference on Computer Vision and Pattern Recognition', IEEE, Champaign, Illinois, USA, pp. 108 – 110. 3DFR.

Gordon, G. G. (1991), Face recognition based on depth maps and surface curvature, *in* 'Proceedings of SPIE, Geometric Methods in Computer Vision', Vol. 1570, San Diego.

Gordon, G. G. (1995), Face recognition from frontal and profile views, *in* 'The International Workshop on Automatic Face and Gesture Recognition', Zurich, Germany.

Graps, A. (1995), 'An introduction to wavelets', *IEEE Computational Science and Engineering* **2**(2), 50–61.

Guéziec, A. (1995), Surface simplification with variable tolerance, *in* 'Proceedings of Medical Robotics and Computer Assisted Surgery (MRCAS'95)', pp. 132–139.

Gühring, J., Brenner, C., Böhm, J. & Fritsch, D. (2000), Data processing and calibration of a cross-pattern stripe projector, *in* 'ISPRS Congress 2000', Vol. 33:5 of *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences(IAPRS)*, International Society for Photogrammetry and Remote Sensing, Amsterdam, Netherlands.

Guo, G., Li, S. Z. & Chan, K. (2000), Face recognition by support vector machines, *in* 'FG '00: Proceedings of the Fourth IEEE International Conference on Face and Gesture Recognition', IEEE Computer Society, Washington DC, USA.

Hanley, J. A. & McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve', *Radiology* **143**, 29–36.

Harris, C. & Stephens, M. (1988), A combined corner and edge detector, *in* '4th Alvey Vision conference', Manchester, pp. 147 – 151.

Hartley, R. (1994), 'Projective reconstruction and invariants from multiple images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(10), 1036 – 1041.

Hartley, R., Gupta, R. & Chang, T. (1992), Stereo from uncalibrated cameras, *in* 'In Proceedings of the Conference on Computer Vision and Pattern Recognition', Urbana-Champaign, Illinois, USA, pp. 761 – 764. SM.

Hartley, R. I. & Sturm, P. (1997), 'Triangulation', *Computer Vision and Image Understanding: CVIU* **68**(2), 146–157.

Hartley, R. & Zisserman, A. (2000), *Multiple View Geometry in Computer Vision*, Cambridge University Press.

Heckbert, P. & Garland, M. (1997), Survey of polygonal surface simplification algorithm, Technical Report CMU-CS-97-TBD, The School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

Heikkila, J. & Silven, O. (1997), A four-step camera calibration procedure with implicit image correction, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition', pp. 1106 – 1112.

Heisele, B., Ho, P. & Poggio, T. (2001), Face recognition with support vector machines: Global versus component-based approach, *in* 'International Conference on Computer Vision (ICCV'01)', Vol. 2, Vancouver, Canada, pp. 688–694.

Heseltine, T., Pears, N. & Austin, J. (2002), Evaluation of image pre-processing techniques for eigenface based face recognition, *in* 'The Proceedings of the Second International Conference on Image and Graphics, SPIE vol. 4875', Vol. 4875, Society of Photo-Optical Instrumentation Engineers, pp. 677 – 685.

Heseltine, T., Pears, N. & Austin, J. (2004*a*), Combining multiple face recognition systems using fisher's linear discriminant, *in* A. K. Jain & N. K. Ratha, eds, 'Biometric Technology for Human Identification: Proceedings of SPIE. vol 5404', Vol. 5404, The International Society for Optical Engineering (SPIE), pp. 470–481.

Heseltine, T., Pears, N. & Austin, J. (2004*b*), Three-dimensional face recognition: A fishersurface approach, *in* 'International Conference on Image Analysis and Recognition ICIAR(2)2004', Porto, Portugal.

Heseltine, T., Pears, N. & Austin, J. (2004*c*), Three-dimensional face recognition: An eigensurface approach, *in* 'International Conference on Image Processing ICIP-2004', IEEE, Singapore.

Heseltine, T., Pears, N. & Austin, J. (2004*d*), Three-dimensional face recognition using surface space combinations, *in* 'British Machine Vision Conference BMVC 2004', Kingston University, London.

Hesher, C., Srivastava, A. & Erlebacher, G. (2003), A novel technique for face recognition using range imaging, *in* 'Seventh International Symposium on Signal Processing and its Applications (ISSPA 2003)', Paris, France.

Hildreth, E. C. (1983), *The measurement of Visual Motion*, Cambridge: MIT Press.

Hill, H., Schyns, P. G. & Akamatsu, S. (1997), 'Information and viewpoint dependence in face recognition', *Cognition* **62**, 201–222.

Hoppe, H. (1996), 'Progressive meshes', *Proceedings of Computer Graphics (SIGGRAPH'96)* pp. 99–108.

Horaud, R. & Csurka, G. (1998), Self-calibration and euclidean reconstruction using motions of a stereo rig, *in* 'Proceedings Sixth International Conference on Computer Vision', IEEE Computer Society Press, Los Alamitos, Ca, Bombay, India, pp. 96–103.

Horn, B. (1987), 'Closed-form solution of absolute orientation using unit quaternions', *Journal of Optical Society of America* 4(4), 629–642.

Howell, A. J. & Buxton, H. (1996), Facial recognition using radial basis function neural networks, *in* 'In Proceedings of British Machine Vision Conference', BMVA, Edinburgh, pp. 455 – 464. FR.

Irani, M. & Anandan, P. (1999), All about direct methods, *in* W. Triggs, A. Zisserman & R. Szeliski, eds, 'Vision Algorithms: Theory and Practise', Springer-Verlag.

Jacobs, C. E., Finkelstein, A. & Salesin, D. (1995), Fast multiresolution image querying, *in* 'Proceedings of SIGGRAPH 95', In Computer Graphics Proceedings, Annual Conference Series, 1995, ACM SIGGRAPH, New York, Los Angeles, California, USA.

Jain, A. K. & Dubes, R. C. (1988), *Algorithms for Clustering Data*, Prentice-Hall, New Jersey.

Jenkin, M. R. M., Jepson, A. D. & Tsotsos, J. K. (1991), 'Techniques for disparity measurement', *CVGIP: Image Understanding* **53**(1), 14 – 30.

Johnson, A. E. (1997), Spin-Images: A Representation for 3-D Surface Matching, PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

Johnson, A. E., Carmichael, O., Huber, D. & Hebert, M. (1998), Toward a general 3-d matching engine: Multiple models, complex scenes and efficient data filtering, *in* 'Proceedings of 1998 Image Understanding Workshop (IUW)', pp. 1097 – 1107.

Johnson, A. E. & Hebert, M. (1997), Recognizing objects by matching oriented points, *in* 'Proceedings of Computer Vision and Pattern Recognition 1997 (CVPR '97)', pp. 684–689.

Johnson, A. E. & Hebert, M. (1999), 'Using spin images for efficient object recognition in cluttered 3d scenes', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(5), 433–449.

Ju, X. & Naftel, A. (1999), Facial shape recovery by feature driven stereo analysis, *in* '10th British Machine Vision Conference (BMVC) '99', pp. 533–542.

Kaya, Y. & Kobayashi, K. (1972), A basic study on human face recognition, *in* S. Watanabe, ed., 'Frontiers of Pattern Recognition', Academic Press, New York, pp. 265 – 289.

Kim, Y. S., Lee, J. & Ha, Y. (1997), 'Stereo matching algorithm based on modified wavelet decomposition process', *Pattern Recognition* **30**(6), 929–952.

Kimmel, R. & Sapiro, G. (2003), 'The mathematics of face recognition', *SIAM News*.

Kingsbury, N. (2000*a*), Complex wavelets and shift invariance, *in* 'IEE Colloquium on Time-Scale and Time-Frequency Analysis and Applications', Iee, London.

Kingsbury, N. (2000*b*), 'Complex wavelets for shift invariant analysis and filtering of signals', *Journal of Applied Computation and Harmonic Analysis*.

Kingsbury, N. & Magarey, J. (1997), Wavelet transforms in image processing, *in* 'Signal Processing and Prediction I', European Association For Signal Processing - EURASIP, ICT Press, Prague 1997, pp. 23–34.

Kirby, M. & Sirovich, L. (1990), 'Application of the karhunen-loeve procedure for the characterization of human faces', *IEEE Transactions on Pattern Analysis and Machine Learning* **12**(1), 103 – 108.

Klette, R., Koschan, A., Schlüns, K. & Rodehorst, V. (1995), Evaluation of surface reconstruction methods, *in* 'Proc. New Zealand Image and Vision Computing Workshop '95', Lincoln, Canterbury, New Zealand, pp. 3–12.

Kobayashi, M. (2000), Wavelet analysis: Applications in industry, *in* T.-X. He, ed., 'Wavelet Analysis and Multiresolution Methods', Vol. 212 of *Lecture Notes in Pure and Applied Mathematics*, Marcel Decker.

Kohonen, T. (1988), *Self Organization and Associative Memory*, Springer Verlag, Berlin.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J. & Abidi, M. A. (2005), 'Recent advances in visual and infrared face recognition - a review', *Computer Vision and Image Understanding* **97**, 103–135.

Kontanzad, A., Bokil, A. & Lee, Y. W. (1993), 'Stereopsis by constraint learning feed-forward neural networks', *IEEE Transactions on neural networks* **4**(2), 332 – 342.

Kwong, J. N. S. & Gong, S. (1999), Learning support vector machines for a multi-view face model, *in* 'British Machine Vision Conference (BMVC'99)', Nottingham, UK.

Lam, L. & Suen, C. Y. (1997), 'Application of majority voting to pattern recognition: An analysis of its behavior and performance', *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans* **27**(5), 553 – 568.

Lanitis, A., Taylor, C. & Cootes, T. (1997), 'Automatic interpretation and coding of face images using flexible models', *IEEE Transactions on Pattern Analysis and Machine Learning* **19**(7), 743–756.

Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. (1997), 'Face recognition: A convolutional neural network approach', *IEEE Transactions on Neural Networks* **8**(1), 98–113.

Lawton, W. (1993), 'Applications of complex valued wavelet transforms to subband decomposition', *IEEE Transactions on Signal Processing* pp. 3566–3568.

Lee, D. D. & Seung, S. S. (1999), 'Learning the parts of objects by non-negative factorization', *Nature* **401**, 788–791.

Lee, J.-J., Shim, J.-C. & Ha, Y.-H. (1994), 'Stereo correspondence using hopfield neural network of new energy function', *Pattern Recognition* **27**, 1513 – 1522.

Lee, J., Moghaddam, B., Pfister, H. & Machiraju, R. (2004), Finding optimal view for 3d face shape modelling, Technical Report TR2004-024, Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, Massachusetts 02139. Published in: IEEE International Conference on Automatic Face and Gesture Recognition (FG'04).

Lengagne, R., Fua, P. & Monga, O. (2000), '3d stereo reconstruction of human faces driven by differential constraints', *Image And Vision Computing* **18**(4), 337 – 343.

Levine, M. D., O'Handley, D. A. & Yagi, G. M. (1973), 'Computer determination of depth maps', *Computer Graphics and Image Processing* **2**, 131 – 150.

Levoy, M. (2000), 'Faq list compiled as part of the publicity material for the digital michaelangelo project'.

Li, Q., Ye, J. & Kambhametu, C. (2004), Linear projection methods in face recognition under unconstrained illuminations: A comparative study, *in* '2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04) - Volume 2', Vol. 2, pp. 474–481.

Li, S. Z. & Jain, A. K., eds (2005), *Handbook of Face Recognition*, Springer.

Lin, S. H., Kunga, S. Y. & Lin, L. J. (1997), 'Face recognition/detection by probabilistic decision-based neural network', *IEEE Transactions on Neural Networks* **8**(1), 114 – 132.

Lin, T. & Barron, J. L. (1994), Image reconstruction error for optical flow, *in* 'Vision Interface', Banff National Park, Alberta, Canada, pp. 73 – 80.

Lin, Y., Gong, S. & Liddell, H. (2003), 'Constructing facial identity surfaces for recognition', *International Journal of Computer Vision* **53**(1), 71–92.

Lina, J. (1996), 'Image processing with complex daubechies wavelet'.

Lina, L. M. & Mayrand, M. (1993), Complex daubechies wavelets, Technical Report UdeM-PHYSNUM-ANS-15, Laboratory of Nuclear Physics, University of Montreal.

Liu, C. H., Collin, C. A. & Chaudhuri, A. (2000), 'Does face recognition rely on encoding of 3-d surface? examining the role of shape-from-shading and shape-from-stereo', *Perception* **29**, 729–743.

Liu, C. & Wechsler, H. (2000), 'Evolutionary pursuit and its application to face recognition', *IEEE Transactions on Pattern Analysis and Machine Analysis* **22**, 570–582.

Loève, M. M. (1955), *Probability Theory*, Van Nostrand, Princeton.

Lu, X. (2003), Image analysis for face recognition, Personal notes, Dept. of Computer Science & Engineering, Michigan State University, East Lansing, MI, 48824.

Lu, X., Colbry, D. & Jain, A. K. (2004*a*), Matching 2.5d scans for face recognition, *in* 'Proceedings of International Conference on Biometric Authentication', Hong Kong, pp. 30–36.

Lu, X., Colbry, D. & Jain, A. K. (2004*b*), Three-dimensional model based face recognition, *in* 'Proceedings of the IEEE International Conference on Pattern Recognition', Cambridge, UK, pp. 362–366.

Lu, X. & Jain, A. K. (2005*a*), Deformation analysis for 3d face matching, *in* 'Proceedings of Seventh IEEE Workshop on Applications of Computer Vision(WACV) 2005', Breckenridge, Colorado, USA, pp. 99–104.

Lu, X. & Jain, A. K. (2005*b*), Integrating range and texture information for 3d face recognition, *in* 'Proceedings of Seventh IEEE Workshop on Applications of Computer Vision(WACV) 2005', Breckenridge, Colorado, USA, pp. 156–163.

Lu, X. & K.Jain, A. (2005), Multimodal facial feature extraction for automatic 3d face recognition, Technical Report MSU-CSE-05-22, Computer Science and Engineering Dept., Michigan State University.

Magarey, J. (1997), Motion Estimation using Complex Wavelets, PhD thesis, University of Cambridge, Department of Engineering, Trinity College, Cambridge.

Magarey, J. & Dick, A. (1998), Multiresolution stereo image matching using complex wavelets, *in* '14th International Conference on Pattern Recognition (ICPR)', Vol. 1, pp. 4–7.

Magarey, J., Dick, A., Brooks, M. & Newsam, G. (1999), Incorporating the epipolar constraint into a multiresolution algorithm for stereo image matching, *in* 'Applied Informatics '99, 17th IASTED International Conference'.

Magarey, J. & Kingsbury, N. (1995), Motion estimation using complex wavelets, Technical Report TR-226, Cambridge University Engineering Department.

Magarey, J. & Kingsbury, N. (1996), An improved motion estimation algorithm using complex wavelets, *in* 'Proceedings of IEEE International Conference on Image Processing', IEEE, pp. 969 – 972.

Magarey, J. & Kingsbury, N. (1998*a*), 'Motion estimation using a complex-valued wavelet transform', *IEEE Transactions on Signal Processing* **46**(4), 1069 – 1084.

Magarey, J. & Kingsbury, N. (1998*b*), 'Motion estimation using complex wavelets', *IEEE Transactions on Signal Processing, special issue on Wavelets and Filter Banks* **46**(4), 1069 – 1084.

Mallat, S. G. (1989*a*), 'Multifrequency channel decomposition of images and wavelet models', *IEEE Transactions on Acoustics, Speech adn Signal Processing* **37**, 2091 – 2110.

Mallat, S. G. (1989*b*), 'A theory for multiresolution signal decomposition: a wavelet representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674 – 693.

Mallat, S. G. & Zhong, S. (1992), 'Characterization of signals from multiscale edges', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 710 – 732.

Marapane, S. B. & Trivedi, M. M. (1989), 'Region-based stereo analysis for robotic applications', *IEEE Transactions of Man Cybernet* **19**, 1447 – 1464.

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

Marr, D. & Poggio, T. (1979), 'A computational theory of human stereo vision', *Proceedings of the Royal Society of London B* **204**, 301 – 308.

Marshall, A. D. (1994), Vision systems, Lecture Notes.

Martínez, A. M. & Kak, A. C. (2001), 'Pca versus lda', *IEEE Transactions of Pattern Analysis and Machine Intelligence* **23**(2), 228–233.

Matthies, L. (1992), 'Stereo vision for planary rovers: stochastic modeling to near real-time implementation', *International Journal of Computer Vision* **8**(1), 71 – 91.

Medioni, G. & Nevatia, R. (1985), 'Segment-based stereo matching', *Computer Vision, Graphics and Image Processing* **31**, 2 – 18.

Medioni, G. & Waupotitsch, R. (2003), Face modeling and recognition in 3-d, *in* 'Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG '03)', pp. 232–233.

Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J. (2000), *Wavelet Toolbox: For use with MATLAB*, version 2 edn, The Math Works Inc.

Moghaddam, B., Jebara, T. & Pentland, A. (1999), Bayesian modeling of facial similarity, *in* 'Advances in Neural Information Processing Systems 11', MIT Press.

Moghaddam, B. & Pentland, A. (1997), 'Probabilistic visual learning for object representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 696–710.

Moghaddam, B., Wahid, W. & Pentland, A. (1998), Beyond eigenfaces: probabilistic matching for face recognition, *in* '3rd International Conference on Automatic Face and Gesture Recognition', pp. 30 – 35.

Moravec, H. P. (1977), Towards automatic visual obstacle avoidance, *in* 'Fifth International Joint Conference on Artificial Intelligence', Vol. 5:1, pp. 584 – 589.

Nakayama, K. (1996), 'Binocular visual surface perception', *Proceedings of the National Academy of Sciences* **1996**, 634–639. Colloquium Paper.

Nasrabadi, N. M. & Choo, C. Y. (1992), 'Hopfield network for stereo vision correspondence', *IEEE Transactions on Neural Networks* **3**(1), 5 – 13.

Nefian, A. V. & III, M. H. H. (1998), Hidden markov models for face recognition, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', Seattle, USA, pp. 2721–2724.

Ohta, Y. & Kanade, T. (1985), 'Stereo by intra- and inter-scanline search', *IEEE Transactions on Pattern and Machine Analysis* **7**(2), 139 – 154.

Okutomi, M. & Kanade, T. (1993), 'A multiple-baseline stereo', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 353 – 363.

Oliensis, J. & Govindu, V. (1999), 'An experimental study of projective structure from motion', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(7), 665 – 671.

Owens, R. (1997), 'Computer vision (it412) lecture notes'. Lecture 11.

Pan, H. (1996*a*), General stereo image matching using symmetric complex wavelets, *in* 'Wavelet Applications in Signal and Image Processing IV', Vol. volume 2825 of Proceedings of SPIE, pp. 697–721.

Pan, H. & Magarey, J. (1999), 'Multiresolution phase-based bidirectional stereo matching with provision for discontinuity and occlusion'.

Pan, H.-P. (1996*b*), 'Uniform full-information matching using complex conjugate wavelet pyramids', *Zeitschrigt fuer Photogrammetrie und Fererkundung (German Journal of Photogrammetry and Remote Sensing* **100**, 64 – 93.

Park, M. J., Choi, M. G. & Shin, S. Y. (2002), Human motion reconstruction from inter-frame feature correspondences of a single video stream using motion library, *in* 'SCA '02: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer Animation', ACM Press, New York, NY, USA, San Antonio, Texas, pp. 113 – 120.

Pastor, L., Rodrguez, A. & Insa, D. R. (1999), Wavelets for object representation and recognition in computer vision, *in* E. V. B. Y & M. P, eds, 'Bayesian Inference in Wavelet Based Models', Springer.

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

Pentland, A., Moghaddam, B. & Starner, T. (1994), View-based and modular eigenspaces for face recognition, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)', Seattle, WA.

Phillips, P. J. (1998), Support vector machines applied to face recognition, *in* M. S. Kearns, S. Solla & D. Cohen, eds, 'Advances in Neural Information Processing Systems', Vol. 11, MIT Press, pp. 803–809.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J. & Worek, W. (2005), Overview of the face recognition grand challenge, *in* 'IEEE Conference on Computer Vision and Pattern Recognition 2005', San Diego, CA, USA.

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

Phillips, P. J. & Newton, E. M. (2002), Meta-analysis of face recognition algorithms, *in* 'Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)', Washington, USA.

Pilu, M. (1997), A direct method for stereo correspondence based on singular value decomposition, *in* 'IEEE International Conference on Computer Vision and Pattern Recognition', Puerto Rico.

Poggio, T. & Girosi, F. (1990), Networks for approximation and learning, *in* 'Proceedings of IEEE', Vol. 78, pp. 1481 – 1497.

Polikar, R. (1999), 'The engineer's ultimate guide to wavelet analysis: The wavelet tutorial'.

Pollard, S. B., Mayhew, J. E. W. & Frisby, J. P. (1985), 'Pmf: a stereo correspondence algorithm using a disparity gradient limit', *Perception* **14**, 449 – 470.

Pollefeys, M. (2000), 'Tutorial on 3d modelling from images'. In conjunction with ECCV 2000, Dublin, Ireland.

Pontil, M. & Verri, A. (1998), 'Support vector machines for 3d object recognition', *IEEE Transactions of Pattern Analysis and Machine Intelligence* **20**, 637 – 646.

Prasad, L. & Iyengar, S. S. (1997), *Wavelet Analysis with Applications to Image Processing*, CRC Press LLC.

Read, J. C. (2002), 'A bayesian approach to the stereo correspondence problem', *Neural Computation* **14**(2), 1371 – 1392.

Rizvi, S., Phillips, P. J. & Moon, H. (1998), The feret verification testing protocol for face recognition algorithms, Technical Report NISTIR 6,281, National Institute for Standards and Technology.

Roobaert, D. & Hulle, M. M. V. (1999), View-based 3d object recognition with support vector machines, *in* 'Proceedings of IEEE International Workshop on Neural Networks for Signal Processing (NNSP99)', Madison, Wisconsin.

Rottensteiner, D. F. (2001), Semi-automatic extraction of buildings based on hybrid adjustment using 3D surface models and management of building data in a TIS, PhD thesis, Vienna University of Technology, Vienna, Austria.

Ruf, A., Csurka, G. & Horaud, R. (1998), Projective translations and affine stereo calibration, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE Computer Society Press, pp. 475–481. CC.

Ruiz-Correa, S., Shapiro, L. G. & Melia, M. (2001), A new signature-based method for efficient 3d object recognition, *in* 'Proceedings of Computer Vision and Pattern Recognition 2001 (CVPR '01)', pp. 769–776.

S. G, M. & Hwang, W. L. (1992), 'Singularity detection and processing with wavelets', *IEEE Transactions on Information Theory* **38**, 617 – 643.

Samaria, F. (1993), Face segmentation for identification using hidden markov models, *in* 'British Machine Vision Conference (BMVC)', British Machine Vision Association, BMVA Press, pp. 399–408.

Samaria, F. & Harter, A. C. (1994), Parametrisation of a stochastic model for human face identification, *in* 'Proceedings of Second IEEE Workshop on Applications of Computer Vision'.

Sanger, T. D. (1988), 'Stereo disparity using gabor filters', *Biological Cybernetics* **59**, 405 – 418.

Scharstein, D. & Szeliski, R. (2001), A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Technical Report MSR-TR-2001-81, Microsoft Research, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA.

Scharstein, D., Szeliski, R. & Zabih, R. (2001), A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *in* 'IEEE Workshop on Stereo and Multi-Baseline Vision (in conjunction with IEEE CVPR 2001', Kauai, Hawaii, pp. 131–140.

Schroeder, W., Zarge, J. & Lorenson, W. (1992), 'Decimation of triangular meshes', *Proceedings of Computer Graphics (SIGGRAPH'92)* pp. 65–70.

Scott, G. & Higgins, H. L. (1991), 'An algorithm for associating the features of two patterns', *Proceedings of Royal Society London* **B244**, 21–26.

Shakhnarovic, G. & Moghaddam, B. (2004), Face recognition in subspaces, *in* S. Z. Li & A. K. Jain, eds, 'Handbook of Face Recognition', Springer Verlag 2004. Technical Report TR2004-01 for Mitsubishi Electric Research Laboratories, http://www.merl.com.

Shashua, A. (1994), 'Projective structure from uncalibrated images: Structure from motion and recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(8), 778 – 790.

Shim, M. (2000), Wavelet-based stereo, *in* S.-W. Lee, H. H. Bülthoff & T. Poggio, eds, 'Biologically motivated computer vision', Vol. 1811 of *Lecture Notes in Computer Science*, Biologically motivated computer vision : first IEEE International Workshop, Springer, pp. 326–335.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H. & Heeger, D. J. (1992), 'Shiftable multiscale transforms', *IEEE Transactions on Information Theory* **38**(2), 587 – 607.

Sirovich, L. & Kirby, M. (1987), 'Low-dimensional procedure for characterization of human faces', *Journal of the Optical Society of America A* 4(3), 519 – 524.

Smith, S. (1995), Susan - a new approach to low level image processing, Technical Report Internal Technical Report TR95SMS1, Defence Research Agency, Chobham Lane, Chertsey, Surrey, UK.

Spies, H. & Ricketts, I. (2000), Face recognition in fourier space, *in* 'Vision Interface 2000', Montreal, Canada, pp. 38–44.

Strang, G. & Nguyen, T. (1997), *Wavelets and Filter Banks*, Wellesley - Cambridge Press.

Strela, V., Heller, P. N., Strang, G., Topiwala, P. & Heil, C. (1995), The application of multiwavelet filter banks to image processing, unpublished.

Sturm, P. (1997), 'Self-calibration of a moving zoom-lens camera by pre-calibration', *Image and Vision Computing* **15**(8), 583–589.

Sturm, P. & Quan, L. (1995), Affine stereo calibration, *in* '6th International Conference on Computer Analysis of Images and Patterns (CAIP'95)', Prague, Czech Republic, pp. 838 – 843.

Suthankar, G. (1997), Face recognition: A critical look at biologically-inspired approaches, http://www-2.cs.cmu.edu/ gitars/16-721/final/final.html.

Swets, D. L. & Weng, J. (1996*a*), Discriminant analysis and eigenspace partition tree for face and object recognition from views, *in* 'International Conference on Automatic Face and Gesture Recognition', pp. 192 – 197.

Swets, D. L. & Weng, J. (1996*b*), 'Using discriminant eigenfeatures for image retrieval', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(8), 831 – 836.

Szeliski, R. (1999), Prediction error as a quality metric for motion and stereo, *in* 'Seventh International Conference on Computer Vision (ICCV '99)', Kerkyra, Greece, pp. 781–788.

Taubin, G. (1995), Curve and surface smoothing without shrinkage, *in* 'ICCV', pp. 852–857.

Tezopoulos, D. (1983), 'Multilevel computational processes for visual surface reconstruction', *International Journal of Computer Vision, Graphics and Image Processing* **24**, 52 – 96.

Todd, J. T. (2002), Perception of three-dimensional structure, *in* M. A. Arbib, ed., 'The Handbook of Brain Theory and Neural Networks', second edn, MIT Press, pp. 868–871.

Tolba, A. S., El-Baz, A. H. & El-Harby, A. A. (2005), 'Face recognition: A literature review', *International Journal of Signal Processing* **2**(1), 88–103.

Torr, P. H. S. (2002), A structure and motion toolkit in matlab: "interactive adventures in s and m", Technical Report MSR-TR-2002-56, Microsoft Research, Cambridge.

Torr, P. H. S. & Zisserman, A. (1999), Feature based methods for structure and motion estimation, *in* 'Workshop on Vision Algorithms', pp. 278 – 294.

Torres, L., Reutter, J. Y. & Lorente, L. (1999), The importance of color information in face recognition, *in* 'IEEE International Conference on Image Processing', Kobe, Japan.

Troje, N. F. & Vetter, T. (1996), Representations of human faces, Technical Report 41, Max-Planck Institute for Biological Cybernetics, Germany. 3DF.

Tsai, R. Y. (1986), An efficient and accurate camera calibration technique for 3d machine vision, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition', Miami Beach, Florida, pp. 364 – 374.

Tsalakanidou, F., Tzovaras, D. & Strintzis, M. G. (2003), 'Use of depth and colour eigenfaces for face recognition', *Pattern Recognition Letters* **24**, 1427–1435.

Turk, G. (1992), 'Re-tiling polygonal surfaces', *Proceedings of Computer Graphics (SIGGRAPH'92)* pp. 55–64.

Turk, M. A. & Pentland, A. P. (1991*a*), Face recognition using eigenfaces, *in* 'Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Hawaii, pp. 586–590.

Turk, M. & Pentland, A. (1991*b*), 'Eigenfaces for recognition', *Journal of Cognitive Neuroscience* **3**(1), 71 – 86.

Uchida, N., Shibahara, T., Aoki, T., Nakajima, H. & Kobayashi, K. (2005), 3d face recognition using passive stereo vision, *in* 'Proceedings of the 2005 IEEE International Conference on Image Processing (ICIP '00)', Vol. II, pp. 950–953.

Valens, C. (1999), 'A really friendly guide to wavelets', http://perso.wanadoo.fr/polyvalens/clemens/wavelets/wavelets.html.

Valentine, T. (1991), 'A unified account of the effects of distinctiveness, inversion and race in face recognition', *Quarterly Journal of Experimental Psychology* **43A**, 161–204.

Vapnik, V. N. (1998), *Statistical Learning Theory*, John Wiley & Sons, New York.

Vincent, E. & Laganière, R. (2002), An empirical study of some feature matching strategies, *in* 'Proc. 15th International Conference on Vision Interface', Calgary, Canada, pp. 139–145.

Voth, D. (2003), 'Face recognition technology', *IEEE Intelligent Systems* **18**(3), 4–7.

Weng, J., Ahuja, N. & Huang, T. S. (1992), 'Matching two perspective views', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 806 – 825.

Weng, J. & Swets, D. (1999), Face recognition, *in* A. Jain, R. Bolle & S. Pankanti, eds, 'Biometrics: Personal Identification in Networked Society', Kluwer Academic, Boston, MA, chapter 1, pp. 67–86.

Weyrauch, B., Huang, J., Heisele, B. & Blanz, V. (2004), Component-based face recognition with 3d morphable models, *in* 'First IEEE Workshop on Face Processing in Video', Washington D. C., USA.

Wilson, R., Calway, A. D. & Pearson, E. R. S. (1992), 'A generalized wavelet transform for fourier analysis: the multiresolution fourier transform and its application to image and audio signal analysis', *IEEE Transactions on Information Theory* **38**, 674 – 690.

Wilson, R. G. (1994), Modeling and Calibration of Automated Zoom Lenses, PhD thesis, Department of Electrical and Computer Engineering,Carnegie Mellon University.

Wilson, R. G. & Shafer, S. A. (1993), A perspective projection camera model for zoom lenses, *in* 'Proceedings of the 2nd Conference on Optical 3-D Measurement Techniques', Zurich, Switzerland.

Winkler, J. (n.d.), 'Lecture notes: Wavelets and filter banks'.

Wiskott, L., Fellous, J. M., Kruger, N. & von der Malsburg, C. (1997), 'Face recognition by elastic bunch graph matching', *Pattern Analysis and Machine Learning* **19**(7), 775–779.

Wu, Y. (2004), 'Stereo and multiview geometry'. ECE432 - Advanced Computer Vision Notes, Electrical and Computer Engineering, Northwestern University, Evanston, IL60208.

Wundrich, I. J., von der Malsburg, C. & Würtz, R. P. (2000), Image representation by the magnitude of the discrete gabor wavelet transform, IEEE Transactions on Image Processing, in revision.

Würtz, R. P. (2002), Face recognition: Neurophysiology and neural technology, *in* M. A. Arbib, ed., 'The Handbook of Brain Theory and Neural Networks', second edn, MIT Press, pp. 434–437.

Xu, G. (1997), 'A unified approach to image matching and segmentation in stereo, motion, and object recognition via recovery of epipolar geometry', *Journal of Computer Vision Research.*

Yambor, W. S., Draper, B. A. & Beveridge, J. R. (2002), Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures, *in* H. Christensen & J. Phillips, eds, 'Empirical Evaluation Measures in Computer Vision', World Scientific Press, Singapore.

Yang, M.-H. & Ahuja, N. (2000), A geometric approach to train support vector machines, *in* 'IEEE Conference of Computer Vision and Pattern Recognition (CVPR2000)', Hilton Head Island.

Yoda, I. & Sakaue, K. (2003), 'Utilization of stereo disparity and optical flow information for the computer analysis of human interactions', *Machine Vision and Applications* **13**, 185–193.

Zhang, J., Yan, Y. & Lades, M. (1997), 'Face recognition: Eigenface, elastic matching and neural nets', *Proceedings of the IEEE* **85**(9), 1423 – 1435.

Zhang, L. (2003), Face recognition from still images: An overview and two proposed methods, Research Proficiency Exam, Advisor: Dimitris Samaras, Computer Science Department, State University of New York at Stony Brook.

Zhang, Z. (1994), 'Iterative point matching for registration of free-form curves and surfaces', *International Journal of Computer Vision* **13**(2), 119–152.

Zhang, Z. (2000), 'A flexible new technique for camera calibration', *IEEE Transactions on Pattern Recognition and Machine Intelligence* **22**(11), 1330 – 1334.

Zhang, Z., Deriche, R., Faugeras, O. D. & Luong, Q.-T. (1995), 'A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry', *Artificial Intelligence* **78**(1-2), 87–119.

Zhao, W. (1999), Subspace methods in object/face recognition, *in* 'International Joint Conference on Neural Networks'.

Zhao, W., Chellappa, R. & Krishnaswamy, A. (1998), Discriminant analysis of principal components for face recognition, *in* 'International Conference on Automatic Face and Gesture Recognition', pp. 336 – 341.

Zhao, W., Chellappa, R. & Phillips, P. J. (1999), Subspace linear discriminant analysis for face recognition, Technical Report CAR - TR - 914, Centre for Automation Research, University of Maryland.

Zhao, W., Chellappa, R., Phillips, P. & Rosenfeld, A. (2003), 'Face recognition: A literature survey', *ACM Computing Surveys* **35**(4), 399 – 458.

Zhao, W., Chellappa, R., Rosenfeld, A. & Phillips, P. J. (2000), Face recognition: A literature survey, Technical Report CAR-TR-948, University of Maryland, Centre for Automation Research (CfAR), College Park, MD 20742-3275.

Zhao, W., R.Chellappa & Phillips, P. J. (1999), Subspace linear discriminant analysis for face recognition, Technical Report CAR-TR-914, Centre for Automation Research, University of Maryland, College Park, MD 20742-3275.

Zisserman, A. (1997), 'Geometric framework for vision i: Single view and two view geometry'.