**Thesis Submitted for the Degree of**
**Master of Engineering**

# Content Based Image Pose Manipulation

**Author: Seán Stephen Begley**

**Supervisor: Professor Paul F. Whelan**

**Dublin City University**
**School of Electronic Engineering**
**November 2008**

*I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Engineering by research is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.*

Signed: _____ ID No.: 51324128

*Candidate*

Date: $28^{th}$ November 2008

## Acknowledgements

First of all, I would like to thank Prof. Paul F. Whelan for supervising me and providing me with valuable direction and support throughout the course of my work. I would like to thank Science Foundation Ireland for funding this research. I would also like to thank Dr. John Mallon who was constantly available to offer advice and opinions. Thanks are also due to the members of the Vision Systems Group who provided assistance and input when it was asked of them especially Ms. Patricia Moore, Mr. Aubrey Dunne and Mr. Brendan Byrne. Finally, I would like to thank my family and friends who provided moral support, friendship and most importantly a link to the real world!

# Contents

# Content Based Image Pose Manipulation

**Seán Begley**

**Abstract**

This thesis proposes the application of space-frequency transformations to the domain of pose estimation in images. This idea is explored using the Wavelet Transform with illustrative applications in pose estimation for face images, and images of planar scenes. The approach is based on examining the spatial frequency components in an image, to allow the inherent scene symmetry balance to be recovered. For face images with restricted pose variation (looking left or right), an algorithm is proposed to maximise this symmetry in order to transform the image into a fronto-parallel pose. This scheme is further employed to identify the optimal frontal facial pose from a video sequence to automate facial capture processes. These features are an important pre-requisite in facial recognition and expression classification systems. The underlying principles of this spatial-frequency approach are examined with respect to images with planar scenes. Using the Continuous Wavelet Transform, full perspective planar transformations are estimated within a featureless framework. Restoring central symmetry to the wavelet transformed images in an iterative optimisation scheme removes this perspective pose. This advances upon existing spatial approaches that require segmentation and feature matching, and frequency only techniques that are limited to affine transformation recovery. To evaluate the proposed techniques, the pose of a database of subjects portraying varying yaw orientations is estimated and the accuracy is measured against the captured ground truth information. Additionally, full perspective homographies for synthesised and imaged textured planes are estimated. Experimental results are presented for both situations that compare favourably with existing techniques in the literature.

# Glossary of Acronyms

| **Acronym** | – | **Explanation** |
|---|---|---|
| 2D | – | Two Dimensional |
| 3D | – | Three Dimensional |
| CCSAE | – | Constrained Curvature Symmetry-Axis Estimation |
| DLT | – | Direct Linear Transform |
| DOF | – | Degree Of Freedom |
| EBPR | – | Energy Balancing Planar Rectification |
| FOV | – | Field of View |
| GVF | – | Gradient Vector Flow |
| L-M | – | Levenberg-Marquardt |
| LS | – | Least Square |
| RMS | – | Root Mean Square |
| ROI | – | Region of Interest |
| SD | – | Standard Deviation |
| STFT | – | Short Time Fourier Transform |
| WBSC | – | Wavelet-Based Symmetry Coefficient |
| WFPR | – | Wavelet-based Face Pose Removal |

# Chapter 1

# Introduction

Biometrics are continually growing in their everyday usage. Examples include fingerprint and iris scanners for secure access, face recognition in security systems, forensic dental analysis for identification, and recently, gait analysis for person recognition. The increased prominence of biometrics may be attributed to the need for better security and law enforcement measures, and to the advancement in the technology that is used. Faces in particular provide an appealing biometric feature that may be used for person verification or recognition. Faces exhibit large variations from person to person which makes computer recognition of faces an attractive approach (Goldstein et al., 1971). Another motivation for using faces is that the techniques involved are non-intrusive, compared to fingerprint and iris recognition techniques. However, in order for face recognition systems to operate optimally, subjects must maintain neutral expressions and look directly into the camera. Variations in illumination, pose, and expression are not well tolerated and so must be removed Gao and Leung (2002).

This research, being conducted in conjunction with the Computer Vision and Imaging Laboratory (CVIL) in NUI Maynooth, addresses the latter two of these issues. This thesis deals directly with the problem of removing pose from face images using featureless rectification techniques in order to facilitate improved expression removal. Throughout this thesis, methods operating on geometric or image features such as corners, will simply be referred to as feature-based techniques. Methods that operate on images as a whole, without explicit extraction of geometric features will be known as featureless

techniques. Experiments demonstrate improved featureless rectification performance compared with similar methods in the literature.

Currently, the estimation of planar transformations, known as homographies, using featureless techniques is confined to the estimation of affine homographies. This is due to the global nature of frequency domain transformations that can not capture variations in frequency response across an image. Uniquely, the space-frequency domain, obtained through a continuous wavelet transform, can capture frequency responses at each location in an image, allowing for the estimation of perspective homographies in a featureless capacity. A thorough examination of planar image transformations and their estimation in spatial and Fourier image spaces is carried out. An algorithm is developed and applied to the problem of featureless perspective homography estimation in the space-frequency domain. Presented results on synthesised and real images captured with out of plane motions indicate that the proposed algorithm does indeed achieve good perspective pose estimation and removal.

## 1.1 Background and Motivation

This section outlines the development of the work that led to the ideas presented in this thesis. The section is divided in two, detailing the different streams investigated in the research.

### 1.1.1 Removal of Pose from Face Images

Many different methods of face recognition are currently being researched, each of which chooses different aspects of the face to use as the discriminating vector. However, as it is noted in Gao and Leung (2002), each facial recognition approach investigated in the report had decreased recognition rates when dealing with subjects portraying non-neutral expressions and varying pose. Evidence of this may also be found in the literature where typically, only images depicting fronto-parallel neutral faces are used in experimental verification of recognition systems. These suggest that the removal of facial expression and pose would improve facial recognition rates.

This becomes an issue for the user, as slight variations in pose are natural,

and neutral expressions tend to be difficult to achieve. To achieve optimal face recognition performance, efforts have to be made by system creators to remove any expression and random pose that the subject shows after image capture, and allow the user to be imaged as naturally as possible. This will allow greater accessability and ease of use for the subjects.

The process of pose and expression removal is sequential, since generally, in order for the expression removal algorithms to be successful, frontally imaged faces are required as input. Thus, it is pose removal from face images that is considered in this thesis.

### 1.1.2 Image Rectification

Planar pose estimation is generally tackled in one of two domains. Firstly, the spatial domain, where strong geometric features are easily segmented and matched from views of the imaged plane. The relative positions of these features give information about the geometry between the two views. And secondly, the frequency domain where densely textured scenes are transformed into the frequency domain using Fourier and similar transforms. Affine relations may then be extracted from the Fourier transforms of the two views. However, there are associated drawbacks with each of the two domains.

There are many scenarios within which features may be difficult to accurately extract, such as in densely textured scenes or in images with a high noise level. Furthermore, the extracted features in each of two related views may be difficult to match because of the high noise level or the high number of possible feature correspondences. These form the most significant drawbacks to the spatial domain techniques and is why one would consider featureless techniques that are available with the frequency domain.

On the other hand, although some of the limitations of the spatial domain, namely the feature extraction and matching problem, are avoided using the frequency domain, some of the better aspects are also lost. The varying spatial depth of planar objects in scenes can only be accurately captured using full perspective transformation models. As a corollary to this, to estimate a full perspective planar transformation model between two views of a plane, local spatial information is required. The frequency domain, due to its inher-

ent global image transform, fails to capture spatial changes and therefore is deemed unsuitable for estimating perspective planar transformations because information on spatial changes is needed for perspective pose estimation.

Currently, no method exists that captures the spatially varying depth of a planar structure in an image within a featureless framework. Essentially, at present, a choice has to be made whether to use a geometric-feature based technique and capture spatially varying depth, or use a featureless technique and accept the limitation that depth information is lost in the global transform that is employed.

Each has its advantages and disadvantages. Using the geometric-feature based technique, the problems of feature extraction and feature matching between views must be tackled, but will result in being able to capture the spatially varying depth between two views. On the other hand, using the frequency-domain, featureless techniques, the problems associated with feature extraction and matching are avoided but depth information is lost and only affine relations between two views may be derived.

## 1.2 Objective

The main aim of this thesis is to remove pose from images in a featureless manner. The system developed will have to be robust to image noise and to lighting variation, two problems that commonly occur in imaged scenes. Two application domains will be examined within which pose removal will need to be accomplished. The author aims to develop a face pose removal system. The problem will be tackled in a featureless framework since feature identification and localisation on faces is difficult to accomplish. The inherent symmetry in faces will be exploited to aid in removing the pose. And secondly, the more general case of pose removal from images of planar textures will be examined and tackled in a featureless framework, again based on symmetry.

For this reason, two associated problems have to be tackled. Initially, the axis of symmetry of the face or plane will have to be determined, so an examination of axis estimation techniques will be carried out. The most suitable axis estimation algorithm will be improved upon to increase the efficiency and accuracy for inclusion in the rectification systems. This axis of symmetry es-

timation may find other useful application in finding and removing in-plane rotations of planar objects as well as in image segmentation where the axis of an object best indicates its position.

Following on from this, a measurement of "how symmetric" objects appear to be will be carried out. This is because, when objects are imaged in a fronto-parallel view, the measurement of their symmetry value will be at a maximum. An investigation into types of symmetry measures and their application areas will be carried out to select the most suitable for the task at hand. The symmetry measure that is selected will need to be robust to noise and lighting variations that will exist in the application domains. Suitable modifications to the symmetry measure will be carried out to achieve this.

A facial pose removal system will then be built upon the symmetry axis estimation and symmetry measurement processes. The author's axis of symmetry estimation may be used to determine in-plane rotation of the subject's faces. This rotation will be removed such that the axis of symmetry is aligned with the vertical axis to allow easier calculation of the symmetry measure. A number of different pose removal techniques will be used to rectify the face images. With each technique, the proposed symmetry measure will be used as an indicator of whether or not the optimum frontal pose has been obtained.

A further application of the novel symmetry measurement technique will be found in determining the optimum fronto-parallel face image from a video sequence. In many applications, faces are passively imaged without any user interaction. The majority of the images taken will be from non-ideal positions, but with the aid of the proposed symmetry measure the number of stored images will be reduced while the quality will be guaranteed by a high symmetry measurement.

Through the examination of pose removal from face images, the space-frequency domain will be utilised as a tool in the process, to overcome the resolution limitations of the short-time Fourier transform. In the literature, few space-frequency domain techniques for pose removal exist. A more thorough investigation into the use of the space-frequency domain for image pose recovery will be carried out. In particular, planar pose estimation will be examined. The featureless pose removal techniques will be further employed to the estimation of out of plane rotations of textured planar surfaces within which

features are difficult to extract. The aim is to bridge the gap between the spatial domain where perspective pose estimation is possible through feature-based techniques, and the frequency domain where only affine pose estimation is currently possible through featureless techniques. Operating in the space-frequency domain, the author aims to remove perspective pose in a featureless framework.

## 1.3 Literature Review

A significant amount of research has been carried out in the field of pose removal from images to date. Publications dealing with issues relevant to this thesis are discussed under headings corresponding to the chapters of this thesis. The topics covered are removal of pose from face images, and planar transformation estimation.

### 1.3.1 Removal of Pose from Face Images

Much research is being conducted to improve recognition rates under varying conditions such as pose, illumination and expression. Varying pose in particular significantly decreases recognition and verification rates (Gao and Leung, 2002). Many approaches can be applied to improve facial recognition, the majority of which may be categorised into one of three major classes;

1. 3D-model based techniques
2. multiple appearance recognition techniques
3. and generic appearance based methods

The first and last of these classes aim to remove pose, while the second attempts to operate on the posed face images using multiple posed-face models. Recognition is typically carried out on frontally imaged faces, however, as is the case in the multiple appearance techniques, recognition is carried out between posed images and other similarly posed images. 3D model based algorithms require the construction of explicit 3D models for each face through a learning process within which multiple views of each subject are utilised. Generic

techniques on the other hand, do not create specific models for each face and typically remove pose in feature space.

Rather than learning individual feature subspaces for each viewing angle, face pose removal is a more desirable goal. Pose removal is generally less restrictive than the individual-pose-model methods, since many poses may be handled. The majority of the research in the literature attempts to remove facial pose as a pre-requisite to face recognition. As such, the majority of the face pose removal algorithms remove any non-frontal view information in the feature space where the subject discrimination takes place and not directly on the image itself. The 3D model based techniques are the exception to the case, where pose removal occurs prior to subject discrimination.

### 3D-model based techniques

Three-dimensional models of each subject may be computed based on the positions of features and the appearance of the face and used to synthesise new views of the face. These 3D-model based techniques tend to be restrictive in that multiple views of each subject are generally required for the systems to be trained. In Sung and Kim (2008), Kahl and Heyden (1998), Jiang et al. (2002), where a 3D model is employed, 3 views at known positions of each subject are used to construct a 2D+3D Active Appearance Model. This is a cumbersome and time consuming approach and yields classification rates lower than through using frontally imaged faces. Similarly in Zhang and Cohen (2002), generic 3D models of faces are used which are morphed based on the locations of features detected from faces imaged in arbitrary positions. A drawback of these approaches, is that multiple images are required to build the face models. Each view of the face must then be registered with the 3D model, which requires accurate feature detection and localisation as well as correspondence matching to a generic model. These all leave the process open to sources of error.

Contrary to these two approaches where multiple views are required, in Jiang et al. (2002) and Blanz and Vetter (2003) only a single view of each face is required to build the 3D face model. In Jiang et al. (2002), features are extracted from each frontally imaged face, the locations of which are used to linearly weight the shape vectors that describe the 3D face model. The

obtained face model is texture mapped with the imaged face and may be re-rendered in a different view. The obvious restriction of this method is that an initial fronto-parallel image of each subject is required, which is generally not available. In Blanz and Vetter (2003), a learned 3D face model is morphed in an optimisation process so that the distance between projected 3D points and the corresponding 2D points is minimised in an iterative process. There are two significant hindrances to using this method. Firstly, some initial manual feature selection is required to start the iteration process. And secondly, due to the high dimension of the parameter space that defines the shape, illumination and color of the face model to be morphed, the algorithm takes 4.5 minutes to execute on a 2GHz Pentium 4 Workstation (Blanz and Vetter, 2003).

### Multiple appearance recognition techniques

A middle-ground between the 3D-model based methods and the generic methods, the multiple appearance based techniques use multiple views of subjects to create multiple posed recognition spaces. Subspaces are created to represent each view of the faces in a database. These subspaces are created by selecting face images exhibiting the same pose and performing dimensionality reduction on this subset of images. Different subspaces are created for each pose onto which the test images may be projected for vector discrimination (Huang et al., 2000). Pose can be automatically estimated using a multi-view pose model with which to compare the test image (Tsukamoto et al., 1994, Sung and Kim, 2004), but typically, the system user determines the most suitable recognition space that best matches the pose of the presented face for recognition. Requiring multiple recognition spaces makes this approach very restrictive and primarily suited to person verification.

### Generic appearance based methods

Generic appearance based techniques do not explicitly create a 3D model with which to re-render a new view of the face. Instead, relations between views are derived in the feature space within which discrimination between subjects is carried out. For example, Lee and Kim (2006) compute a linear transformation between the basis functions of face images in one view and the basis function representation of the frontally imaged faces. This allows them to project the

basis functions of a presented test image into its fronto-parallel view using the learned transformation matrix. This method is however restrictive in that for each viewing angle, a separate transformation matrix has to be calculated at the training stage, and also the pose of the individual has to be known by the user prior to pose removal.

Similar methods exist that operate on the positions of facial features and are the most prominently researched techniques in this field. Active shape models (ASM) use statistical learning methods to build up a reduced dimension subspace of the facial feature vectors (Lanitis et al., 1997, Edwards et al., 1996). Faces portraying varying expressions and in different poses are used to train the system. The subspaces onto which the feature vectors are projected are spanned by vectors that control pose, expression and inter-person variation. Removing components that control pose and expression leaves the user with feature vectors that only contain useful identification information.

The two appearance based techniques mentioned above use the assumption that the pose components are entirely or almost entirely independent of the inter-person variation vectors. This however is not the case, so through the removal of the primary pose components, some of the essential variations that distinguish subjects' identities may also be removed. This is due to the pose and identity components being somewhat correlated.

### Summary

3D model based approaches provide the most realistic pose removal systems since once the model is known any view of the face my be synthesised by rotating the three dimensional model and re-rendering the image with the new view. There are however some significant drawbacks to the algorithms employed. An issue that occurs with both the generic 3D model and the learned 3D model is that of registering the image and the model. Accurate detection of feature points is vital, a feat that is difficult to achieve. Inaccurate texture mapping of the model could have a significant impact on subsequent expression classification and recognition processes.

### 1.3.2 Image Rectification

Rectification is generally referred to as the manipulation of image geometry to attain certain goals. As such, one incarnation can be interpreted as pose normalisation which is synthesising an image so that the object of interest appears to be imaged in its fronto-parallel position, see Fig. 1.1.

The problem of rectification may be grouped into three very distinct streams of rectification research, two of which, are currently being heavily researched. These three broad research streams are:

1. spatial rectification techniques
2. frequency based rectification
3. and an emerging research area, spatial-frequency rectification

The former two are the areas with much research, and it will be shown that the third is a valid and comparable method, with a future of much investigation and experimentation.

Each research stream has a niche application domain and can operate more efficiently within that niche than the other methods. The application domain for each of the three types of rectification will be described, and some of their characteristics will be highlighted. Also, some particular uses will also be given as examples to highlight their operation and distinctiveness. The logical progression from one stream to the next in search of a suitable rectification technique is detailed below.

#### Spatial Domain Rectification

Encompassing image mappings that are determined directly from the structural information in images, this is the largest stream of research in the field. Some indicative methods include: the Direct Linear Transform (DLT) which maps sets of interest points in one view of a plane to a second view's same interest points; maximising the rectilinearity of a single image; using priori data, such as real world parallel lines, orthogonal line pairs and conics, to compute homographies; and using the relative shape of objects to find the transformation.

(a) Original Image taken from arbitrary position

(b) The rectified image

Fig. 1.1: An example of image rectification.

With DLT rectification methods, the most common techniques, the geometry will be directly obtained from spatial relations between corresponding points. This may be viewed as finding the perspective transformation that will map one set of points in order to find the other, and the points are operated on directly.

It was published in 1841 by Grunert, a German mathematician, that a direct solution to the relative pose problem is retrievable from just three point correspondences. Unfortunately, after a complex method of equation substitutions and the solution of a fourth degree polynomial, only a solution with up to four ambiguities was obtained from the problem. In Haralick et al. (1991), a collection of variations on this method were scrutinised and their performance evaluated. It was found that these methods were very susceptible to noise, and the accuracy depended heavily on the order of substitutions. Also, these methods yield a unique transformation for each set of three point correspondences used, resulting in no globally suitable solution. There were also multiple solutions that could work, giving an ambiguity in the final solution. This problem was addressed in Quan and Lan (1998), where new linear least squares solutions for four or more point correspondences were proposed using Grunert's technique, but it was merely as an academic exercise. This again involved a complicated system of equations, but the ambiguity was removed.

A more straightforward least squares solution was shown in Criminisi et al. (1997), where vector algebra methods were employed, and using four or more point correspondences the two-dimensional transformation matrix solution was found. This method has its roots in the research on finding epipolar geometry and camera matrices from point correspondences, known as the camera calibration problem. It is this area that mainly contributes to the advancement of the spatial domain techniques. Initially, homographies were solved for as an intermediate matrix for camera pose determination (Weng, Huang and Ahuja, 1988). This method progressed and was used as an initial estimate for iterative non-linear least squares methods that followed, where various cost functions are minimised. There exist a multitude of cost functions that can be optimised based on algebraic, geometric, and other distances, measured in one or multiple images. Maximum likelihood, or geometric distance minimisation, techniques are among the most popular, as the cost functions are intuitive and easy to implement, and are used to refine the solution found from the linear methods (Weng, Ahuja and Huang, 1988, Hartley and Zisserman, 2003).

Each of the spatial domain techniques are prone to noise effects due to the methods used for feature detection and correspondence matching. Poorly located features, or mismatched features, may have a large effect on both the linear and iterative stages of transformation estimation (Mallon and Whelan, 2007). To overcome such outliers, a method that uses subsets of the points to determine the pose was proposed, and is known as the Random Sample Consensus method (RANSAC). In this method, initially formulated in Fischler and Bolles (1981), a random selection of correspondences is chosen to estimate the transformation. The number of correspondences that support this transformation are then counted. This is repeated until a high enough count is found, and that is determined to be the best solution. Supporting correspondences are then used together to calculate the transformation and all non-supporting correspondences are considered as outliers. Although RANSAC removes part of the noise limitations of spatial domain techniques, errors still remain.

A priori knowledge of the image may be used in order to remove planar pose, which bypasses the need for correspondence matching. One such method uses the knowledge that the line at infinity should be in its canonical position (Collins and Beveridge, 1993). Projecting this line back to its canonical position affinely rectifies the image, which allows for certain direct measurements to be found. However, this method is very susceptible to error in point localisation, as generally the vanishing points on the line at infinity have coordinates of very large magnitude, which can be largely affected by a single pixel error. Other methods use conics in a scene to rectify the image where the conic is reprojected back to its canonical position (Kahl and Heyden, 1998, Jiang et al., 2002).

Similarly, certain classes of images can be rectified by examining the directions of their edges. In an image with strong rectangular shapes, the majority of the edges should be aligned in one of two directions, orthogonal to each other. An iterative scheme is implemented to restore the orthogonality of the line directions, known as the image's rectilinearity (Rosin and Zunic, 2005). When using the rectilinearity of an image to remove its pose, direct measurements of the angles and lengths of the edges in an image are computed and placed in a histogram of length per angle. Due to the pixelated nature of images, line lengths and the angle categories into which they fall can be difficult to determine and this forms the largest source of error with this technique.

Spatial domain techniques are particularly suited to the areas of camera calibration and machine placement of parts. These are areas in which a strong, well defined grid may be used to locate points of interest in each view of the scene, which will then be easily matched. Rectilinearity, as the name suggests is particularly good for enforcing rectangular objects in the real world to appear as rectangles in the images. It is well suited to ordnance survey applications where buildings can be used to rectify the image such that the plane is orthogonal to the camera's central ray Rosin and Zunic (2005).

The principle drawback to using spatial domain techniques is that feature detection and correspondence matching will always be initially required (with the exception of the rectilinearity and RANSAC methods, which only require feature detection), which leaves the processes open to feature detection errors. There are many factors that contribute to the overall error in these systems, but the most significant errors are due to the feature detection and localisation processes employed. A second issue is that much of the information encoded in the image is lost due to the discarding of non-feature aspects.

### Frequency Domain Rectification

Methods that operate in a non feature-based framework overcome the limitations of the spatial domain, namely difficulty of feature extraction and correspondence matching between different views of the same planar scene. This is accomplished in a more restrictive capacity with less degrees of freedom in the planar transformation model. In order to rectify scenes using a featureless approach, for particular applications the input domain is altered. Due to its well understood properties, the frequency domain would be an obvious alternative to the spatial domain. All of the textural information in the image is used. Additionally, a single image may be used for rectification, provided some information about the frequency components of an image in canonical position are known.

This is a relatively new direction in research, which was spawned by the composition of the affine theorem of the Fourier transform (Bracewell et al., 1993). Since then only a few researchers have used frequency transforms as a method for finding object pose, most notably Lucchese and Cortelazzo (1997), and Lucchese (2000). In the series of papers, Lucchese progressively found rotations,

translations, and affine warping transformations between two views. These were all computed using the Fourier transforms of the images to be registered. Lucchese focused on minimising the error between the radial projection of the Fourier transform magnitude for one image with the radial projection of the Fourier transform magnitude of the second. It is an iterative minimisation technique that works well for synthesised data, but not so well for real data. Since only affine relations may be derived from the Fourier transforms, for real images a weak perspective assumption has to be made, but because this weak perspective assumption doesn't hold up for scenes that aren't imaged from very large distances, affine relations are difficult to extract. Translations between two views can be estimated using the phase of the Fourier transformed images.

The affine theorem of the Fourier transform may also be used to directly compute the affine warping from one image to another. Finding corresponding magnitude peaks in two Fourier transformed images of the same scene, will facilitate a solution for the affine matrix to be found using the Direct Linear Transform operating on the locations of the detected Fourier magnitude peaks. The spatial domain affine transform is then readily available as the inverse of the obtained frequency domain affine transform. This process is susceptible to the same feature detection and localisation errors associated with the spatial domain, but allows for simpler registration of heavily textured scenes.

Both of these techniques may be used to align images, within which features are difficult to extract, for the purposes of mosaicking, where only a portion of the image is used to estimate the relative pose between the images. This will work well provided that the images are taken at a similar scale, the majority of the information in one image portion is in the other, and that the perspective warping is not too severe. Although the Fourier domain techniques work well on densely textured scenes, scenes with regular structured textures such as chessboard patterns, or sparse textures will give poor performance. This is due to the global image transform that is employed to compute the frequency domain representation, which fails to encode spatial information. Also, at present, perspective image rectification is not achievable in the Fourier transform method because a full perspective planar transformation model between two views may only be estimated if there is varying spatial depth in a scene which is lost through the Fourier transform.

A more general approach to global image transformation techniques was proposed in Petrou and Kadyrov (2004), where any global image transformation may be applied to each of the views of the scene. Strong geometric relations then exist between the two sets of transformed data, from which affine relations can be found, (Kadyrov and Petrou, 2006). Lucchese's method in Lucchese (2001c) was shown to be a specialisation of this, as his technique involved the calculation of integrals along horizontal and vertical directions as well as a circus functional which is merely a projection of the Fourier magnitude data integrated along its radius. Because the spatial information is lost through the calculation of the functionals along an entire chord in the image, the ability to derive perspective relations between two images is lost.

Another variation on Lucchese's method uses the Fourier transforms of images of extracted feature points in two views to determine an affine fundamental matrix relation (Lehmann et al., 2007). Some assumptions are made about the setup of the scene to allow for the solution to be found. The two-dimensional captured images are assumed to be orthographic projections of the real world scene, and the motion of the camera is constrained to rotation and translation. If the scales are the same, then using the assumptions made, there will exist a line in each of the Fourier spectra of the two images that will be the same. Using the angles of these lines with respect to their local frequency axes, the fundamental matrix can be computed. The translation may again be found from the phase exponent of the Fourier transform. This approach has been shown to be robust to noise, and more accurate than some other direct methods. The drawback to this technique is that even less of the image information is kept, first features are extracted and then the Fourier transform of the resulting image is found. Much information is lost in both texture and spatial information. Again, no perspective transforms can be handled.

### Spatial-Frequency Domain Rectification

The inability of the techniques of the frequency domain to rectify perspective images is a large drawback. Since only affine approximations to the correct transformation are available, rotations out of the plane are unsolvable. Using the spatial domain, the scene could have severe perspective distortions but still have a solvable transformation model. It is desirable to simultaneously exploit both the structural information from the spatial domain techniques and the

textural information from the frequency domain techniques. That leads us to once again propose the use of a domain that will encode both the spatial and frequency information at once which we call the spatial-frequency domain.

These methods, although only newly emerging, consolidate some of the better aspects of each of the other two domains. All of the spatial and texture information are encoded into each transformed image, which means that theoretically, everything that can be achieved in either of the two other domains can also be accomplished in this single more concise domain. When rectifying scenes with little geometric structure, it would be desirable to use the featureless rectification methods of the frequency domain, but also include the ability of the spatial domain techniques to remove perspective distortions. To do this, an examination of how the frequency components change across an image would be the ideal route.

This may be accomplished using the short time Fourier transform (STFT), where the Fourier transform of small image patches is computed, yielding localised frequency content information. The changes in frequency content provide indications of both affine and perspective warping, and it is these changes that may be used to obtain pose information. A major drawback associated with the STFT is that a tradeoff has to be made between having poor frequency resolution for high spatial resolution, or poor spatial resolution for high frequency resolution Grossmann and Morlet (1984).

Alternative methods presented themselves with wavelet decompositions. Similar to Fourier transforms, wavelet decompositions are found for image patches, but unlike Fourier transforms, no tradeoff between resolution in the space and frequency domains is required. The theoretical mathematics are still at an early stage, but may be developed and used to solve for homographies in a number of situations. These situations may include image mosaicking and camera calibration problems. However, due to the fact that the more complex and useful wavelet transforms have no analytic expressions, and have only been developed as one-dimensional decompositions using filtering techniques, geometric relations between two views are difficult to derive. In order to correctly rectify an image, two-dimensional wavelets or shapelets are required, from which geometric relations will need to be derived.

Wavelets have previously been used extensively to retrieve shape from texture.

In the earlier methods, affine warping between a frontally imaged reference texture patch and a similar patch in the image is estimated (Witkin, 1981, Garding, 1992). From the affine transformation matrix, information about the rotation and angle of tilt of the patch can be found. This was first done using simple area correlation methods, not wavelets, where the known base patch was iteratively warped until it had a strong correlation to the image patch. These methods progressed to use one dimensional wavelet decompositions of the patches, that could be affinely warped, or alternatively the wavelets themselves could be pre-warped (Clerc and Mallat, 1999). Furthermore, smoothness constraints can be put on the textured surfaces to allow more accurate reconstruction, even when the texture element is estimated from the image (Loh and Hartley, 2005). Similar work in Kovesi (2005) used two-dimensional wavelets, or shapelets, to recover the surface normals for textured surfaces. In that body of work, slant and tilt were found from the gradients of both the image and the shapelet together.

Each of the above mentioned methods solve for the affine planar transformation of one surface patch relative to another on a local scale. Surface orientations are derived from the affine transformation relating two surface patches, but planar perspective homographies are not estimated and global solutions to the problem of pose normalisation are unavailable.

In Heikkilä (2002) with a method known as Multi-Scale Autoconvolution (MSA), which loosely falls into the space-frequency domain, the distribution of feature points within an image are used to determine affine transformations. The input image is treated as a probability density function of feature points, from which three points are selected and linearly combined to define the affine relation to the second image. Transformations are parameterised by $\alpha$ and $\beta$ coefficients, which are scaling factors for two of the three coordinate vectors. When fixed values of $\alpha$ and $\beta$ are chosen, a distribution of the affine space may be calculated from the probability density functions of the base image, spatially scaled by the $\alpha$ and $\beta$ coefficients respectively. which is where the space-frequency tie is. Correlating this distribution with the input image and integrating over the whole of the image yields an expectation value. A range of expectation values for varying values of $\alpha$ and $\beta$, known as MSA coefficients, are calculated for each view. Affine relations between the MSA coefficients for each view can then be extracted (Kannala et al., 2005). Similar to the Fourier domain techniques, this method is only suited to solving for affine transformations and so

will not be used.

The space-frequency techniques are ideally suited to correcting for distortion in scenes where strong features are difficult to extract and match across multiple views. Current methods in the literature are at an early stage of examination and are currently constrained to estimating affine relations between two views. Hence, the potential of the space-frequency techniques to also provide perspective cues are examined and exploited in this thesis.

## 1.4 Contributions

A face pose removal system was developed. Because facial geometric-features are difficult to identify, the pose removal was accomplished in a featureless framework. Symmetry was used as a cue to obtain information about the pose of subjects in this framework. A number of significant developments and improvements to the techniques used in the process were accomplished, improving upon both the efficiency and accuracy of the techniques involved.

Symmetry was examined as a cue for determining pose. An algorithm that estimated the axis of symmetry in images was implemented and showed both improved efficiency and accuracy. The algorithm took just 10% of the time that the algorithm in Prasad and Yegnanarayana (2004) required to execute. The accuracy was much improved, because the majority of unimportant background pixels, which only add to the overall error, were removed in a prefiltering stage developed by the author.

Furthermore, a continuous symmetry metric was developed and shown to be more accurate than existing metrics in the literature. We developed a symmetry measure that is both robust to noise and lighting variation. This is accomplished through the examination of frequency components at each position in the image rather than the image graylevels themselves.

A facial pose removal system was built upon the proposed symmetry-axis estimation and symmetry measurement processes. A number of different pose removal techniques were used to rectify the face images. With each technique, the novel symmetry measure was used as an indicator of whether or not the optimum frontal pose had been obtained. Results demonstrate that the au-

thor's pose removal system outperforms similar techniques in the literature, and are measured against the ground truth. A peer reviewed paper on the topic has been accepted for publication.

Retrieving face images from video sequences is a common task in security surveillance systems, however, the quality of the images will vary greatly due to large pose variations. Obtaining and storing only those images that are suitable for person recognition will greatly improve the efficiency of any face recognition system that is employed, as well as reduce the disk storage space required for acquired face images. The proposed symmetry measurement technique was further applied in determining the optimum fronto-parallel face image from video sequences. Results demonstrate that the most suitable images for face recognition are extracted from the video sequences and the rest of the frames discarded. This both saves on storage space and post capture processing with the recognition systems. Also, the quality of the images can be quantified using the symmetry measure.

An examination of the fundamental theory behind the use of the space-frequency domain for planar image pose recovery was carried out. The proposed featureless pose removal technique was employed to estimate out of plane rotations of textured planar surfaces. Operating in the space-frequency domain, we showed how perspective pose can be removed in a featureless framework, improving upon similar methods in the literature that only account for affine pose estimation. Results demonstrate that pose can be recovered to a reasonable degree of accuracy with the author's method.

## Publications Arising

The first publication is directly associated with the methods and techniques discussed in the thesis. The second publication deals with the implementation of a high performance computing (HPC) cluster to assist in the execution of the large experiments required for the thesis.

- *Removing Pose from Face Images.*
  Seán Begley, John Mallon and Paul F. Whelan.
  Fourth International Symposium on Visual Computing (ISVC08)
  Las Vegas, Nevada, USA, 1st-3rd December 2008.

- *Cost-Effective HPC Clustering For Computer Vision Applications.*
  Julia Dietlmeier, Seán Begley and Paul F. Whelan.
  International Machine Vision & Image Processing Conference 2008 (IMVIP08)
  Portrush, Northern Ireland, 3rd-5th September 2008.

## 1.5 Thesis Outline

Chapter 2 is concerned with symmetry in images. Symmetry is extensively used as a cue in the human visual system and will be used for removing pose in our proposed algorithms. Methods for finding the axis of symmetry in images are presented and evaluated, the best of which is chosen as the technique to be used throughout the rest of the thesis. An improvement to the algorithm is presented that reduces the computational overhead of the algorithm by intelligently filtering image components to be used. Experiments are presented demonstrating improved efficiency and accuracy. The second part of the chapter deals with continuous symmetry measures, determining "how symmetric" an image is. Again, symmetry coefficients in the literature are critiqued and the most suitable symmetry measure is selected. The symmetry measure is developed to improve accuracy and experimental results are presented to demonstrate this improved accuracy.

Chapter 3 deals with face rectification or face-pose removal. Three different mapping techniques are used to remove pose using non-model based techniques. For this, an iterative minimisation process is used which is described. Improvements to the algorithm are achieved through the use of the space-frequency domain and the proposed symmetry measure. Evidence of the improvement is presented. The symmetry measure is further employed to recover the most fronto-parallel face image from video sequences. Results are presented which show a high degree of accuracy.

In order to highlight the larger application area of the rectification technique developed in Chapter 3, Chapter 4 presents planar image rectification. A thorough introduction to the area of planar transformation estimation between two views is presented. A discussion on the more commonly used techniques in both the spatial and frequency domains is given. A novel technique, based on the principals introduced in Chapter 3, is introduced to overcome some of

the limitations of both the frequency and spatial domains. The theoretical reasoning behind the choice of this novel space-frequency domain is presented. Experiments validating the use of the space-frequency domain for planar image rectification are carried out, the results of which are presented.

Chapter 5 presents conclusions and possible further work. A summary of the work carried out to date and an evaluation of its contribution to the literature is presented. Directions of possible future contributions are discussed and their viability assessed. Finally an executive summary is given.

# Chapter 2

# Symmetry in Images

Symmetry forms a very important cue to the human visual system, aiding in object recognition, person identification, and also in object pose determination. Objects are said to be bilaterally symmetric if they are invariant under a reflection about a line, known as the axis of symmetry. This is also known as axial symmetry. Objects may also exhibit rotational symmetry if they are invariant under a rotation about their centroid. Axial symmetry in particular will be exploited as an aid to recover pose in images in this thesis. Two important aspects of symmetry are examined as a pre-requisite to determining pose from images. Firstly, the axis of symmetry in images must be determined. Methods for determining the axis of symmetry in images are examined and evaluated. An algorithm that demonstrated good symmetry axis localisation abilities in complex scenes is chosen to be used in subsequent algorithms in this research, Modifications to the selected algorithm are made to improve both efficiency and accuracy. Secondly, despite the inherent opinion of most people that symmetry is either present or not, continuous symmetry measures may be derived. Different methods exist to quantify the level of symmetry an object portrays. These methods will be assessed to determine the most applicable technique to be used in our application. Again, improvements are suggested and supporting evidence of the improvements are presented.

Finding and quantifying symmetry is not a trivial endeavour. Symmetry can be a very strong defining feature for many objects. As such, in machine vision applications, object detection and recognition may rely on the presence or absence of symmetry. Similarly, in manufacturing processes, symmetry may

be used as a quality control where a high level of symmetry would indicate successful production. These tasks all require that both an axis of symmetry as well as a measurement of the detected symmetry relative to that axis are found.

## 2.1 Finding Axes of Symmetry in Images

Symmetry is measured relative to some axis or point, through which all of the image information is transformed. In order to find how symmetric an object is, the reference axis or centre of symmetry needs to be determined. Axial symmetry is the main focus of this body of work. This section investigates a number of methods for obtaining the location of the axis of symmetry in images. The most appropriate method is selected and improved upon. Experimental results are presented demonstrating the accuracy to which the algorithm performs. The modifications to the selected algorithm also improve the computational efficiency of the overall system.

### 2.1.1 Critique of Selected Algorithms

There exist a number of different methods for estimating the location of the axis of symmetry in images. Some operate directly on detected features in the image, while others operate on the image as a whole. Generally, they can be grouped into two major categories, those that operate on geometric or image features, and those that operate on the graylevels of the image.

In Saint-Marc et al. (1993), it is demonstrated that detected edges, fitted with B-spline curves, may be used to directly estimate three different types of symmetry defined as; skew symmetry, parallel symmetry and smooth local symmetries, all of which are specifications of axial symmetry. It is the skew symmetry that interests us. The methods rely heavily on good edge localisation for the B-spline estimation after which, the parametric equations of the splines are directly used to compute the symmetries. If the edge localisation is inaccurate and subsequently the B-spline fitting is noisy, the symmetry axes detected will be incorrect. There are another two significant drawbacks to utilising this technique for estimating axes of skew symmetry. Firstly, the image

in question needs to have strong, well defined edges, which may not always be available, and secondly, the image must be exactly invariant under a reflection with no outlying edge points.

In Marola (1989), an $n$-dimensional transformation of the detected edges of the image is used to obtain information about the image's axes of symmetry. Each point in an edge detected shape yields a vector direction from the shape's centre of mass. This vector is transformed by multiplying its length and the angle it makes with the x axis by $n$. For perfectly symmetric images, transformed points will overlay their corresponding symmetric pairing. The number and direction of all of the axes of symmetry in the image may be computed in this way. For almost-symmetric images, the detected edges are blurred and each point on the blurred line is transformed. A coefficient of symmetry is then calculated to determine whether symmetry with respect to a chosen axis exists. The symmetry coefficient can be maximised in a global optimisation scheme by moving the axis of symmetry about to obtain the correct axis. Similar to the method in Saint-Marc et al. (1993), unless the images being examined are highly symmetric with little noise, this method of determining axes of symmetry is inaccurate.

A method that operates directly on the grayscale values of the image being examined was proposed in Gofman and Kiryati (1996). A symmetry coefficient with respect to an axis of symmetry is calculated within a region of interest operating directly on the grayscale values. This coefficient of symmetry is used in an iterative optimisation scheme to determine the axis of symmetry. Since the symmetry coefficient is not dependent on the size of the window within which it is measured, the region size that supports the highest symmetry value can also be determined. Although the region based nature of the symmetry coefficient ensures the technique is robust to noise, the method is a very time consuming process and is prone to trapping at incorrect local optima.

In Prasad and Yegnanarayana (2004), a feature based approach is used. An altered image domain is calculated, for every position of which, a matching correspondence is found in the same altered image domain. This forms a symmetric point pair. The altered image domain is obtained using the Gradient Vector Flow (GVF) field of the image Xu and Prince (1998). Each symmetric pair gives rise to a position and angle for the axis of symmetry. A 2D histogram is built up where each correspondence votes for their respective de-

termined axis of symmetry. The angle and position of the axis with the highest support is deemed to be the correct axis of symmetry. The obvious drawback to this method is that the computations involved are very intensive, where even background image locations may vote for the axis. Even with the suggested pruning to the pairs of voting points, a very high number of votes are used. This method is however very robust to noise due to the averaging affect of the GVF field that is calculated, and as such, features from the GVF field are easily matched. By eliminating the background pixels from the voting scheme in a different manner to Prasad and Yegnanarayana (2004), few voting pairs need to be found and the algorithm would be more accurate and efficient. This method was chosen to be the symmetry estimation technique that was most promising for implementation and further investigation for this body of work.

### 2.1.2 Implementation

An algorithm is developed that is a variation of that proposed by Prasad and Yegnanarayana (2004). With Prasad's algorithm, every possible pair of positions in the image are pruned down to likely symmetric pairs that then vote for their axis of symmetry, which is very computationally inefficient. For images that are $100 \times 100$ pixels in size, with Prasad's method there are $10^8$ corresponding pairs to be pruned. The pruning reduces the number of voting pairs to approximately 0.05% of the total possible pairings. In general, for an image of size $m \times n$, the total number of possible pairings is of the order $O(m^2n^2)$ which after pruning yields symmetric pairings numbering $O(m\,n)$. Since the absolute maximum number of valid pairs in the image, including duplicate pairs, is exactly $m \times n$ for a perfectly symmetric image, no saving is achieved.

With Prasad's method, a matching pair of points is found as the positions with the closest possible values from the curvature of the GVF field. A match for every point in the image is found in this way. These pairs are then filtered in a post processing stage using the curl, divergence and magnitude of the GVF field. Pairs that are too dissimilar in one of the other three measurements are removed from the voting table. To summarise, initially all correspondences for each point in the image are found, and then the points that are too dissimilar are filtered out.

To increase the efficiency of the algorithm, a constraint on the curvature of the GVF field will be used to eliminate a large proportion of the image pairings in a pre-processing stage. Because background pixels only add to the overall error in the process, only features with a high curvature value, indicating that they have a strong relevance to the image structure, are deemed to be foreground pixels and are used. This subset of points, determined from regions of high curvature in the GVF field, are deemed appropriate with which to vote. This reduces the number of point correspondences that have to be found as well as reducing the search space. This both increases the speed and efficiency of the algorithm as well as improve the axis localisation accuracy. Approximately 10% of the maximum number of valid pairings are used giving $0.1 \times m \times n$ total voting pairs. We will call this proposed algorithm the Constrained Curvature Symmetry-Axis Estimation (CCSAE).

Using a binary tree search algorithm, a standard search algorithm, for both Prasad's method and the CCSAE method provides the most efficient searching mechanism. The total number of comparisons required for finding a single correspondence is $\mathcal{C}\{\log_2(N)\}$ which is the depth of the binary tree, where $N$ is the total number of points ($N = m \times n$ in our case) and the $\mathcal{C}$ (ceiling) operator indicates that the number is rounded up to the next whole number. In total, for Prasad's algorithm the total number of comparisons initially required is $N \times \mathcal{C}\{\log_2(N)\}$. With pre-filtering the curvature of the GVF field, the total number of comparisons required is $0.1 \times N \times \mathcal{C}\{\log_2(0.1 \times N)\}$. This gives a very significant time saving of

$$T_{saving} = \left[0.9 \times N \times \mathcal{C}\{\log_2(N)\} + 0.1 \times N \times \mathcal{C}\{\log_2(10)\}\right]\tau \qquad (2.1)$$

where $\tau$ is the time taken for a single comparison between two numbers. This is verified in the experimentation section where images of various sizes are operated on and the execution time is measured. Of course other time differences need to be considered too. The computation of the binary tree takes longer for a larger input set. Therefore, as is the case with the CCSAE method, using a smaller data set will require less time for the computation of the binary tree.

The algorithm is implemented using the following procedure, a flow-chart representation of which is shown in Fig. 2.1.

First, the Gradient Vector Flow (GVF) field of the edges in the image need to be determined. Strong edges in the input image, $I(x, y)$ are found using a

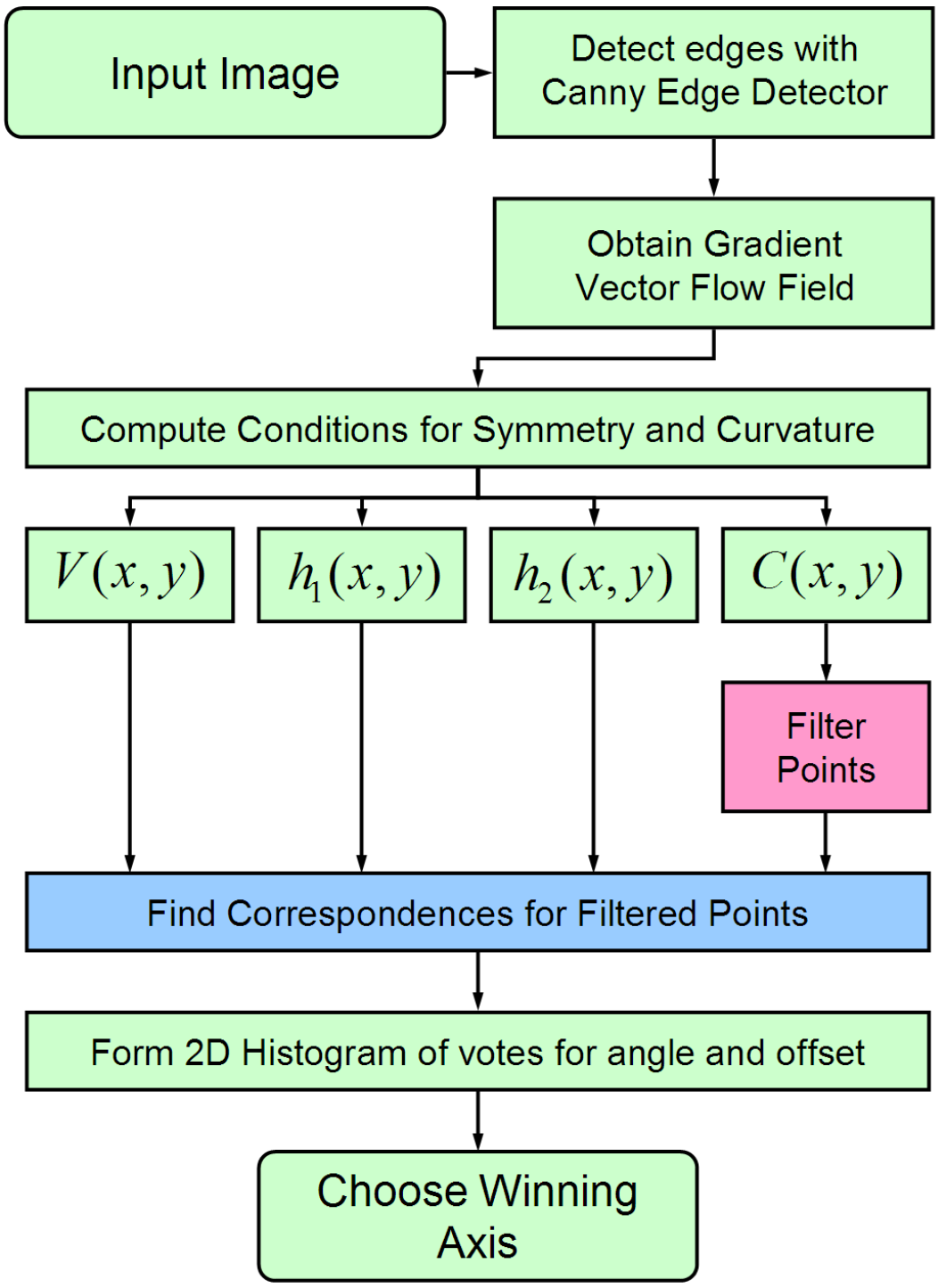



Fig. 2.1: Symmetry Axis algorithm. The standard approach processes are displayed as green blocks. The addition point filtering process is highlighted as a pink block, and the modified point correspondence finding process is shown as a blue block.

Canny edge detector, the output from which is stored in the edge map $f(x, y)$. Operating on the edge map detected from the input image, the GVF field is determined which yields vectors at each position in the image $\mathbf{v}(x, y)$ whose components are stored in $u(x, y)$ and $v(x, y)$. The GVF field is obtained by minimising an energy functional determined from the edge image data. The functional to be minimised is:

$$\varepsilon = \int\int [\mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |\mathbf{v} - \nabla f|^2] dx\, dy \tag{2.2}$$

where $u_x$, $u_y$, $v_x$, and $v_y$ are the partial derivatives of the vector components $u$ and $v$ with respect to $x$ and $y$. The $\mu$ parameter controls the relative effects of the first and second terms in the energy functional and should be selected according to the noise present in the image. The higher the noise level, the higher the value $\mu$ should be. The energy functional is minimised using the algorithm developed in Xu and Prince (1998). The technique iteratively determines the GVF field by estimating the gradient of the image on each iteration and updating the image with the new gradient field. Using the calculus of variations, it can be shown that for $\varepsilon$ to be a minimum, the following two criteria must be satisfied.

$$\mu \nabla^2 u - (u - f_x)|\nabla f|^2 = 0 \tag{2.3}$$

$$\mu \nabla^2 v - (v - f_y)|\nabla f|^2 = 0 \tag{2.4}$$

On subsequent iterations, the GVF field of the image is updated using Equations 2.3 and 2.4 using the iteration scheme given in Xu and Prince (1998). The number of iterations and the step size used in computing the GVF were heuristically chosen to find the optimal balance between the speed of the algorithm and the smoothness of the GVF field.

Three vector measurements on the GVF field are calculated in order to find symmetry between points. The squared magnitude of each vector at every position in the image is calculated $Mag(x, y)$, the divergence of the vector field at each position is found $Div(x, y)$, and the curl of the vector field at each position is also found $Cur(x, y)$. The expressions for determining these from the GVF field are given in equations 2.5, 2.6, and 2.7.

$$Mag(x, y) = |\mathbf{v}(x, y)|^2 = u(x, y)^2 + v(x, y)^2 \tag{2.5}$$

$$Div(x, y) = \nabla \cdot \mathbf{v}(x, y) = u_x(x, y) + v_y(x, y) \tag{2.6}$$

$$Cur(x, y) = \nabla \times \mathbf{v}(x, y) = [v_x(x, y) - u_y(x, y)]\,\overrightarrow{\mathbf{k}} \tag{2.7}$$

These three vector measurements are used to obtain three necessary conditions on the field to demonstrate symmetry between the points $(p_i, q_i)$ and $(p_j, q_j)$. It should be noted that these conditions don't guarantee symmetry, but are necessary if symmetry is present and so form a good indicator when matching points Prasad and Yegnanarayana (2004). The three conditions are:

$$\begin{aligned} C_1: \quad Mag(p_i, q_i) &= \quad Mag(p_j, q_j) \\ C_2: \quad Div(p_i, q_i) &= \quad Div(p_j, q_j) \\ C_3: \quad Cur(p_i, q_i) &= \quad -Cur(p_j, q_j) \end{aligned}$$

Another local feature is also used for matching corresponding symmetric points and is used in the removal of features of little importance. The curvature of the GVF field is found using equation 2.8, and a threshold on the curvature is used to segment out the foreground points of interest. The threshold is determined from the content of the image so that 10% of the points in the image are used for calculating the axis of symmetry.

$$C(x, y) = \frac{1}{|\mathbf{v}|^3}[(v_x + u_y)uv - u_x v^2 - v_y u^2] \tag{2.8}$$

Examples of the outputs from each of the processes in the algorithm are shown in Fig. 2.2. Once the subset of feature points $C_{sub}(x, y)$ are determined from the curvature map, their corresponding points need to be determined. For each detected point $C_{sub}(x_i, y_i)$ in the thresholded curvature image, the corresponding closest curvature value is found at location $C_{sub}(x_j, y_j)$ and the locations of the corresponding pair are stored in a matched point table $P(p_m\{x_i, y_i, x_j, y_j\})$. Each corresponding pair is then assessed using the three conditions that are required for symmetry to be present. Since real images don't exhibit exact symmetry, a threshold on the allowable minor differences between the condition values is selected, e.g.

$$Mag(p_i, q_i) \quad - \quad Mag(p_j, q_j) < \epsilon_1 \tag{2.9}$$

$$Div(p_i, q_i) \quad - \quad Div(p_j, q_j) < \epsilon_2 \tag{2.10}$$

$$Cur(p_i, q_i) \quad + \quad Cur(p_j, q_j) < \epsilon_3 \tag{2.11}$$

$$\tag{2.12}$$

where each $\epsilon$ is heuristically selected for each of the three conditions. If any of the three condition values for the matched pair are too far from ideal, the points are considered non-symmetric and their coordinates are removed from the matched pair table.

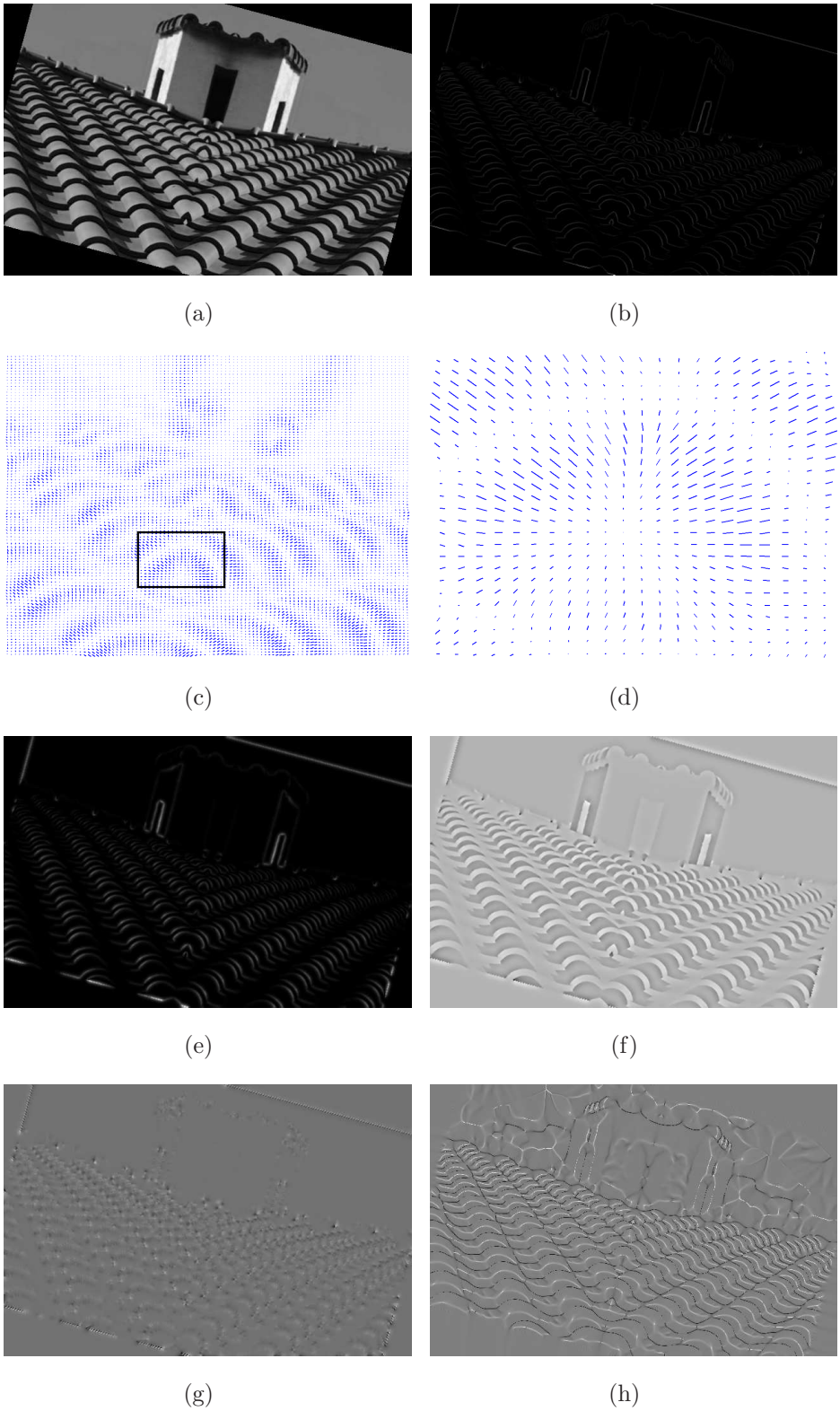


Fig. 2.2: Images from each step in the process. (a) Input image. (b) Edge map. (c) GVF field. (d) Close up of highlighted region of GVF field. (e) Magnitude of GVF. (e) Divergence of GVF. (e) Curl of GVF. (h) Curvature of GVF.

Once all of the detected feature points have been evaluated, the angles and $y$-intercepts of the axes of symmetry through which each point pair are symmetrically related are found relative to the $x, y$ axes of the image. A 2D histogram of the angles and intercepts of the axis of symmetry is formed and the axis with the highest support is found as the position in the histogram that has the most number of points voting for it. This is deemed to be the axis of symmetry. In the event that more than one axis is found to have the same support in the histogram, the continuous symmetry metric, described in the next section, for each axis is calculated and the axis yielding the highest continuous symmetry value is chosen to be the correct axis.

### 2.1.3 Experiments

In order to evaluate the algorithm and determine the accuracy to which it estimates axes of symmetry in images, a number of different experiments are carried out. Also, the other significant advantage of using the CCSAE method is that the algorithm is executed in a much shorter time. This is demonstrated in the first experiment.

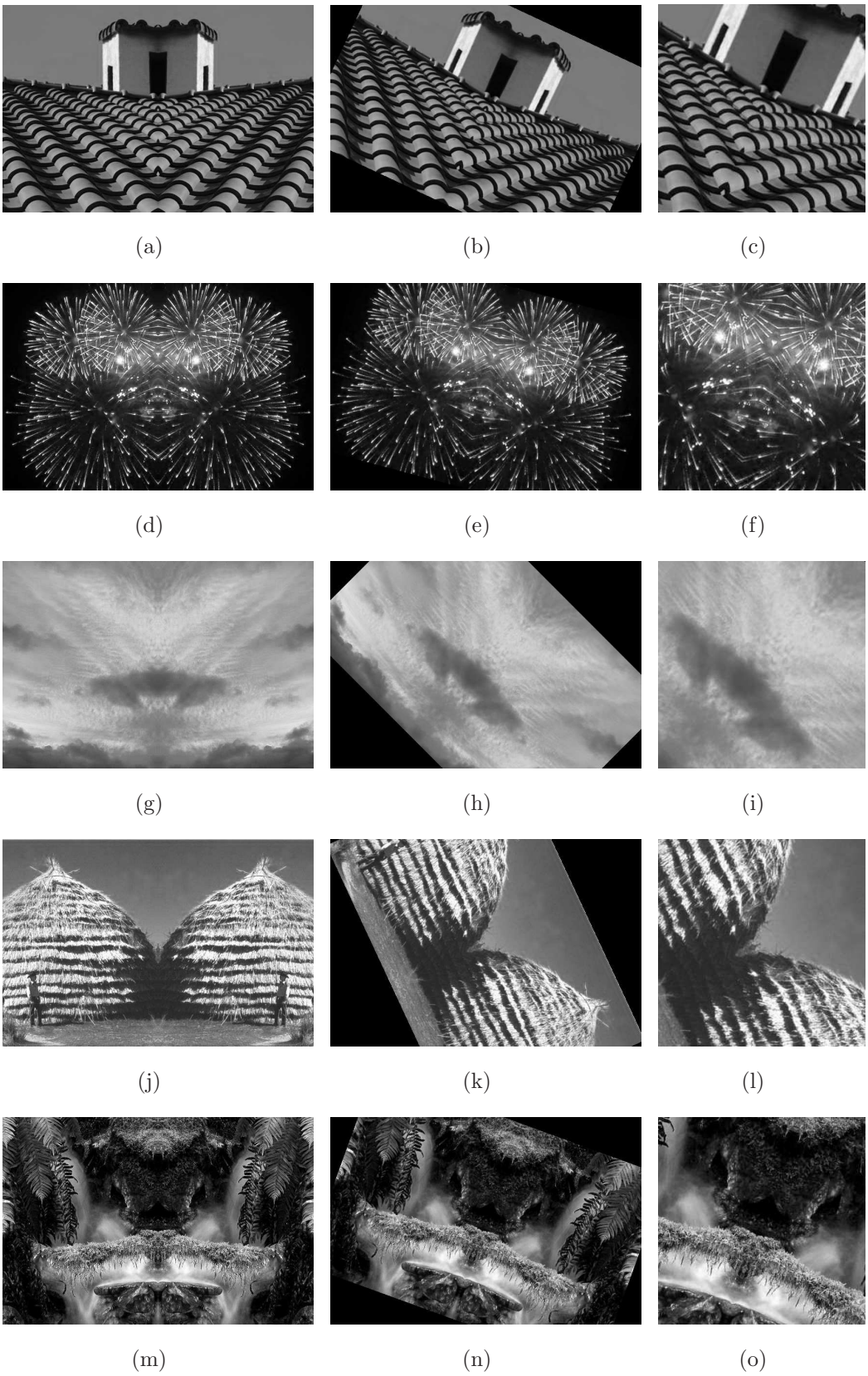
Each of the experiments in this section are carried out on a series of synthetically created symmetric images. Depending on the experiment being run, noise may be added. However, in each experiment, the original image is transformed by a planar transformation and each image is then cropped to remove any zero padding caused by the transformation. A sample of the images used, with example transformations is shown in Fig. 2.3.

#### Experiment 1 - Computational Efficiency

As was stated earlier, pre-filtering the image's associated curvature of the GVF field yields significant time savings. The estimated time saving in determining point matches after the GVF field is computed is estimated by the author to be

$$T_{saving} = \left[0.9 \times N \times \mathcal{C}\{\log_2(N)\} + 0.1 \times N \times \mathcal{C}\{\log_2(10)\}\right]\tau \qquad (2.13)$$

Some amount of time is required to compute the GVF field with each of the two methods, but in each case the parameters for computing the field are



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Fig. 2.3: A sample of the symmetric images. The first column contains the symmetric images. The second column contains the transformed images. The third column contains the cropped images.

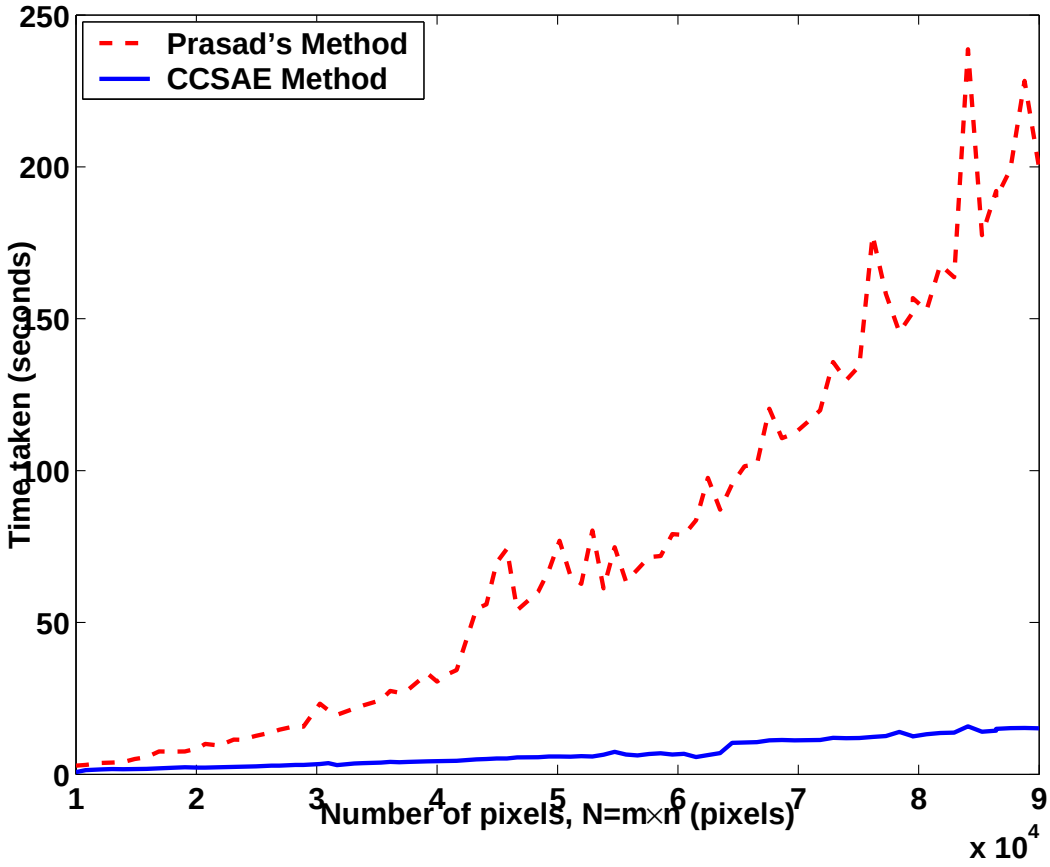



Fig. 2.4: The amount of time taken to compute the axis of symmetry histogram.

identical. The only difference in time will be due to the computation of point matches and their associated binary trees. The times required for estimating the axis of symmetry in images of increasing size was measured for each of the two methods. The results are shown in Fig. 2.4.

From these results it can be seen that the suggested modifications to the algorithm provide a significant efficiency improvement. Part of the improvement is due to the much smaller number of comparisons that are required to match each detected point in the image to its symmetric pair. The remainder of the difference in computation times can be attributed to the larger time required to compute the initial binary tree. Since the computation time of the binary tree has a exponential relationship to the depth of the tree, and since the depth of the comparison tree is $\mathcal{C}\{log_2(10)\}$ times the depth of the CCSAE binary tree, the CCSAE binary tree will be computed in significantly less time.

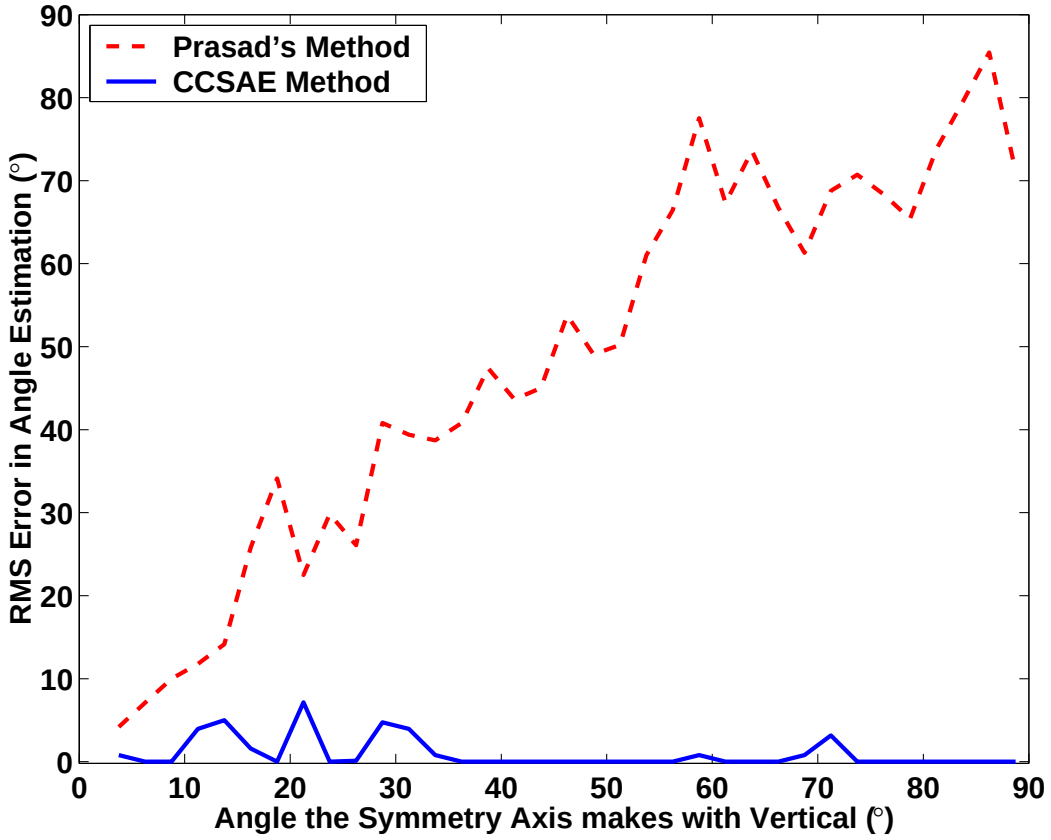



Fig. 2.5: RMS error in estimating the angle of the axis of symmetry.

**Experiment 2 - Perfect Symmetry**

The second experiment is conducted to determine the accuracy to which the algorithm can detect axes of symmetry within images portraying perfect symmetry, where the axis of symmetry runs through the centre of the image at various angles. A collection of 10 perfectly symmetric images is used. Each image is rotated in-plane about its centre point and then the image is cropped to remove any assisting bias that the occluded and included image edges would give. The algorithm is run on each image using a one-dimensional histogram to determine the angle since no offset is present. The process is repeated on each of the 10 images for a range of rotation angles. The results are gathered and the root mean squared error for each angle of rotation across the 10 images is calculated and are shown in Fig. 2.5. The experimental results are compared to the filtered method of Prasad and Yegnanarayana (2004) and are shown to be more accurate while requiring less computation.

The results clearly indicate that the CCSAE algorithm provides a large increase

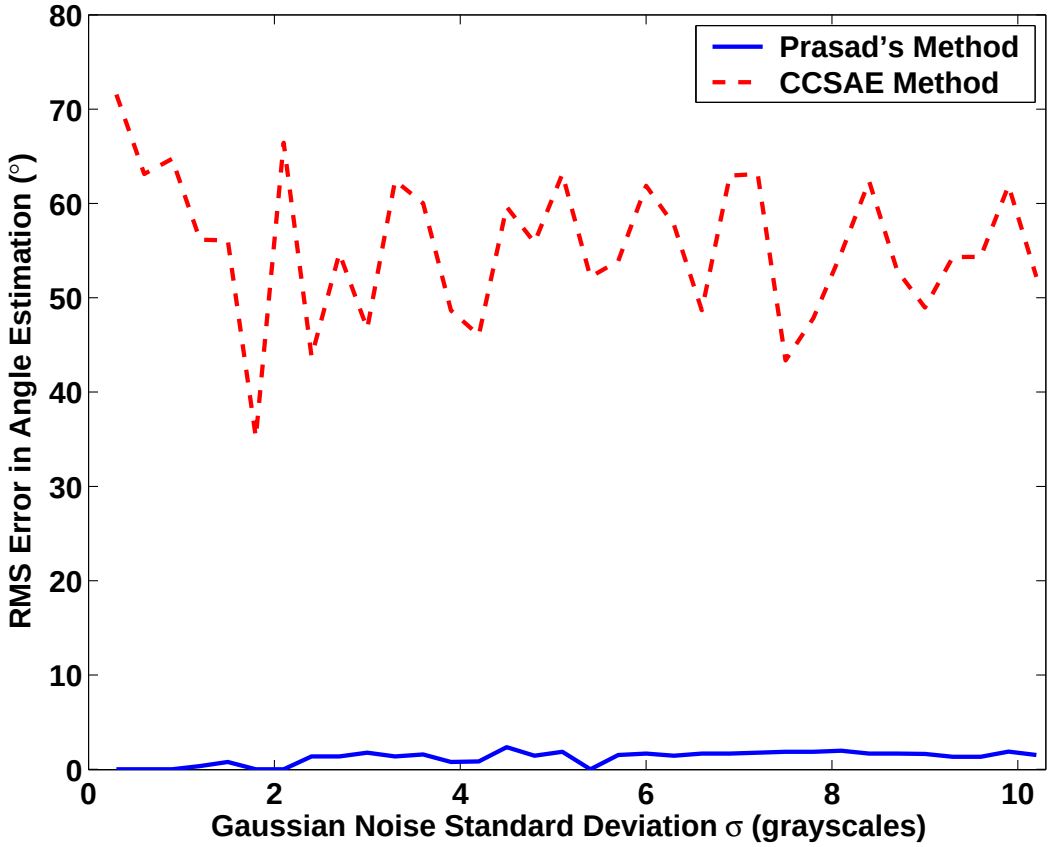



Fig. 2.6: RMS error in estimating the angle of the axis of symmetry.

in accuracy over Prasad's method, although part of the reduced accuracy of the un-altered method may be due to the more complex scenes that are currently used compared to those that were presented in Prasad and Yegnanarayana (2004).

### Experiment 3 - Robustness to Noise

The third experiment demonstrates the noise robustness of the GVF field methods. Gaussian noise of varying levels is added to the images prior to edge map computation. The images are no longer perfectly symmetric in the grayscale sense. This experiment highlights that the CCSAE algorithm accurately estimates the axis of symmetry for in-plane, symmetric objects even in the presence of high noise levels.

As it can be seen from the graph in Fig. 2.6, even with high levels of added gaussian noise, the CCSAE algorithm performs very well. With Prasad's Method, although there is some variation in how the algorithm performs at the different

noise levels, the errors appear to be independent of the level of noise. This indicates that using the GVF field does have a noise filtering effect.

**Experiment 4 - Finding off-centre Symmetry**

The final experiment measures the accuracy of the axis determination algorithm with axes of symmetry that don't run through the image centre. Once again, images that are perfectly symmetric are used, and no noise is added. A 2D histogram is required in order to find the correct axis of symmetry in this case.

For each part of the experiment, each image is transformed by a rotation angle and one of 6 pre-selected translations. Then the algorithm is run on a cropped version of the image (again, the images are cropped to remove any zero padding that results from the rotation and translation). Every matching pair gives rise to an angle and offset value for an axis of symmetry. The 2D histogram of these values is computed and the angle-offset combination that yields the highest vote is deemed to be the axis of symmetry.

The results are shown in Figures 2.7 and 2.8. Fig. 2.7 shows the mean estimated angle error at every angle of rotation of the image for each of the six translations, and Fig. 2.8 shows the mean error in the estimated $y$-intercept of the axis of symmetry. From these results it can be seen that very small angle estimation errors are present, and in fact are even smaller than with the 1D histogram method of Experiment 1. The $y$-intercept values are typically 25 pixels, which is again a very small error, considering that the majority of the correct intercept values are at least an order of magnitude larger than the errors.

The results demonstrate that a very accurate estimation of the orientation and location of the axis of symmetry is accomplished through using the proposed CCSAE technique. This indicates that it is more efficient and accurate to remove points conveying weak structural information prior to finding the corresponding symmetric point pairs.

This technique will aid in removing pose from face images. The face images' axis of symmetry will give information about the in-plane rotation of the subjects, or the tilt of their heads. This tilt can be removed prior to removal of

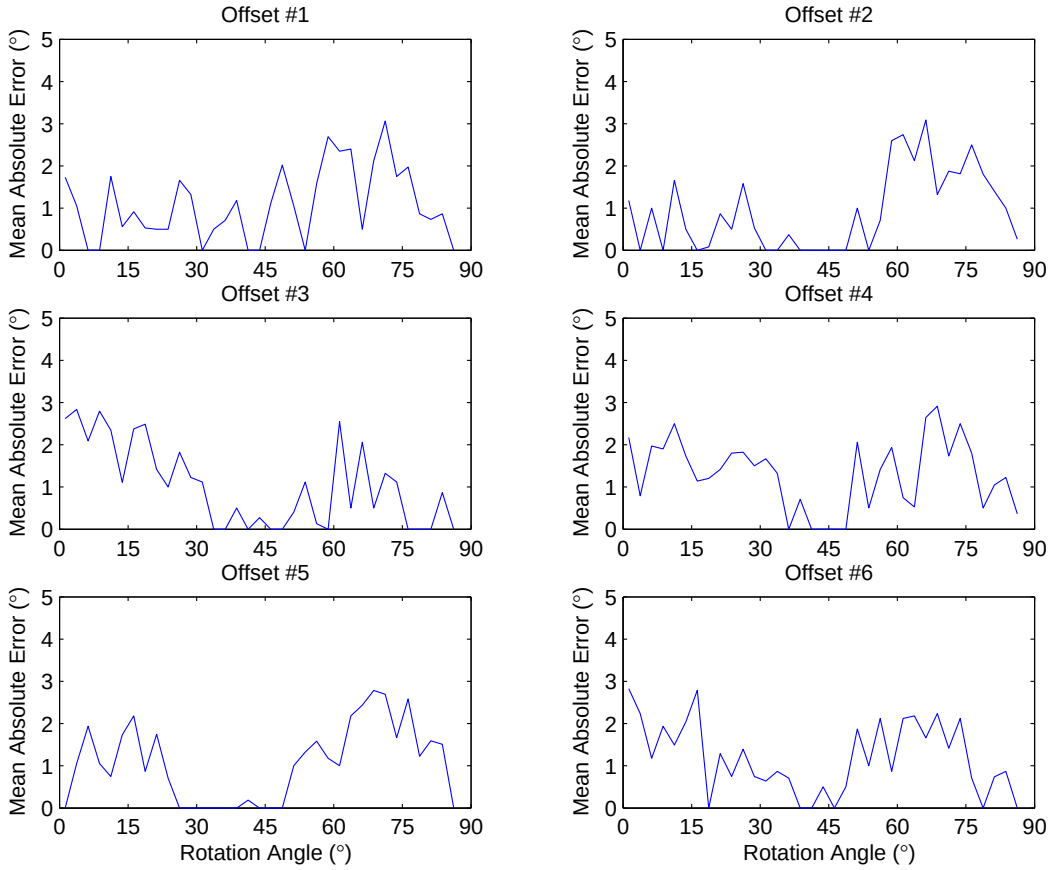



Fig. 2.7: Mean absolute error in estimating the angle of the axis of symmetry.

out of plane rotation of the subjects' heads.

## 2.2 Continuous Symmetry Measures

Symmetric and almost symmetric images can easily be identified by a human observer. Computers on the other hand, require a strong description of what symmetry is. Although symmetry is typically considered as a binary metric, there are levels of "how symmetric" an object appears to be. This section deals with a number of different representations of how symmetry can be quantified. The techniques involved are critiqued and their niche application domains identified. The most suitable approach to quantifying levels of symmetry for our task is identified and suitable modifications are made to it. Detailed experimentation supporting the proposed modifications are given.

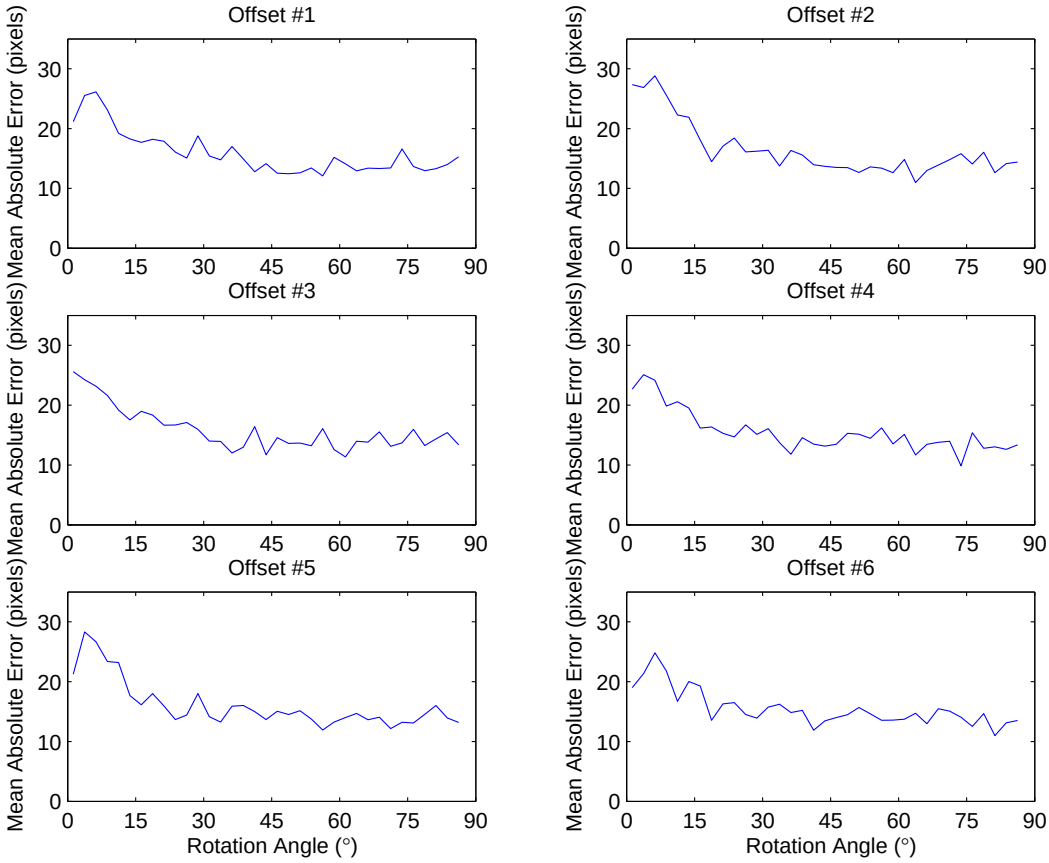



Fig. 2.8: Mean absolute error in estimating the y-intercept of the axis of symmetry.

### 2.2.1 Critique of Selected Algorithms

A measure of "how symmetric" an object appeared to be was first proposed in Marola (1989). In it, Marola used a symmetry coefficient to determine the existence of symmetry. The symmetry coefficient was measured directly on the intensity of edge detected images. The coefficient was used to determine the axis of symmetry in the almost symmetric images where the coefficient was evaluated for a number of different values of the parameters defining the axis, $a$ and $b$. The combination of parameters that yielded the highest symmetry coefficient defined the axis of symmetry for that set of detected edges. Marola didn't fully develop the potential of that symmetry coefficient to also describe the level of continuous symmetry that is present in all images, and instead used it to refine the initial estimate for the position of the axis of symmetry obtained with other methods.

The notion of "continuous symmetry" was first realised in Zabrodsky et al.

(1995). In that publication a symmetry distance is used to quantify how close to perfect symmetry an object appeared to be. The minimum mean squared distance that all of the detected feature-points within a shape needed to be moved to ensure symmetry was used as the continuous symmetry measure. However, that symmetry measure, initially intended for measuring the symmetry of molecules, is only really suited to strong geometric shapes.

A more suitable symmetry measure is described in Gofman and Kiryati (1996) where a ratio of the symmetric and antisymmetric components of a function may be used to obtain a continuous symmetry measure. Elementary algebra tells us that functions are even if $f(x) = f(-x)$ and are odd if $f(x) = -f(-x)$, which are directly equivalent to symmetric and antisymmetric representations respectively. It is the proportion of the overall signal that is symmetric that determines the measure of symmetry. The ratio of symmetric to antisymmetric components was exploited as their continuous symmetry measure. One drawback to using the algorithm is that light gradients have an effect on the symmetry coefficient. However, since the symmetry is measured for large regions of the object image, it is very robust to noise. Because the algorithms employed in this thesis aim to avoid the use of geometric-feature detection and localisation, this symmetry measure is deemed to be the most suitable.

### 2.2.2 Implementation

Although the algorithm in Gofman and Kiryati (1996) is very robust to noise, it is susceptible to errors due to lighting gradients present in the images being examined. Also, the continuous symmetry measure does not vary smoothly as the object being viewed deviates from perfect symmetry, with the error function showing discontinuities. This becomes problematic when the coefficient of symmetry is being used in iterative minimisation processes to restore symmetry and hence pose. For these reasons, some minor modifications to the algorithm are proposed which are outlined further below.

The algorithm is implemented using the following procedure. For ease of notation, the axis of symmetry is assumed to run through the image centre in a vertical direction, but the symmetry coefficient can be measured with respect to any axis of symmetry.

Firstly, each image is decomposed into its symmetric and antisymmetric components about the $y$ axis as follows:

$$I_{sym}(x,y) = (I(x,y) + I(-x,y))/2 \tag{2.14}$$

$$I_{asym}(x,y) = (I(x,y) - I(-x,y))/2 \tag{2.15}$$

In Gofman and Kiryati (1996) the continuous symmetry measure about the vertical $y$ axis is calculated as:

$$S\{I(x,y)\} = \frac{\|I_{sym}\|^2}{\|I_{sym}\|^2 + \|I_{asym}\|^2} \tag{2.16}$$

This symmetry measure will achieve its maximum value of 1 when the image in question is perfectly symmetric and will achieve its minimum value of 0 when the image is perfectly asymmetric about the vertical axis. Due to the nature of the optimisation algorithm that we employ, the level of asymmetry in the image will be minimised in the iterative optimisation scheme which equates to maximising the symmetry. The function for calculating the antisymmetric coefficient is similar to Eqn. 2.16, but uses the antisymmetric component of the image in the numerator.

$$AS\{I(x,y)\} = \frac{\|I_{asym}\|^2}{\|I_{sym}\|^2 + \|I_{asym}\|^2} \tag{2.17}$$

To achieve optimum pose normalisation performance, the antisymmetry coefficient being used should form a smooth continuous curve with no discontinuities and with as few, and preferably zero, local minima as the image deviates further from symmetry. This should also be true in the case where illumination variations are present. Because the symmetry coefficient of Gofman and Kiryati (1996) is calculated directly on the grayscale pixel values of the image, which we call dense-matching, the symmetry coefficient is prone to local minima trapping. One approach to combat this drawback is to use Gaussian blurred images in the cost function. The Gaussian blurring of images removes the high frequency noise content of the images, making them robust to noise. But as with the symmetry coefficients measured directly on the grayscale values of the images these Gaussian blurred images can not overcome lighting variations. An alternative approach is required.

One such approach would be to examine the frequency response at every position in the image. This allows higher frequency noise components to be discarded from the estimation and also removes the bias of lighting imbalances

as it is the frequency content not the grayscale values that are being examined. These criteria can be achieved with the use of a wavelet transformation of the image space which gives localised frequency information.

The wavelet transformation of a one-dimensional function is given as:

$$W(a,b) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{|a|}} \psi^* \left( \frac{x-b}{a} \right) dx \tag{2.18}$$

where $W(a,b)$ is the wavelet decomposition of the function $f(x)$ at position $b$ using the mother wavelet $\psi$ scaled in width by parameter $a$, and the $*$ operator indicates the complex conjugate of the mother wavelet function. An infinite range of values for $a$ exists, but in practice only a very small set of values are chosen. A similar expression may be used to describe the wavelet decomposition of an image using a two dimensional wavelet decomposition in which case 3 parameters are used, one for scale and two for position. It is possible to have two different scales for the mother wavelet in the $x$ and $y$ directions separately for two dimensional decomposition, but in practice this is not done.

Here our proposed symmetry coefficient is defined. The symmetry coefficient is measured on the $2D$ wavelet decomposition of the image in question, the equation for which is given in Eqn. 2.19 where $W_{sym}(s,x,y)$ and $W_{sym}(s,x,y)$ are the symmetric and antisymmetric components of the wavelet transformed image respectively at position $(x,y)$ and using a wavelet scaled by $s$. The coefficient will be called the Wavelet-Based Symmetry Coefficient (WBSC).

$$S\{W(s,x,y)\} = \frac{\|W_{sym}(s,x,y)\|^2}{\|W_{sym}(s,x,y)\|^2 + \|W_{asym}(s,x,y)\|^2} \tag{2.19}$$

### 2.2.3 Experiments

To demonstrate the advantages of using the wavelet transformed image domain over directly using the raw grayscale values or gaussian filtered values, three experiments are presented. With each experiment, three symmetry coefficients are calculated for each orientation of the texture-mapped surface:

(i) a coefficient calculated directly on the grayscale values

(ii) a coefficient calculated on the gaussian blurred grayscale values

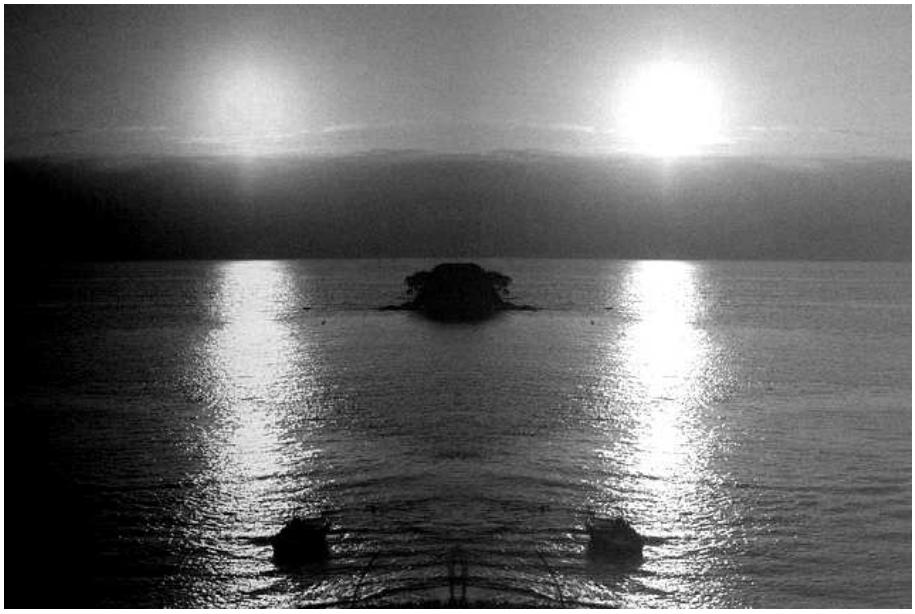


Fig. 2.9: Sample fronto-parallel symmetric image from the database with Gaussian noise and lighting imbalance added.

(iii) and the WBSC calculated on the wavelet transformed image values

Each of the three surfaces that the image is mapped to demonstrate the three approximations to the shape of a human face that will be used in the subsequent chapter, namely a planar surface, a cylinder and an ellipsoid.

**Experiment 1 - Image Mapped to a Planar Surface**

One of the fronto parallel symmetric images from the database of images was selected, to which, a lighting imbalance and gaussian noise, with a standard deviation of 5 grayscales, was added. This experiment will highlight the robustness to noise and lighting variation of the proposed technique. The image used is depicted in Fig. 2.9. The altered image was then texture-mapped to a generated plane. This plane was synthetically rotated about its vertical axis using a planar transformation matrix. The asymmetry was measured in each position using the WBSC cost function with 10 wavelet scales, the direct cost function, and the gaussian-filtered image cost function. Fig. 2.10 shows the results of the experiment which are scaled to have the same mean for display purposes.

From the graph in Fig. 2.10 it can be seen that the cost function based directly

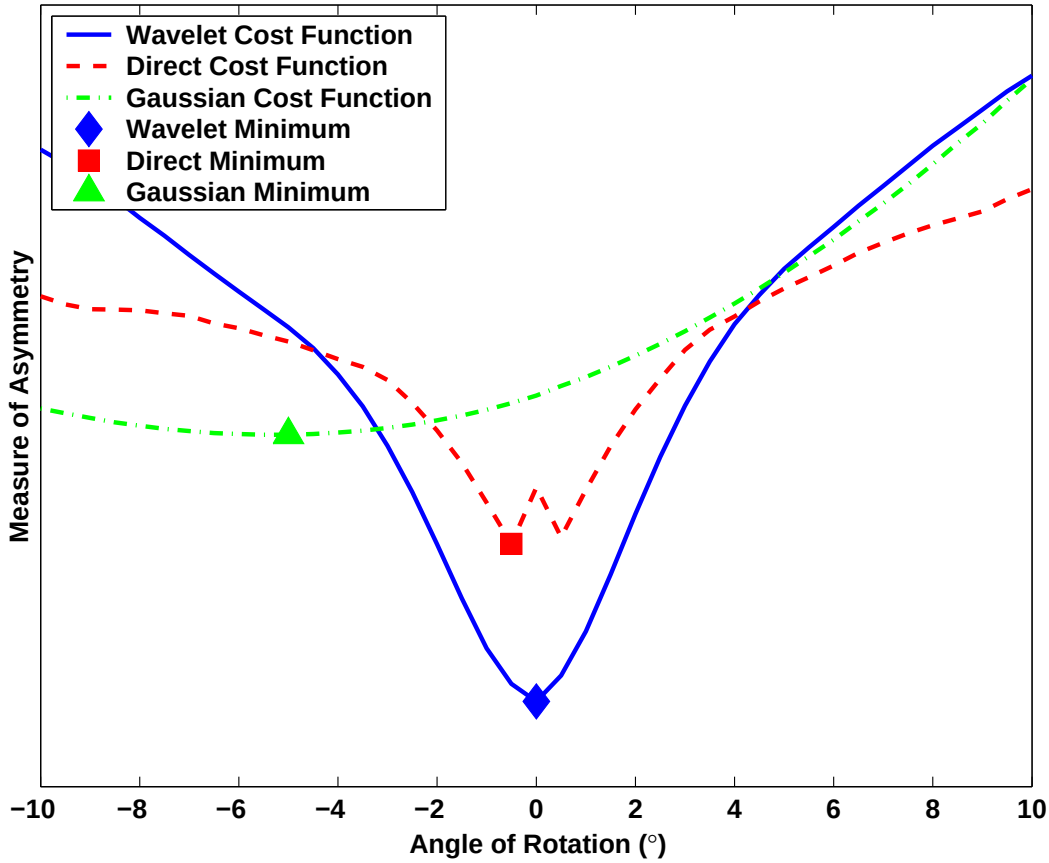



Fig. 2.10: Symmetry-cost measured on the three different input spaces, described in Sec. 2.2.3, as the textured plane is rotated about its yaw axis. The costs are scaled for display purposes.

on the grayscale values exhibits two local minima, neither of which is correct. The gaussian filtered image achieved its minimum at -5 degrees while the wavelet transformed data cost was the only method to achieve its minimum at the correct location. The WBSC cost function manifold is also both smooth and without local minima. This indicates that the wavelet transformed image cost function could be utilised in iterative minimisation processes and it is robust to lighting variations and noise.

### Experiment 2 - Image Mapped to a Cylinder

The same fronto-parallel symmetric image from Experiment 1 is again selected. This image is then texture-mapped to one half of the curved surface of a cylinder, with the other half being left blank. The cylinder is rotated about its vertical axis and the projection of the texture mapped cylinder onto the image

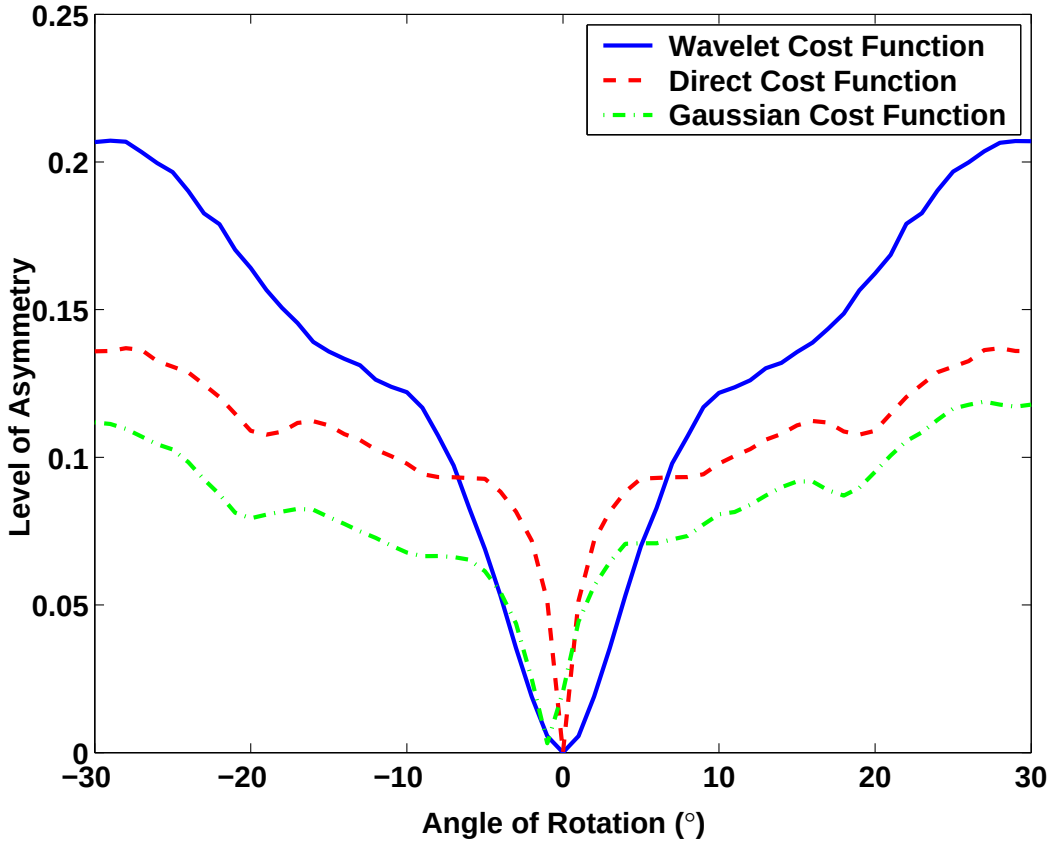



Fig. 2.11: Symmetry-cost measured on three different input spaces as a texture mapped cylinder is rotated about its yaw axis and projected onto the image plane.

plane is taken. The asymmetry coefficient of the projection of the cylinder onto the image plane is measured in each position using the proposed WBSC cost function, the direct cost function, and the gaussian-filtered image cost function. Fig. 2.11 shows the results of the experiment. Again, the number of wavelet sizes chosen was 10.

From the graph in Fig. 2.11 it can be seen that the cost function based directly on the grayscale values exhibits a number of minima, only one of which is correct. This indicates that the symmetry cost function calculated directly on the grayscale values is prone to local minima trapping. The gaussian filtered image achieved its minimum cost at -2 degrees, an incorrect value, and also displays a number of local minima and so it suffers from the same problems that exist with the directly measured cost function. The WBSC cost function is the only method to achieve a single minimum which is also at the correct location. The cost function is smooth and without incorrect local minima,

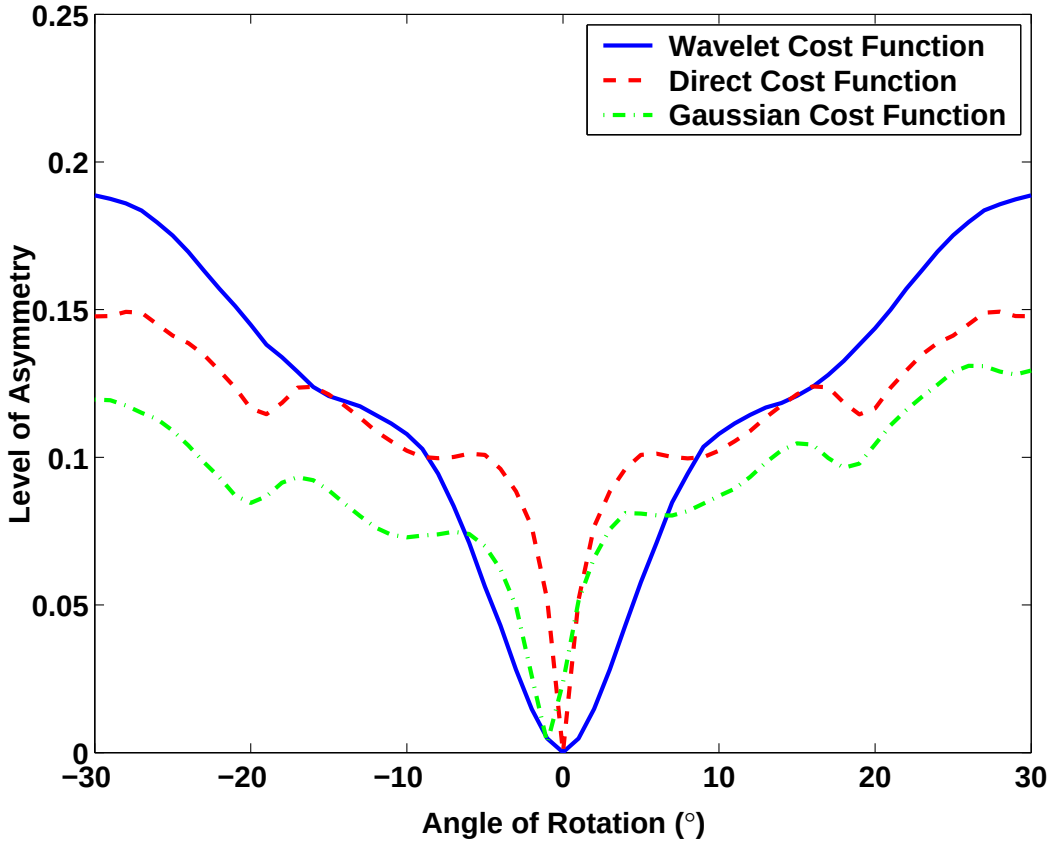



Fig. 2.12: Symmetry-cost measured on three different input spaces as a texture mapped ellipsoid is rotated about its yaw axis and projected onto the image plane.

demonstrating its superiority over the other two methods.

### Experiment 3 - Image Mapped to an Ellipsoid

Once again, the fronto parallel symmetric image from Experiment 1 is used. The image is texture-mapped to one half of an ellipsoid with the other half being left blank. The texture-mapped ellipsoid is rotated about its vertical axis and a projection onto an image plane is created. The asymmetry of the imaged ellipsoid in each position is measured using the WBSC cost function with 10 scales of wavelet, with the direct cost function, and with the gaussian-filtered image cost function. Fig. 2.12 shows the results of the experiment.

Similar to the results of Experiment 2, it can be seen from the graph in Fig. 2.12 that the WBSC cost function is the most suitable choice for use in an iterative minimisation process. The cost function based directly on the grayscale values

exhibits a number of local minima, only one of which is correct. Because of the proximity of other local minima, this correct solution will only be obtained in an iterative minimisation scheme if an initial estimate very close to the correct solution is provided. The gaussian filtered image achieved its minimum cost at the wrong angle as well as having a number of incorrect local minima. Optimisation processes based on the cost functions measured directly on the grayscale values and the gaussian blurred grayscale values are susceptible to trapping at local minima. The wavelet transformed cost function on the other hand is the only method to display no incorrect local minima and achieve a smooth cost manifold. The WBSC cost function is thus very suitable for use in iterative optimisation processes.

## 2.3 Conclusions

Two aspects of symmetry in images were examined: (i) finding axes of symmetry in digital images was discussed, with a critique of current algorithms also being carried out, and (ii), measuring the level of symmetry present in images on a continuous rather than on a binary scale was examined.

The algorithm in Prasad and Yegnanarayana (2004) demonstrated good symmetry axis localisation and was selected. The algorithm used the gradient vector flow field of the detected edges in an image to find corresponding symmetric point pairs. Each of these point pairs then voted for an axis of symmetry. Modifications were made to the algorithm to improve both its efficiency and accuracy. Namely a restriction on the allowable curvature of the GVF field that points must achieve to vote for their respective axis of symmetry. This significantly reduced the search space within which points are matched yielding a vast improvement in efficiency. Removing voting background pixels also reduced sources of error, resulting in improved accuracy. In the majority of situations, the removal of corresponding pairs in areas of low GVF curvature removes very little structural information. However, in situations where the entire image can be considered as being part of the foreground of the captured scene, removing 90% of the image pixels based on the curvature threshold would result in the removal of structural information. This may cause less accurate axis localisation than would be possible using all of the corresponding point pairs. The difference in accuracy would be small, since the corresponding

pairs with the highest curvature are kept with the filtering technique, resulting in the majority of the structural information being retained.

Finding the axis of symmetry in an image allows for the extraction of further measurements relating to the symmetry. The level of symmetry present in the image is one of these measures. An examination on describing levels of symmetry in images was carried out. Current methods in the literature were critiqued and discussed, with strengths and weaknesses of each algorithm outlined. A symmetry coefficient that suited the application's needs was selected. Modifications to the calculation of the coefficient were made to account for uneven lighting conditions as well as noise, two situations that regularly occur when capturing face images. This came about through a wavelet transformation of the image domain, yielding localised frequency information, which allows for the featureless removal of pose from face images even with lighting variation and noise. The result is a wavelet-based symmetry coefficient (WBSC).

Results demonstrate that the selected axis estimation algorithm estimates the axis of symmetry to a very high degree of accuracy with the largest source of error being the quantisation of the angles in the histogram. The axis of symmetry of symmetric objects with off centre axes of symmetry in the image are also detected. The algorithm was shown to be very robust to noise due to the filtering effect of the GVF field calculation. It is also shown that the suggested alterations to the algorithm improve on both the efficiency and the accuracy of the algorithm. Once again, the largest source of error is in the quantisation of the histogram bins.

Results for the continuous symmetry measure demonstrate improved smoothness in the coefficient manifold for images that are deviating from perfect symmetry. The three experiments that were carried out show that the suggested WBSC cost function improves upon the comparison cost functions. No incorrect local minima are found with the WBSC cost function, and the single minimum that is achieved is the correct one. The WBSC cost function is also shown to be robust to lighting variations and noise. It is demonstrated that the alterations to the symmetry coefficient make it suitable to be used in an iterative optimisation process to recover pose.

The aspects of symmetry examined in this chapter may be further explored in a pose normalisation framework. Determining the position and the orientation

of the axis of symmetry allows a partial pose removal from the image. The symmetry axis may be aligned along the vertical axis of the image, thus removing in-plane rotation and translation. The wavelet based symmetry coefficient may then be used in an iterative minimisation process to remove out-of-plane rotations. At this point, the translation, in-plane rotation, and out-of-plane rotation will have been fully removed.

# Chapter 3

# Removal of Pose from Face Images

## 3.1 Introduction

For face recognition systems to operate optimally, subjects must maintain a neutral expression and look directly into the camera. Efforts have to be made by system creators to remove expression and pose, and allow users to be imaged in near frontal positions with any expression. Expression classification and removal is being carried out by collaborators on this research (Ghent and McDonald, 2003). It is pose removal that is considered in this chapter. A full discussion of current pose removal techniques was given in the literature review of Chapter 1.

Two systems are created to cope with varying pose in images. The first system removes facial pose after the images have been captured. This comes about through an optimisation of the wavelet-based symmetry coefficient (WBSC), proposed in Section 2.2.2, measured on re-rendered views of the subject. The second system aims to filter out non-frontal images through an online calculation of the WBSC. Images that deviate too far from ideal values are rejected.

Capturing a database of images for experimental verification is described in section 3.2. Section 3.3 describes the proposed algorithm to remove facial pose. Experimentation for the proposed pose removal algorithm is demonstrated in Sections 3.4, with both quantitative and subjective results shown. Section 3.5

describes the image filtration process with experimental results demonstrated. Finally, Section 3.6 gives a conclusion and direction for future work.

## 3.2 Capturing the varying face images

For experimental validation of the proposed algorithm, a database of subjects in various poses was required. Due to the collaboration with the Computer Vision and Imaging Laboratory at NUI Maynooth, a database of face images portraying various expressions and captured from different view points was captured. An experimental setup was built to allow the database to be captured. The aim of the experimental apparatus was to ensure consistency between the angles of rotation of each subject. The setup is described below.

To ensure a high degree of accuracy with the measurement of the angle of rotation, a high-precision Newport RV120-PE computerised rotation stage with a resolution of 1/1000th of a degree was used. Onto this, a chair was mounted so that the axis of the rotation stage would run through the centre of the subjects' heads. A narrow headrest was fitted to the chair to restrict the movement of each subject while the image capture process was taking place. A Newport ESP100 motion controller was used to connect the rotation stage to a computer. An Hitachi CCD camera (model KP-M1EK), attached to the same computer through a Matrox "Meteor 2" (model MC4) video capture card, was positioned directly in front of the rotation stage for image capture.

A system was developed, with the aid of the provided Newport C++ header files and the OpenCV toolbox (OpenCV, 2001), that allowed simultaneous control of both pieces of equipment. The software was designed so that the rotation stage would step through a series of motions from $-10$ to $+10$ degrees in steps of 1 degree, at each position capturing an image. Once the image capture process was complete, the stage was returned to its resting position of 0 degrees.

Some initial training of the subjects was also required. Since we required each subject to portray various expressions, a trained expert supervised the learning of the facial muscle movements required to form each expression through the Facial Action Coding System (FACS) (Bartlett et al., 1999). Throughout the process, the trained expert remained and monitored the expressions. The

whole process of capturing one set of 20 images took approximately 20 seconds, the majority of which time was spent on the rotation stage movements. The complete database consisted of 5 subjects portraying 5 expressions in each of 20 views, 500 images in total. Some examples of the captured images are shown in Fig. 3.1.

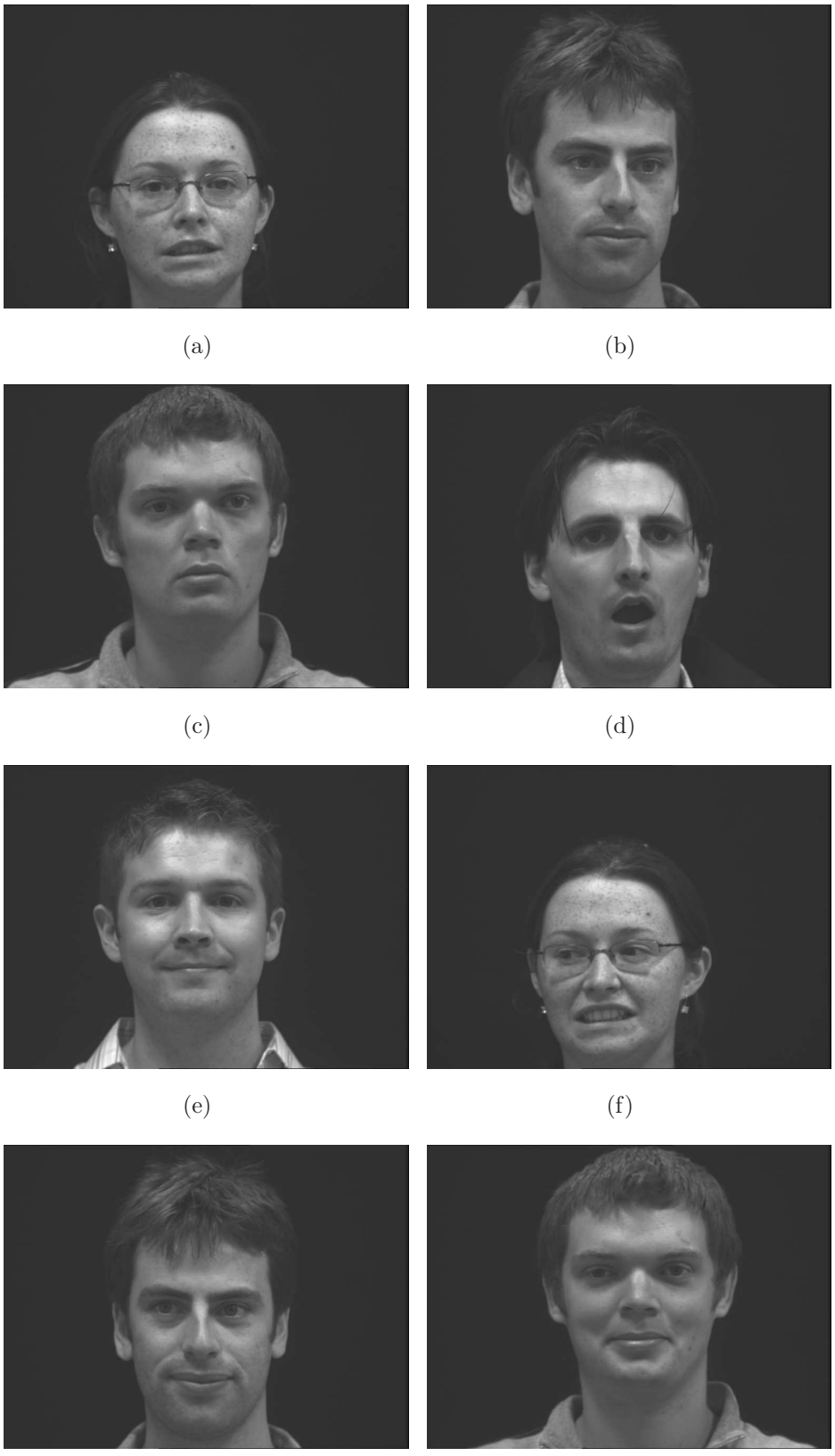
## 3.3 Proposed Pose Removal Algorithm

Since faces exhibit a great deal of symmetry, symmetry will be used to restore the pose of each subject back to a fronto-parallel view. Each image is rectified by calculating the WBSC cost function, developed in Section 2.2.2, on each re-rendered view of a model texture mapped with the face image. The cost function is minimised in an iterative optimisation process.

Pose removal will be carried out using a technique where the face image will be re-rendered in a new view based on a measure of the symmetry. Three different techniques for generating the new view to be rendered are considered. The face images will be texture-mapped onto a planar surface, a cylinder, and an ellipsoid respectively, to form the basis for the three different view creation techniques. The WBSC cost function is then optimised to determine the optimum view to remove the pose variation.

Although some degree of additional distortion may result from projecting the face images onto a plane, cylinder and ellipsoid, the reason these three shapes were chosen is that they represent approximations to the shape of a face. The plane is the simplest approximation to the shape of a face, and is also the simplest shape to transform. Moving on from this approximation, a cylinder better represents the shape of a head, and mapping the face onto one side of this will be a close approximation to the correct curved shape of the face. To fully capture the shape of the face, the curvature of the face in both the horizontal and vertical directions needs to be considered. This is available in ellipsoid shapes. The ellipsoid shape stops just short of using a $3D$ model of the face onto which the face image could be mapped, which would require accurate feature detection and correspondence matching for image and shape registration, which is one of the problems we are trying to avoid.

The algorithm operates as follows. The face is segmented from the background



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 3.1: Sample Images from the database of subjects portraying various expressions and poses.

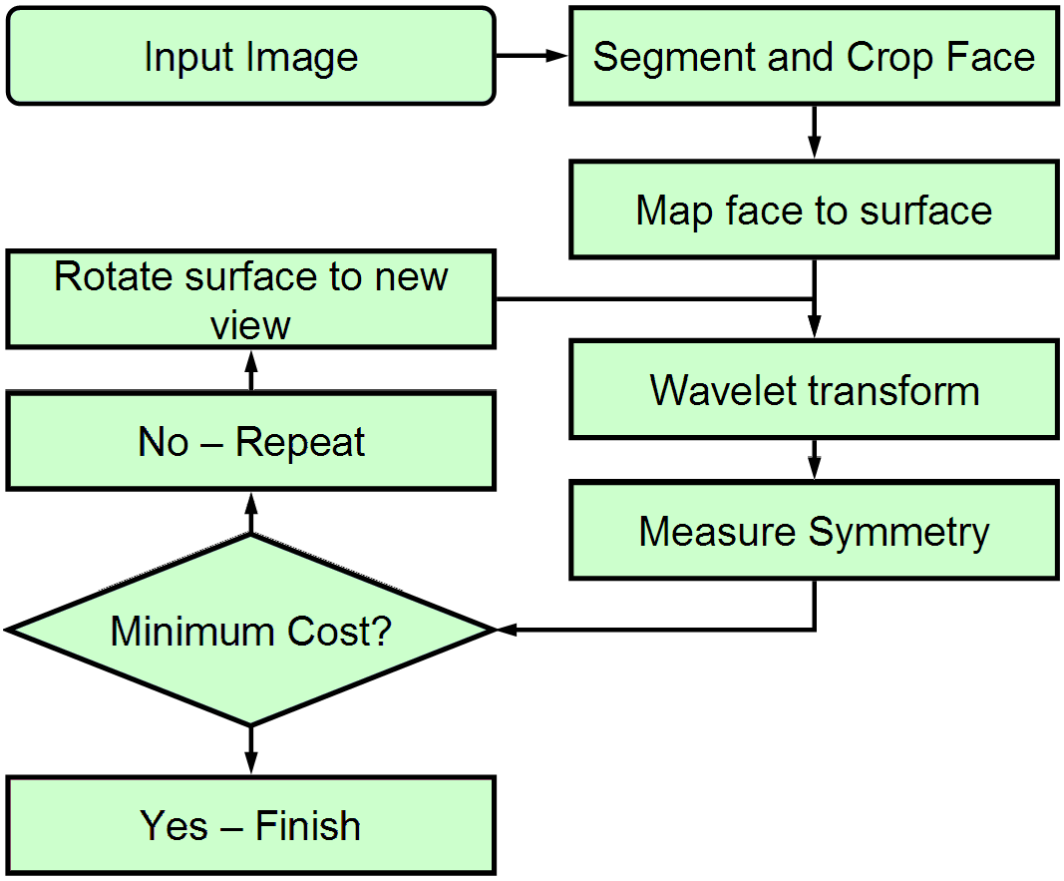



Fig. 3.2: Flow Diagram representation of the Wavelet-based Face Pose Removal (WFPR) algorithm

of the image using a background subtraction technique. Each detected face region is then cropped. The cropped face image is texture mapped to one of the three surface types. In the case of the cylinder and ellipsoid techniques, the cropped image is texture mapped to one half of the surface with the other half being left blank.

The texture-mapped surface is then rotated to a new viewing position and the projection of the surface onto the image plane is taken. The continuous wavelet transform of the image is calculated to obtain the frequency response of the new view at each position in the image. At this point, the WBSC cost function of the re-rendered view is calculated. A non-linear optimisation scheme is employed to minimise the symmetry cost of the wavelet transformed views and obtain the rectified image. The algorithm, called the Wavelet-based Face Pose Removal (WFPR) algorithm, is outlined in Fig. 3.2. A diagrammatic representation of each of the three view creation techniques is given in their respective experimentation sections.

## 3.4 Experiments

Three experiments are carried out, each of which are compared directly to the dense matching and gaussian dense matching methods, as described in Section 2.2.3. The first experiment uses a planar approximation to the face images in the rectification process. Faces are not planar, and projecting the face images onto a plane may introduce additional distortions, but the distortions will be minor for small out-of-plane rotations of the faces. In the second and third experiments, the face images are mapped to a cylinder and an ellipsoid respectively. In each case, the WFPR algorithm is run on each image and for the second and third experiments the angle of rotation of the mapped shape is recorded. These captured angles are compared to the ground truth information for quantitative evaluation. No rotation angle for the planar approximation method can be derived, since the matrix of internal camera parameters is unknown. Without this camera matrix, the planar transformation solutions obtained can not be decomposed into constituent rotation matrices about the $x$, $y$ and $z$ axes. In each experiment, the comparison methods are also employed, and it should be noted that with the coarse to fine gaussian filtered approach a number of iterations of the process are required, with less gaussian blurring being employed on each successive iteration.

### 3.4.1 Experiment 1 - Planar Approximation

Initially, each image is processed to segment and crop the face from the background. At this point, although faces are not planar, the face is treated as a plane with planar transformations being applied for the remainder of the experiment. Since the motion of the subjects is restricted to rotations about the yaw axis, solutions are restricted to planar transformations that deviate from $I_{3x3}$, the identity matrix, in only the $x$ perspective parameter.

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ g & 0 & 1 \end{pmatrix} \tag{3.1}$$

New views are created by rotating the textured plane about its vertical axis, and taking the projection of the rotated textured plane onto the image plane. The re-rendering scheme can be seen in Fig.3.3.

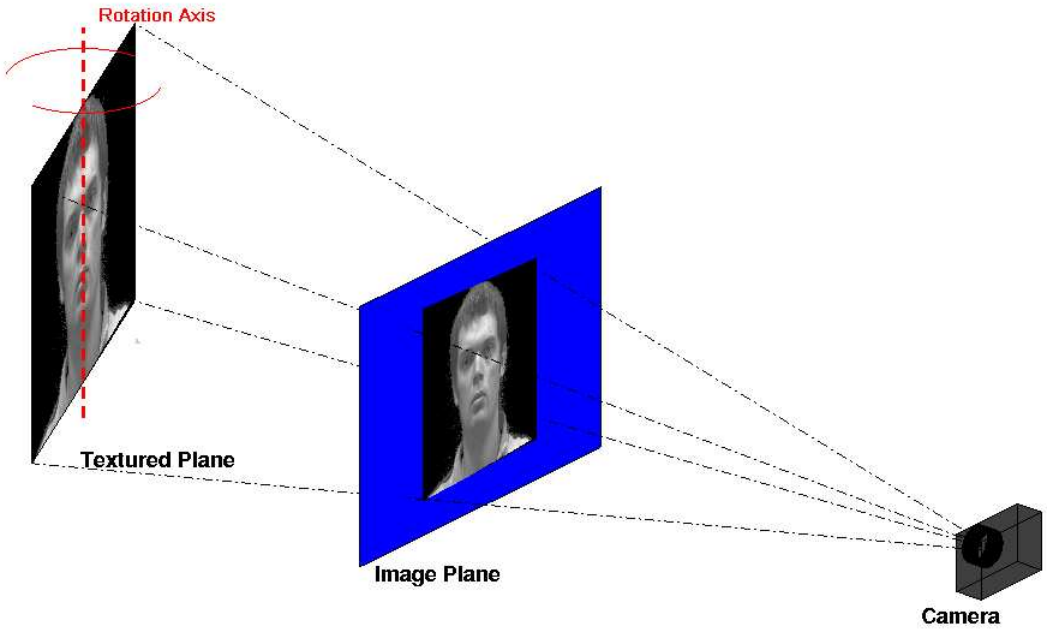



Fig. 3.3: Creating synthetic views of the face based on a planar approximation.

The WFPR algorithm is used to remove the pose from each face image. Each time the algorithm converges to a solution, the resulting image is stored. Due to a lack of priori camera information, the rotation angle can not be extracted from the retrieved perspective transformation. It should be noted that an alternative error matric could be employed in place of the rotation angle. Each pose removed image could be compared against the image that is known to be in the fronto-parallel position from the ground truth data, and a correlation measure between the two images could be calculated. In order for a metric of this nature to have meaning, a more restricted environment with even lighting conditions would have to be employed to ensure even lighting of the faces in all orientations. Only subjective results are presented here.

From the results in Figures 3.4, 3.5, and 3.6, it can be seen that the face images are rotated in the correct direction. The WFPR algorithm performs well, removing a large portion of the pose. There are however some distortions in the face images due to the planar approximation made.

### 3.4.2 Experiment 2 - Cylinder Mapping

Once again, each image is initially processed to segment each face from the image and crop the image to the face region. The cropped face region is then

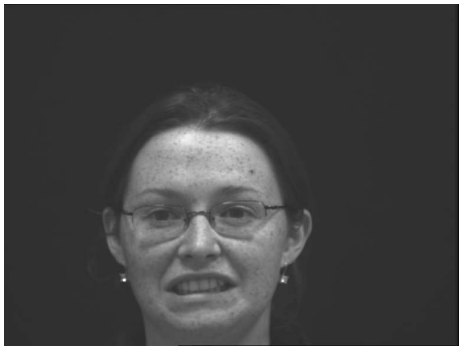
(a)



(b)

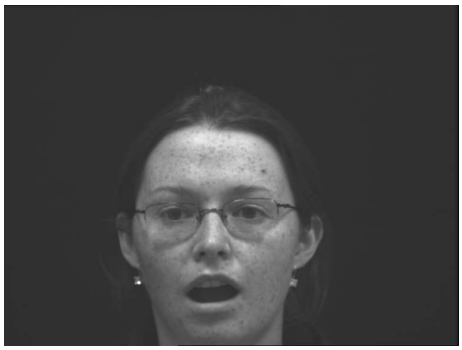
(c)

(d)



(e)

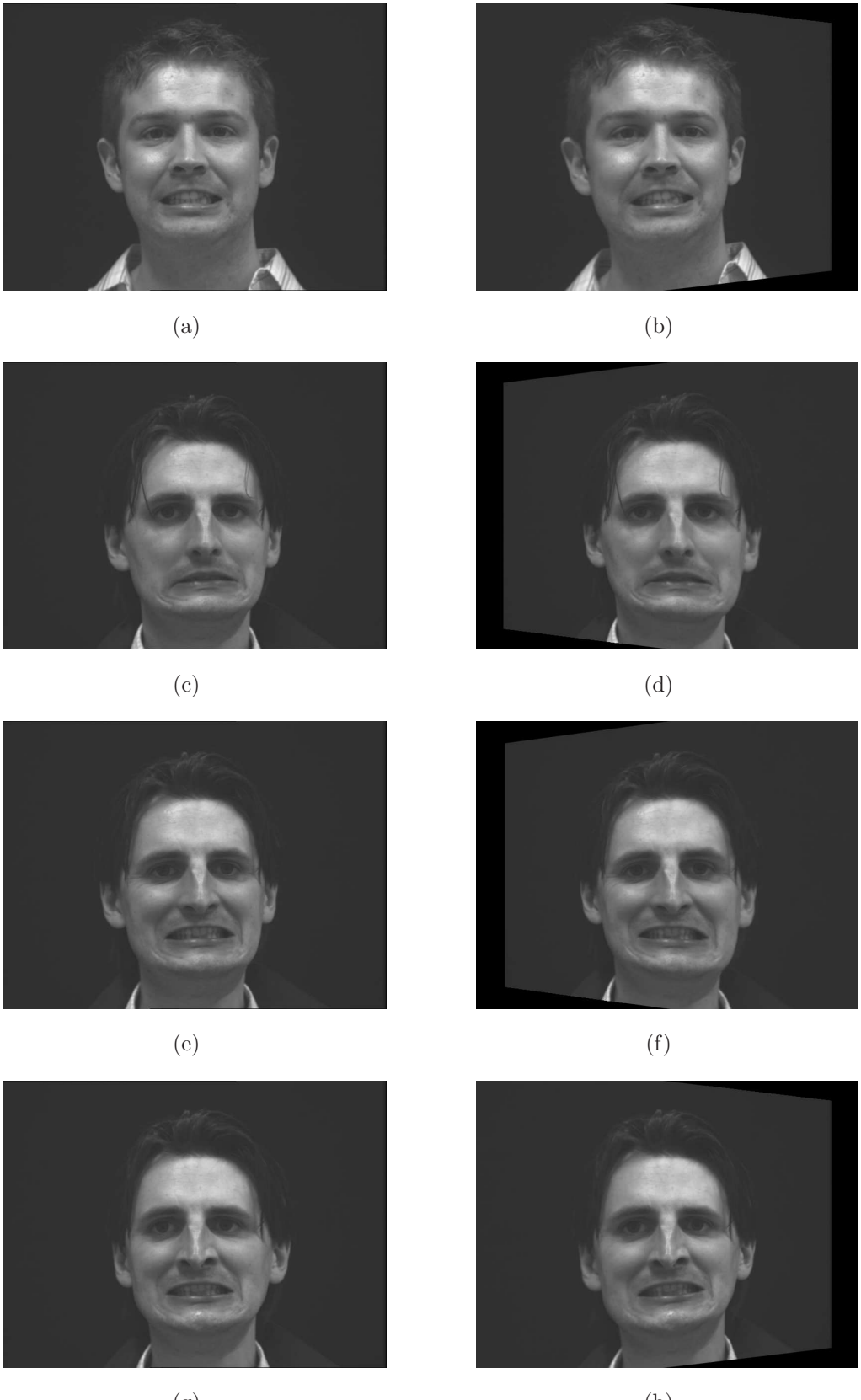
(f)



(g)

(h)

Fig. 3.4: A sample of images that are rectified using the planar view generation technique. Column 1 displays the input images. Column 2 displays the output images.



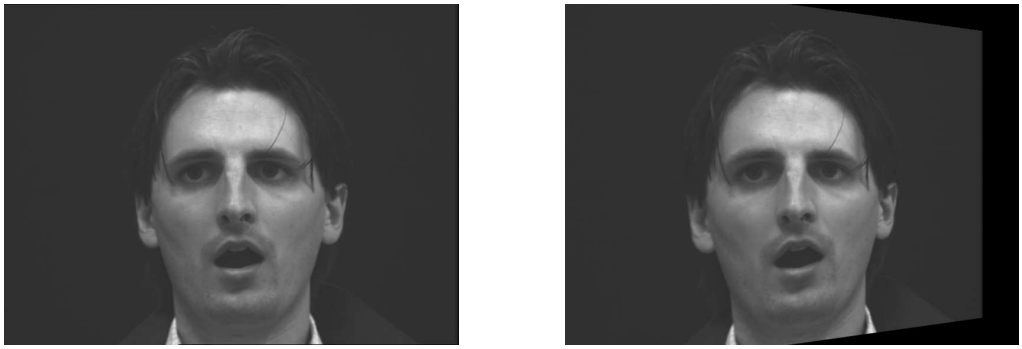
(a) (b)

(c) (d)

(e) (f)

(g) (h)

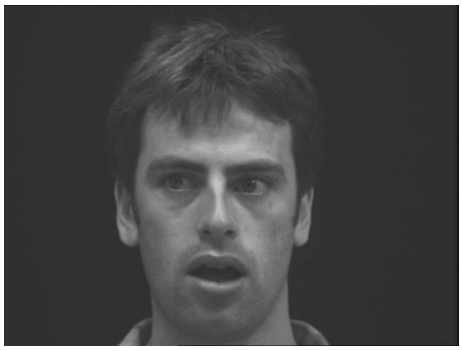
Fig. 3.5: A sample of images that are rectified using the planar view generation technique. Column 1 displays the input images. Column 2 displays the output images.



(a)



(b)

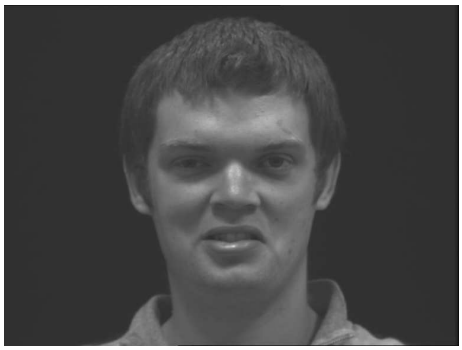


(c)

(d)

(e)

(f)



(g)

(h)

Fig. 3.6: A sample of images that are rectified using the planar view generation technique. Column 1 displays the input images. Column 2 displays the output images.

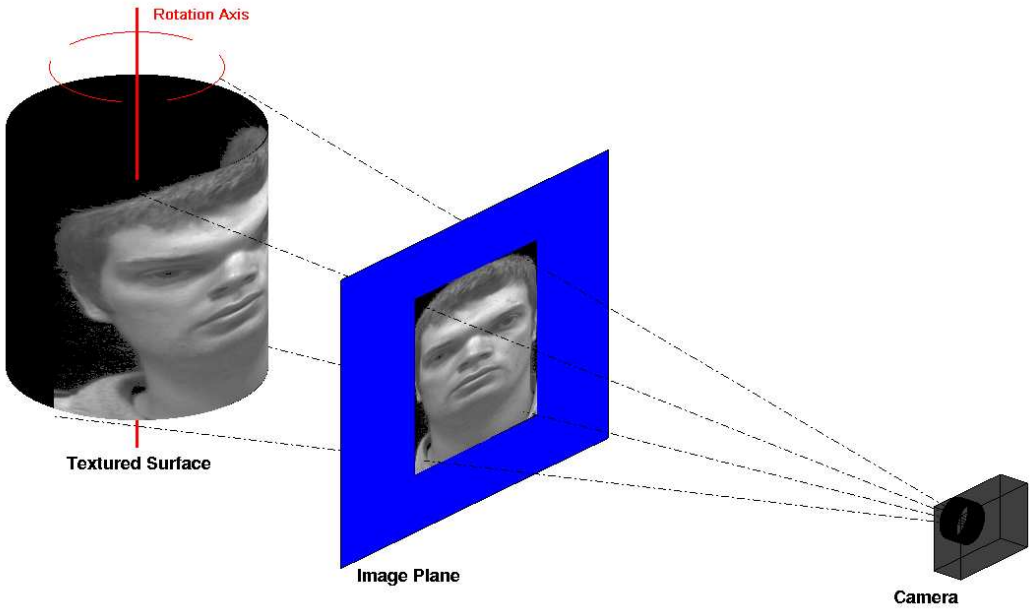



Fig. 3.7: Creating synthetic views of the face based on a cylinder mapping.

mapped to one half of a cylinder, the other half is left blank. The cylinder is rotated about it's vertical axis and the projection of the mapped cylinder onto the image plane is then used to create the new view as shown in Fig. 3.7.

A rectification of the face image is obtained with the WFPR algorithm. Similar rectifications are achieved with each of the comparison algorithms. The coarse to fine approach is run 5 times, with progressively less gaussian blur at each stage and the previous result being used to initialise the next phase. For each of the algorithms used, the angle of rotation is recorded for each result obtained. There are 25 images of subjects in each orientation that are rectified, from which each error is computed. The error at each orientation is computed as the mean absolute error across the 25 images of that orientation. The results are presented in Fig. 3.8.

From Fig. 3.8, it can be seen that the WFPR algorithm achieves better rectification results than the two comparison methods for rotations of the subject between $-4$ and $+4$ degrees. Also, the errors obtained with the proposed WFPR algorithm increase more smoothly with increasing rotation angle than the comparison methods. A bias in both the dense matching and gaussian blurred schemes is visible in Fig. 3.8 resulting in a perceived improved performance over the WFPR algorithm. However, the WFPR algorithm is more stable and is also without bias.

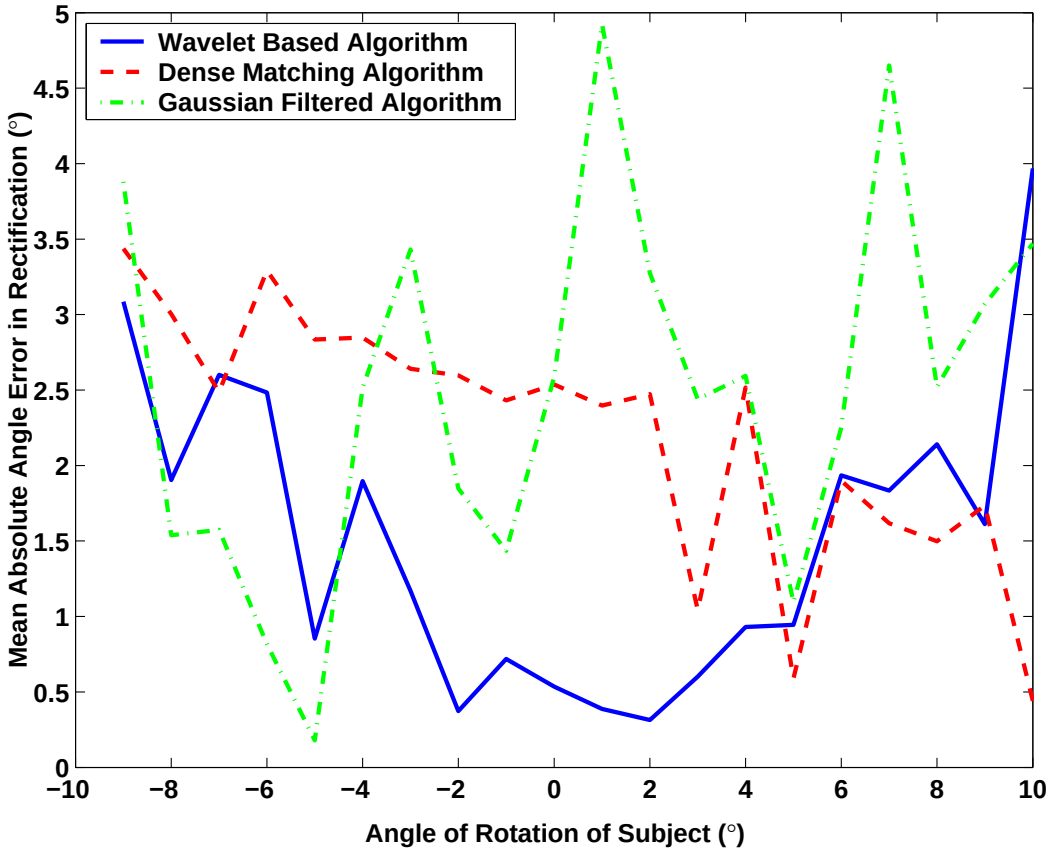
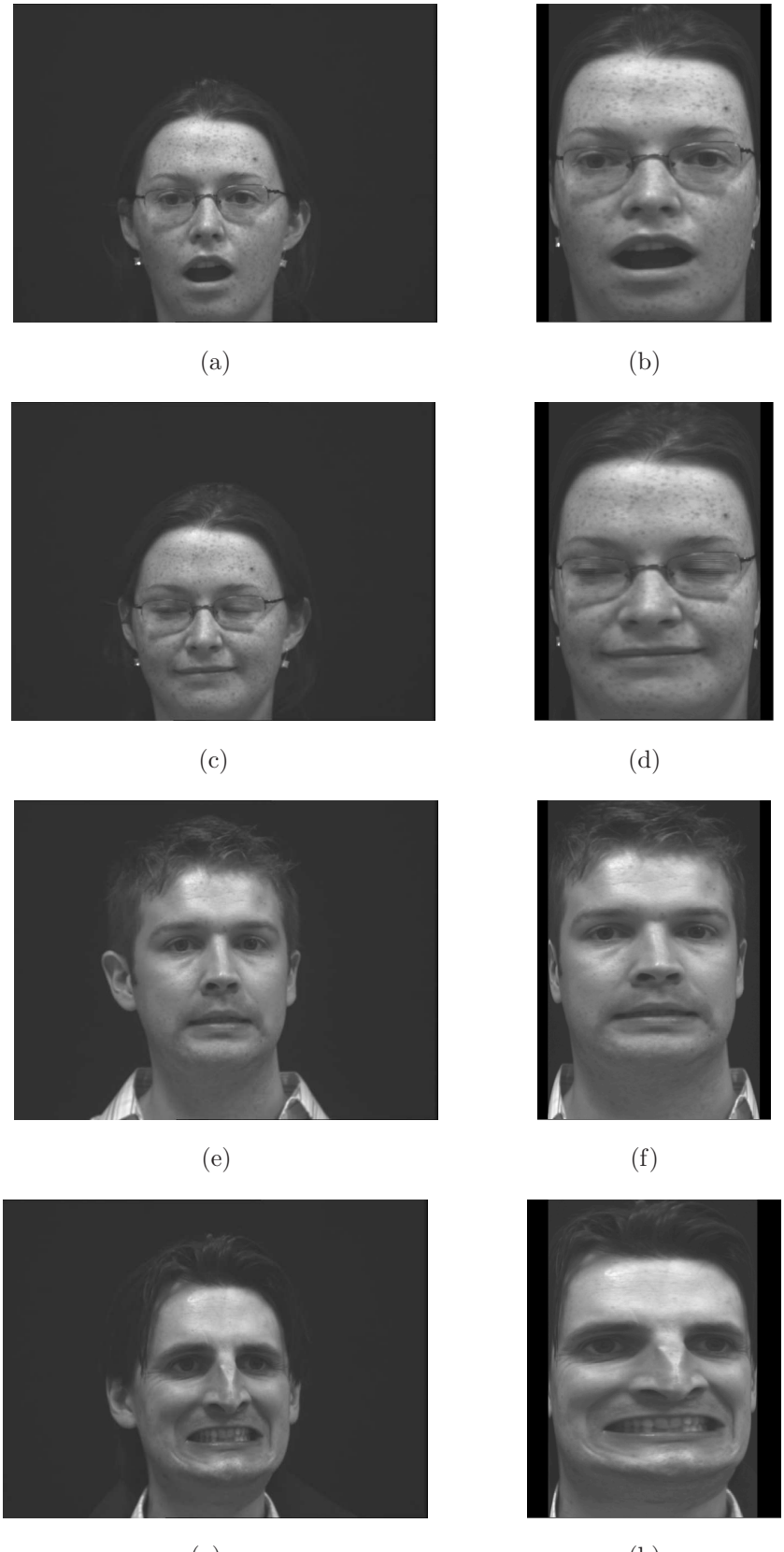



Fig. 3.8: Mean absolute angle error in rectifying faces with the texture-mapped cylinder

To further demonstrate the rectification abilities of the WFPR algorithm, additional subjective results are demonstrated in Figures 3.9 and 3.10. From the output image results, it can be seen that the face images appear to be imaged from a more fronto-parallel view point than the input images. Additionally, the images appear less distorted than those obtained with the planar approximation technique.

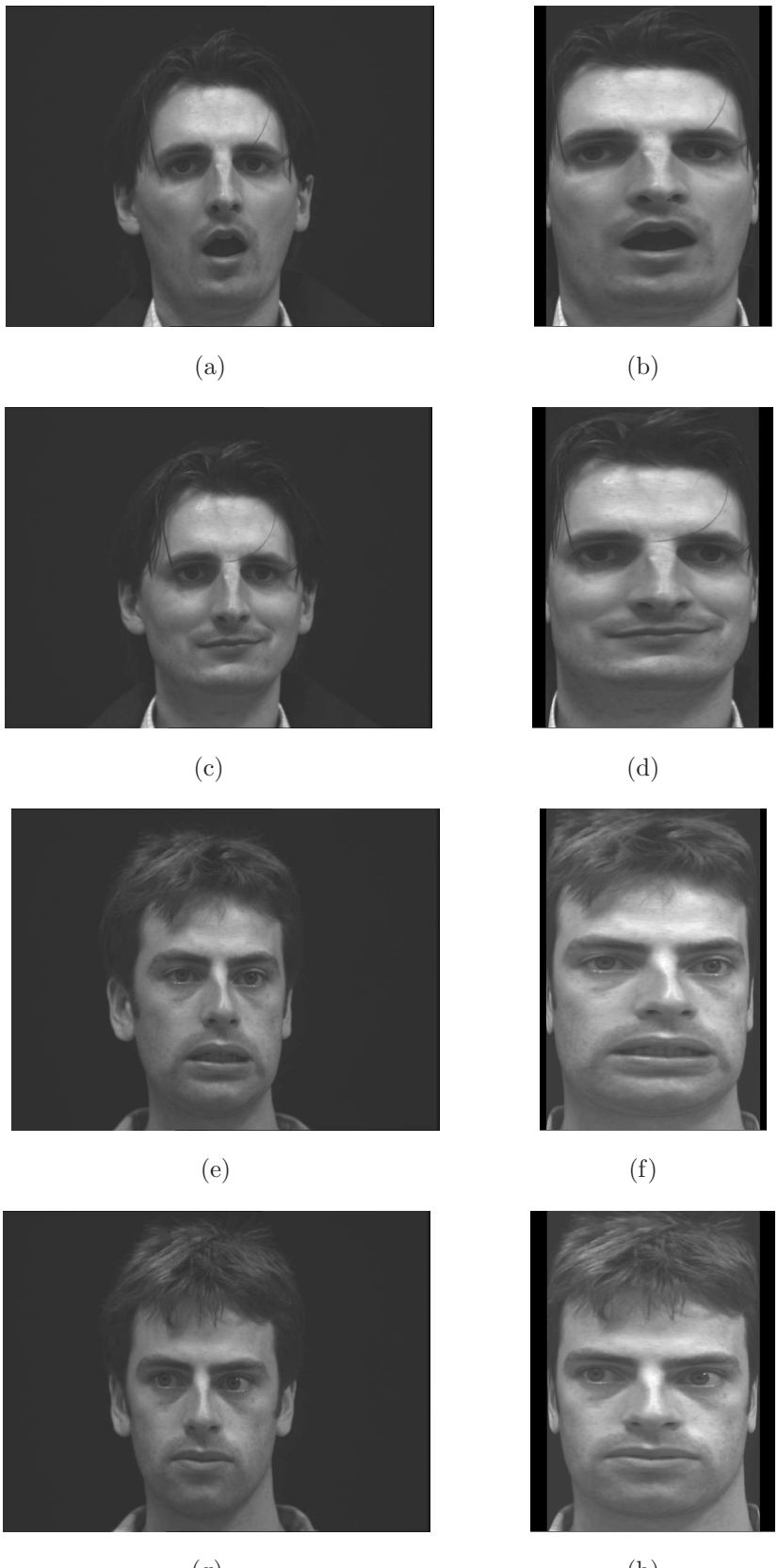
### 3.4.3 Experiment 3 - Ellipsoid Mapping

Once the faces have been segmented and cropped from the images, they are texture-mapped to one half of an ellipsoid with a vertical diameter 1.5 times the two horizontal diameters which are equal, which is the approximate shape of a human head. To accomplish this, the face images are registered with the ellipsoid shape such that the axis of symmetry of the face is aligned vertically with the long axis of the ellipsoid. New views of the subject are then created by



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3.9: A sample of images that are rectified using the textured-cylinder view generation technique. Column 1 displays the input images. Column 2 displays the output images.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3.10: A sample of images that are rectified using the textured-cylinder view generation technique. Column 1 displays the input images. Column 2 displays the output images.

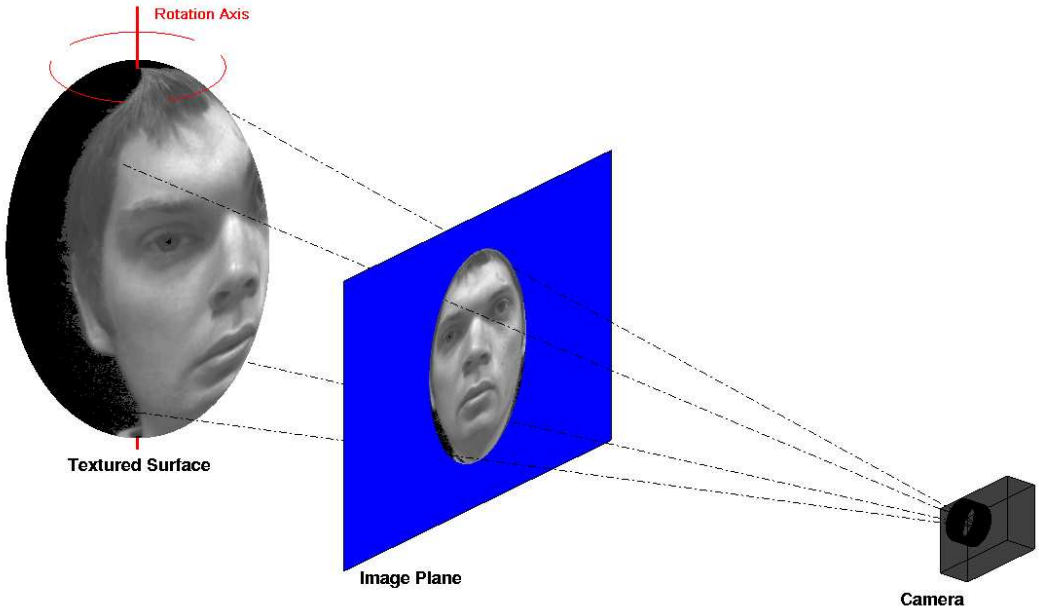



Fig. 3.11: Creating synthetic views of the face based on an ellipsoid mapping.

rotating the ellipsoid, texture-mapped with the face image, about its vertical axis. A projection of the texture mapped ellipsoid onto the image plane is then taken.

The WBSC cost function and the comparison cost functions are then computed on the projection of the ellipsoid onto the image plane. These costs are minimised by obtaining the angle that provides the minimum cost in an optimisation scheme. There are 25 images of subjects in each orientation that are rectified, from which, each error is computed. The error for each orientation is computed as the mean absolute error of the 25 images. A graph of the results is shown in Fig. 3.12.

As it can be seen from Fig. 3.12, the WFPR algorithm performs better than the comparison methods for the majority of the presented angles. The algorithm is more stable and displays a more smooth error curve, which provides the user with a higher confidence in the algorithm. For the larger positive angles of rotation, the comparison methods demonstrate superior performance compared to the proposed WFPR technique. This can be attributed to the bias that is evident for each of the comparison techniques, which may have resulted from unbalanced lighting conditions. The WFPR technique is more robust to unbalanced lighting conditions and as such shows no bias.

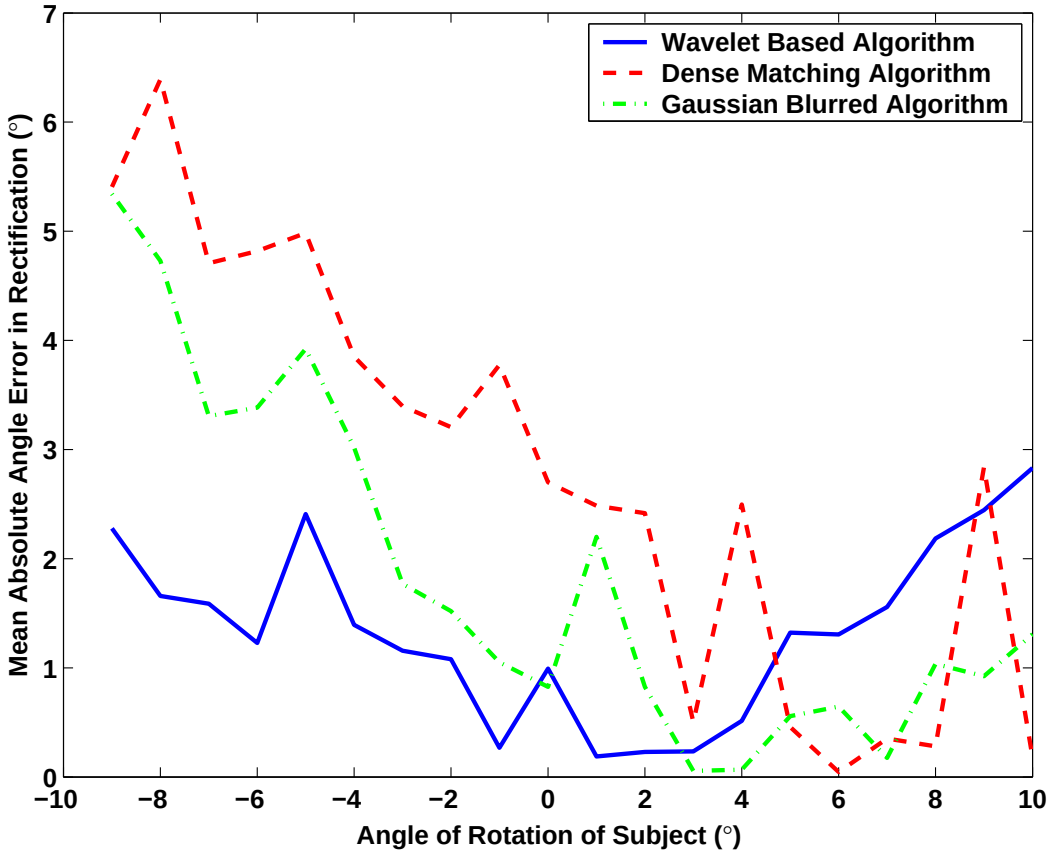
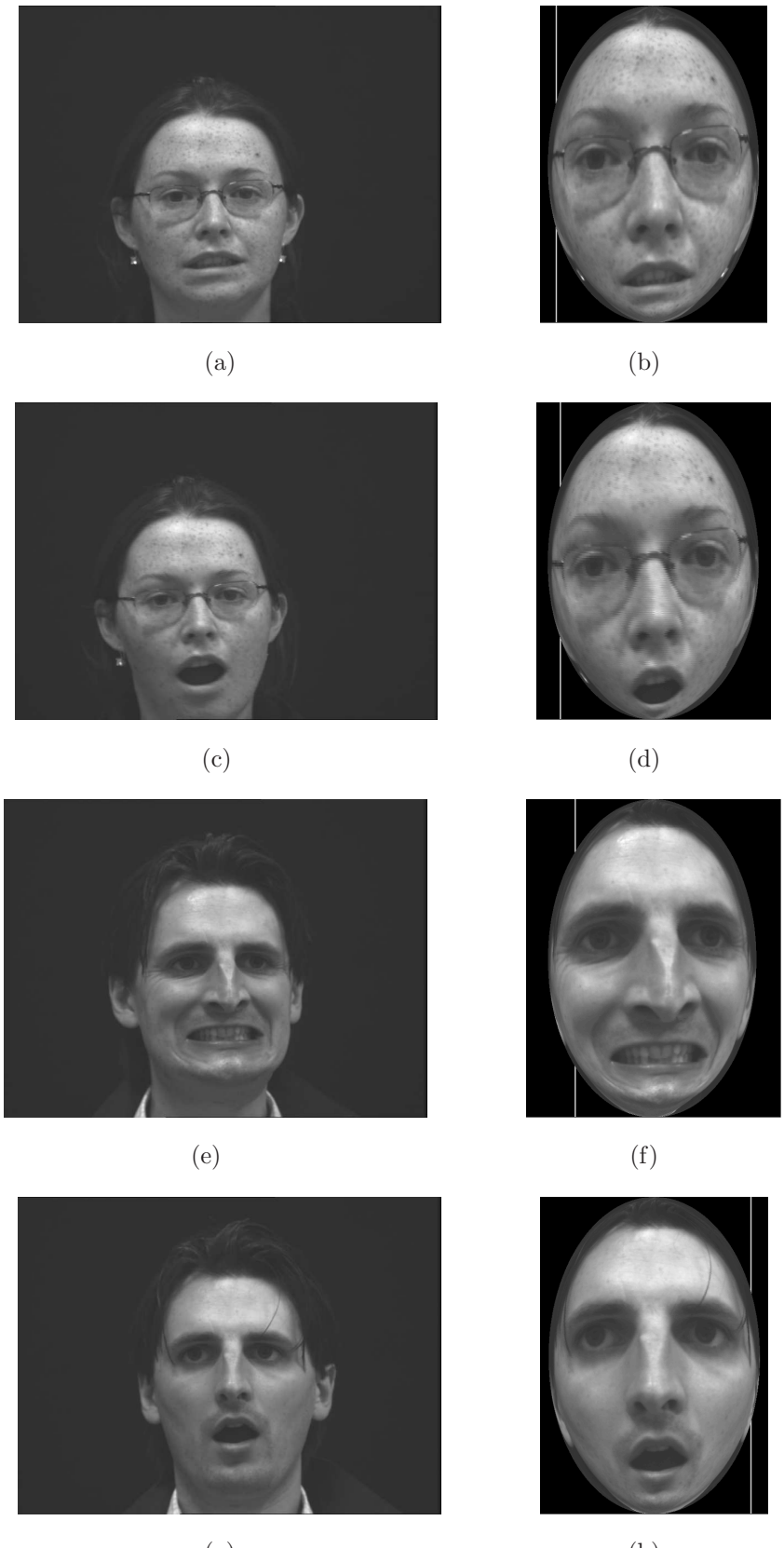



Fig. 3.12: Mean absolute angle error in rectifying faces with the texture-mapped ellipsoid.

Additional qualitative results are demonstrated in Figures 3.13 and 3.14. From the output images, it can be seen that the subjects appear to be imaged from a nearly fronto-parallel view point. Additionally, the images appear less distorted than those obtained with both the planar approximation technique and the cylinder-mapping technique. This is due to the closer approximation to the correct shape of the head that is achieved with the ellipsoid mapping technique over the other two techniques.

## 3.5 Selective Retrieval of Faces from Video

It has been shown in the preceding sections of this chapter that it is possible to remove pose from face images after they have been captured, with few restrictions on the input images. The algorithm is robust to noise and lighting variations and does not rely on the accuracy of feature detection or point correspondence algorithms. However, for each image that is processed,



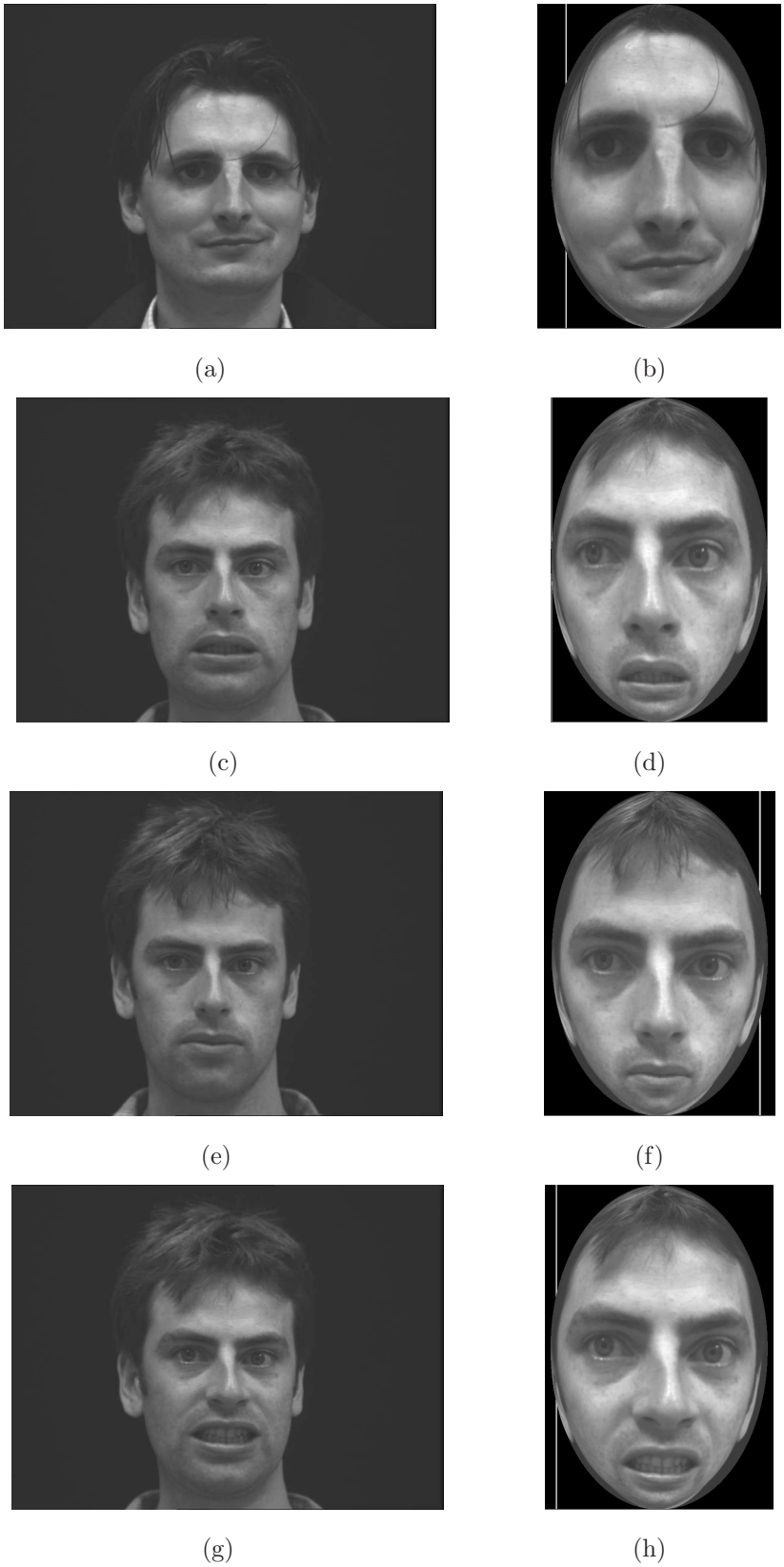
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 3.13: A sample of images that are rectified using the textured-ellipsoid view generation technique. Column 1 displays the input images. Column 2 displays the output images.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3.14: A sample of images that are rectified using the textured-ellipsoid view generation technique. Column 1 displays the input images. Column 2 displays the output images.

a certain amount of computation time is required. Also, a system without intelligent image selection abilities will need to store all captured images as well as all processed images. These two hindrances cause a significant drain on computing resources such as disk space, processing power and available RAM. To overcome these inefficiencies, an image selection algorithm is required.

The suitability of captured images of subjects' faces for the purposes of face recognition, can vary greatly due to pose. Although the pose can be removed, it would be more efficient to select only those images that appear to be imaged from close to fronto-parallel. Two approaches can then be applied, which differ in the number of input frames that will be stored. Either, all input images are stored and only those achieving a high quality score are processed for pose removal, or, only those input images achieving a high quality score are stored, all of which are then processed for pose removal. In both cases the computational saving is two-fold. Firstly, the space required to store the processed images, and possibly the input images is reduced. And secondly, the total computation time required to remove pose from these images with a high quality measure is reduced due to the lower number of images.

The fronto-parallel quality of the images can be assessed using the WBSC from Section 2.2.2. Each face is segmented from its input image and the face region cropped from the image. The symmetry cost measure can then be found on the wavelet transformed image of the subject's face. Low values of the symmetry cost indicate good quality images. Hence, a threshold on the allowable quality of the images can be enforced prior to image storage and pose removal can then be performed on the stored images. To highlight the selective abilities of the proposed pre-processing algorithm, an experiment was carried out.

### 3.5.1 Experiment

Using the database of captured images detailed in Section 3.2, an experiment was carried out. Each image was assessed for the pose of each subject based on the proposed WBSC cost selection technique. A high degree of symmetry indicates a good subject pose relative to the camera.

For each of the 500 images in the database, the face region was segmented and cropped from the background. Then the WBSC was calculated for each image.

The quality of each image was recorded in the symmetry value obtained. Figure 3.15, on the proceeding page, shows a subset of the captured and cropped face images each with a colour border around them. The colour of the border indicates how symmetric each image is based on the WBSC, and hence, how fronto-parallel each image is. Additional results are presented in Appendix A.

From the results, it can be seen that a large majority of the images could be ignored for face recognition purposes. Images with blue to navy borders indicate almost fronto-parallel subjects. These images are suitable for storage for further pose removal and person identification. It can be seen that only 4 from every 20 images would be stored for pose removal. This would have a computation time saving of 80% for pose removal.

The experimental results highlight the time and storage space saving that can be achieved through intelligent selection of input images to pose removal algorithms. An online system could be developed to quickly asses security footage to select and store face images of high quality only, and filter out the poorer quality images. This will provide better quality face databases with which to compare unseen test images, as well as reduce the total storage required.

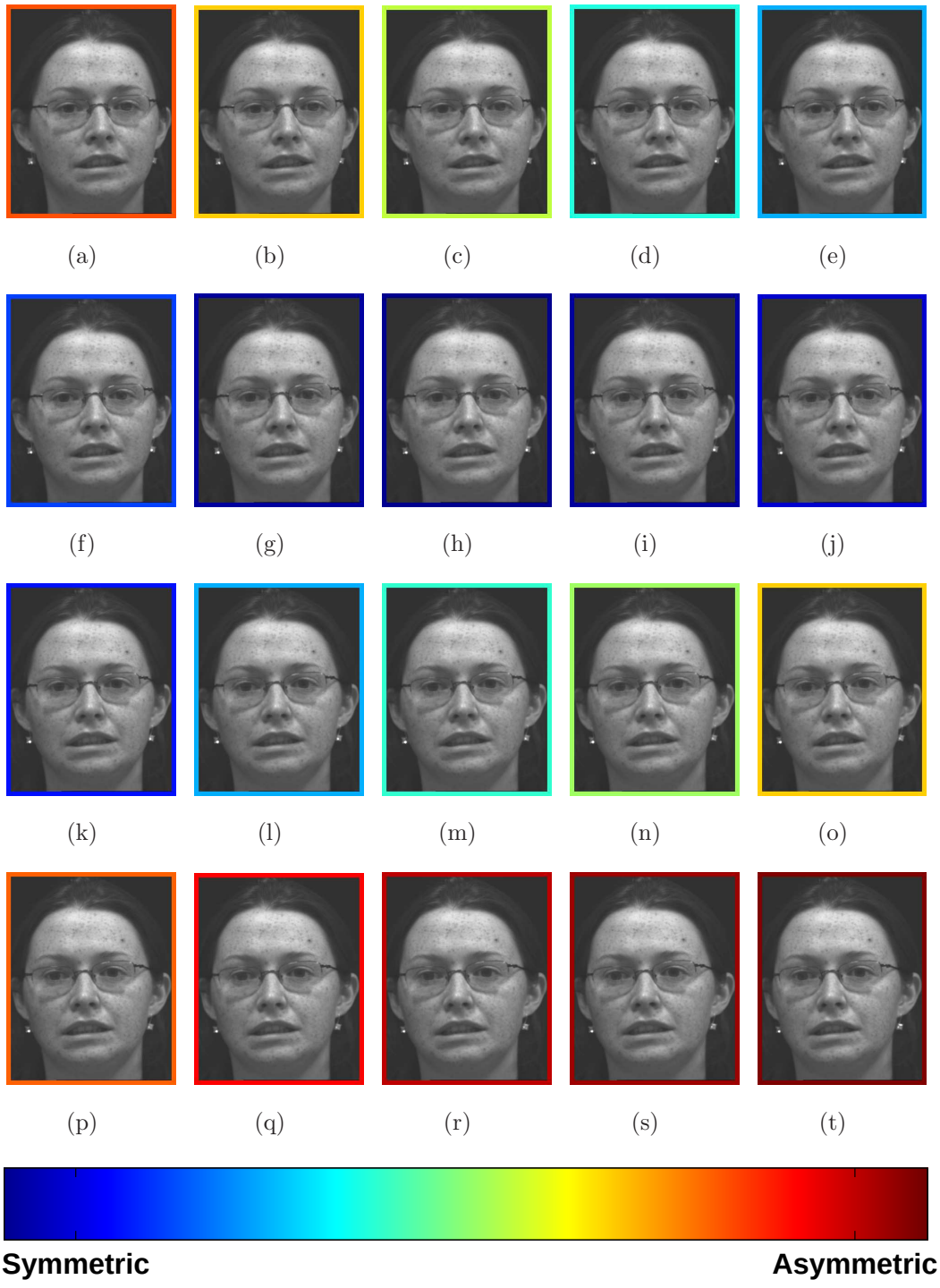



Fig. 3.15: Sequence #1 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

## 3.6 Conclusions and Future Work

A novel method for facial pose removal was presented. No 3D face model was required to be trained and no statistical learning process was used, the WFPR algorithm operates on each image individually. The Wavelet-based Face Pose Removal algorithm was implemented with the WBSC cost. Experiments were carried out on a database of real face images to remove pose. Although the subjects in the database have their motion restricted to yaw rotations, in plane rotations could also be handled with the use of the CCSAE technique. The results demonstrate improved rectification over the comparison methods. From both the subjective and numerical results, it can be seen that the ellipsoid rectification algorithm provides the most realistic results with little distortion. These improved results were achieved with few restrictions on the input data.

In order to reduce the amount of computation time required in obtaining pose removed face images, an image selection algorithm was designed. Images were assessed based on the WBSC calculated on each image. The images portraying high degrees of symmetry, were deemed to be almost fronto-parallel and could be directly used in person recognition or further processed to remove remaining pose. The results demonstrated that differentiability between fronto-parallel and otherwise posed images is easily achievable with the selection method. This selection algorithm could be used as a pre-processing stage for face database acquisition from video sequences to ensure the storage of high quality images.

To further highlight the advantages of using the wavelet based rectification algorithm, it would be desirable to benchmark the accuracy rate of a recognition system before and after pose removal. Also, a comparison between the accuracies of all current recognition systems that operate on posed face images and the accuracy rate of a recognition system operating on the images obtained using the WFPR techniques would be carried out.

# Chapter 4

# Removal of Pose from Imaged Planar Textures

## 4.1 Introduction

Through the examination of pose removal from face images in the preceding chapter, the continuous wavelet transform was used to obtain the frequency response at every position in images. Having information on how the frequency response changes across an image gives valuable information about the structure of the imaged scene. Further investigation into the recovery of pose from views of an imaged plane are given in this chapter. A full literature review was carried out in Chapter 1, but further more detailed examinations of selected current methods are presented in this chapter to highlight the progression of thought that led to the use of the space-frequency domain.

The relative orientations of two views of an imaged plane may be described using an eight parameter spatial transformation matrix known as an homography. This transformation matrix is a valuable measurement, because information about the scene's structure, the camera's orientation, and the camera's internal parameters may be extracted from one or more of these homographies. Therefore it is essential that there are accurate methods available with which to estimate homographies.

To summarise the literature review of Chapter 1, there are two general approaches to estimating planar homographies. Firstly, using spatial domain

techniques, strong features are extracted and matched in two views of the planar scene. Using linear estimation techniques the homographies relating views are estimated. The solutions can be refined using non-linear iterative solution refinement techniques. And secondly, using featureless frequency based techniques, a Fourier transform of each view is calculated. With these Fourier transforms, affine relations between the two views can be established using non-linear iterative techniques. The limitations to using each approach are as follows. With the spatial domain techniques, strong features need to be extracted and matched across multiple views to estimate perspective transformations, a task that is very difficult to achieve in cluttered or noisy scenes. And with the featureless, frequency-domain techniques, although the feature extraction and correspondence matching problems are eliminated, no perspective transformations may be estimated at present.

This chapter presents a novel approach to planar transformation estimation. Perspective transformations will be estimated in a featureless framework through the use of the space-frequency domain. This incorporates the better aspects of both the spatial domain and the frequency domain techniques. At present, no other technique operates in a featureless regime to estimate perspective planar transformations. The proposed technique will be shown to be capable of estimating full perspective transformations.

This chapter is organised as follows. Section 4.2 describes the relevant current methods of estimating homographies in both the spatial and frequency domains to give a fuller understanding of each domain's advantages and limitations. The proposed planar transformation estimation technique is described in detail in Section 4.3, with mathematical derivations and algorithmic details provided. In Section 4.4, experiments are presented that highlight the advantages of using the proposed technique. The proposed method is compared directly against another featureless rectification technique, showing favourable results. And finally, in Section 4.5, conclusions are drawn from the experimentation, with a summary given.

## 4.2 Current Planar Transformation Estimation Techniques

The principal idea behind all of the methods is to find the $3 \times 3$ matrix that is the linear transformation matrix described in Appendix B. A more in depth description of each of the main branches of image rectification will be given to clarify how each solves for this matrix. A complete comparison will be given in the last section.

### 4.2.1 Spatial Domain Methods

Spatial domain techniques operate directly on geometric features, and require no image representation transformation. The most popular and straightforward technique is outlined in Appendix C, the Direct Linear Transform (DLT).With the DLT, a transformation between two view of the same imaged plane may be obtained. The required steps to estimate the transformation matrix, or homography, are as follows:

(i) Detect geometric features, such as corners, in both views of the imaged plane with a corner detector

(ii) Establish correspondences between the detected features in both views

(iii) Use the DLT to estimate H, the transformation matrix

This method can be used to directly calculate all planar image transformations in an efficient manner. The spatial domain techniques are prone to erroneous results due to noise issues with the feature detection and matching processes. The Direct Linear Transform and other spatial domain techniques are detailed in Hartley and Zisserman (2003), but will be given a brief discussion here.

#### The Direct Linear Transform

There are two main sources of error associated with using the DLT method, poor feature localisation and correspondence mismatches. Poor feature localisation comes about when a detected feature in one view of the imaged plane is

detected in the incorrect location, possibly due to noise or pixelisation. Mismatched features may come about because similar but unrelated features in two views are mistakenly deemed to be similar enough to be the same feature. These two sources of error will introduce error into the system of equations to be solved, resulting in an erroneous solution.

In the case where the $A$ matrix, the formation of which is described in Appendix C, is overdetermined, with more than 4 point correspondences found, each mismatch or poorly located feature adds to the overall noise in the solution. For this reason, large numbers of corresponding point pairs are used with the assumption that if error in the localisation of points is gaussian distributed with a mean of zero, then the error due to poor feature localisation will also tend to zero Mallon and Whelan (2007).

For the exact solution scenario, where exactly 4 point correspondences will be used, poor feature localisation may lead to a situation where features may not be linearly independent of the other detected features. This leads to a situation where the $A$ matrix is under-determined and a null-space of 2 or more dimensions will be found. This yields a family of solutions for the transformation matrix, where no single solution uniquely defines the transformation between the two sets of detected points. Methods exist to combat the problems of poor feature localisation and correspondence mismatches which are described below.

### Improving the Technique and Refining DLT Estimate

The hindrances to using the standard DLT method, typically involve the problems with feature localisation, and correspondence mismatches. Other *robust* methods have been developed that account for these noise errors and mismatches. One of the most successful of these is the *RANSAC* method, developed in Fischler and Bolles (1981). The algorithm selects a random subset of the detected feature points in one image, and using the DLT, estimates the transformation between these and a random selection of the imaged points in the second image. This transformation is used to transform one set of points onto the other. The number of points supporting this estimation, the number of points that have overlying correspondences, are counted and stored. Further random estimates are carried out until a high enough count is achieved.

All supporting point pairs are then used to find the direct linear solution. This overcomes the initial issues with automatically matching features in the presence of noise. However, due to the non-zero distance tolerance between projected and measured points when obtaining supporting point pairs, the optimum solution is not immediately obtained. Further refinement of the estimate is required to achieve the optimum solution.

There exist a number of methods with which the homography matrix estimation may be refined. A selection of these methods are described in detail in Hartley and Zisserman (2003). These involve finding the minimum possible distance between the transformed points from one view and the imaged points in the second view. These refined estimates come about through minimising other, non-algebraic distance measures such as the geometric error in one image, the symmetric geometric error, the re-projection error and the Sampson error which is an approximation to geometric error. Although direct solutions with some of these techniques may be found, they are more complex to compute and as such are used less often than the simple algebraic method. Generally, these techniques are used to refine the initial linear estimate in an iterative optimisation such that the geometric or similar errors are minimised improving upon the initial algebraic solution.

There are number of other image features that can be used to accomplish the estimation of the homography between views. Rather than using detected corners in the estimation process, imaged lines and conics may be used. Similar DLT methods are employed to estimate the corresponding line and conic transformation matrices which can be reformulated to yield the point transformation homography. The issues involved with estimating transformations in this manner are the same issues that crop up for the point based methods, but additional line and conic estimation errors are also present increasing the overall error. The main purpose for which these techniques are used is in image rectification. Generally, the orientations of lines have to be known, i.e. parallel or orthogonal, or the correct shape of the imaged conics have to be known for the rectification to be possible.

### 4.2.2 Frequency Domain Methods

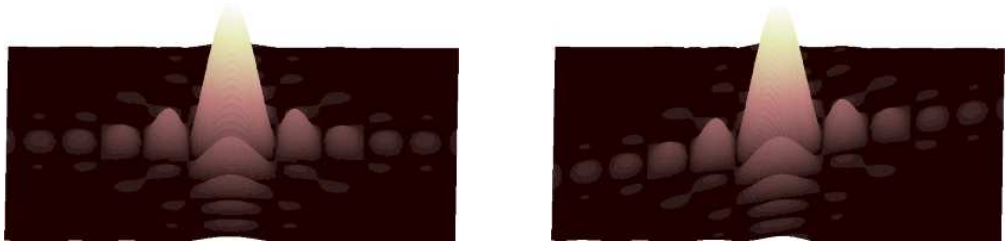
The frequency domain methods are all formulated around the relation between how an image warps in the spatial domain and how its corresponding Fourier transformation warps in the frequency domain Bracewell et al. (1993). It emerged that the relation is quite simple in the case that only affine warping is being considered. The estimation of affine transformations can consequently be accomplished in a featureless framework, removing the dependence of pose estimation techniques on feature detector and point matching algorithm accuracies.

A full derivation of the relationship between the spatial and frequency domain image warping is given Appendix D. The methods with which the frequency domain is used for pose estimation are briefly covered to highlight their applications and limitations. The drawback to using the frequency domain techniques is that, at present, no perspective transformations may be estimated. A discussion on the possibilities of using the frequency domain to estimate perspective transformations in the future is given at the end of this section.

#### The DLT and the Fourier Transform

Taking just the magnitude of the Fourier transformed images, corresponding magnitude peaks may be matched to sub-pixel accuracy across two images, see Fig. 4.1. This requires that the magnitudes of one Fourier image are scaled to that of the other, so that corresponding magnitude peaks will have the same magnitude value. The magnitude plots are easily matched since the magnitudes are related by the determinant of the affine transformation matrix. The determinant and therefore the relative scaling of the Fourier magnitudes may be found through comparing the magnitudes of the Fourier plots at the zero frequency. At this point, the homography between the two sets of Fourier magnitude peaks may be found. This relates directly to the affine transform of the images in the spatial domain through the relation given in 4.1, where $A_{spatial}$ is an affine transformation matrix that operates in the spatial domain, and $A_{freq}$ is an affine transformation matrix that operates in the frequency domain.

$$A_{spatial} = A_{freq}^{-1} \tag{4.1}$$



(a) Original un-warped image (b) The affine-warped image

Fig. 4.1: The Fourier magnitude plots at low frequencies of two related images

There are a few drawbacks and complications with the use of the direct linear transform in the frequency domain. These problems are of comparable significance to those that exist with DLT in the spatial domain. These problems are related to the point selection, and point matching methods which are the same methods that are used in the spatial domain. Perspective planar transformations also cause significant problems which is also dealt with later in the chapter.

The points to be used in the DLT algorithm are the positions of the magnitude peaks of the Fourier transformed images. The peaks of the Fourier transform may be found to a sub-pixel level by fitting paraboloid shape onto the peak region, and finding the location of its maximum. This detection method is akin to finding chessboard corners in the spatial domain using a saddle shape fitting algorithm. The resulting feature detection errors in both domains will be of similar magnitude, primarily determined by the accuracy of the shape fitting algorithm.

For real images of planar scenes there are a number of difficulties that arise when matching detected Fourier magnitudes for two views. Uneven lighting conditions between views, causes an increase in the Fourier magnitudes in one image. Different background regions in both images causes additional different frequency artifacts to appear in their Fourier transformed images. Loss of information because of pixelisation due to scaling or perspective warping causes additional high-frequency artifacts, while removing other valuable information. These effects are difficult to quantify and depend on the native resolution of the camera, the distance to the plane and environmental conditions. These problems result in making correspondence matching across views difficult because the relation between the magnitudes of the Fourier transformed images

is altered from the ideal affine relation.

### Cost functions and the Fourier Transform

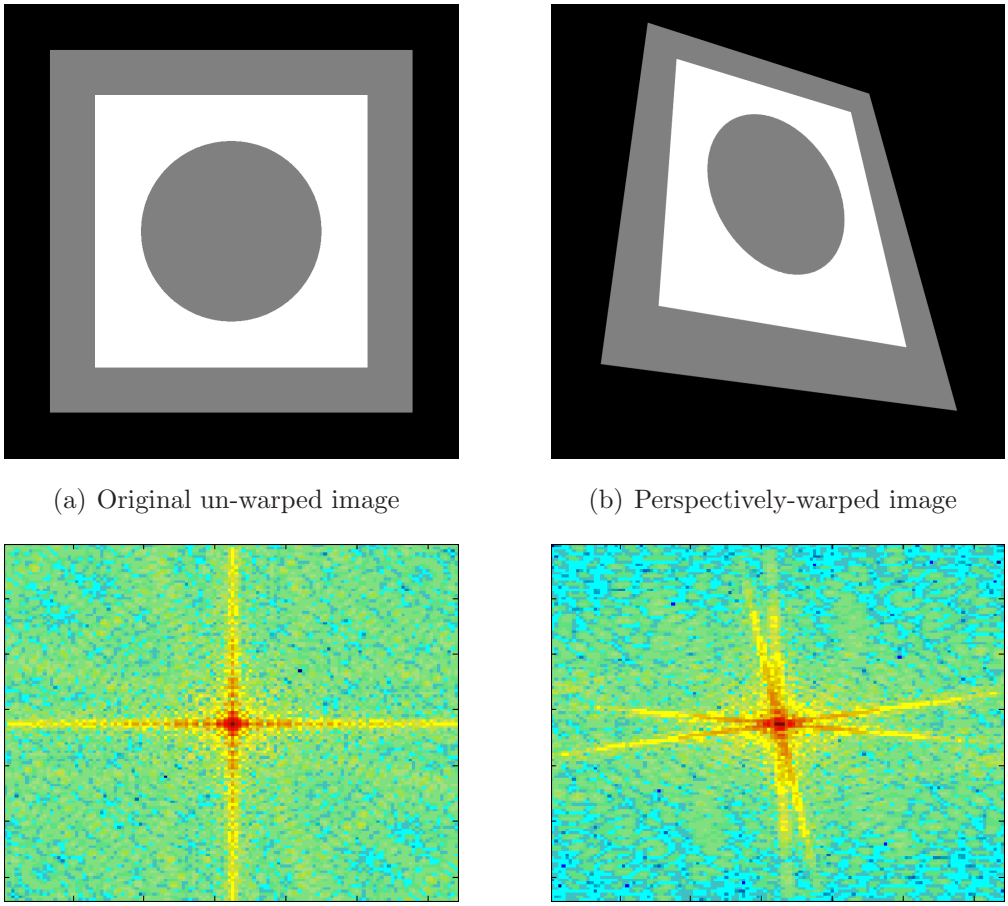
A similar approach used in Lucchese (2000), uses the radial projections of the Fourier transform magnitudes in two images, and minimises the distance between these projections in an iterative optimisation to solve for the transformation. This again is very effective with synthesised data, but tends to fail with real data. One reason for this may be seen in figure 4.2, which highlights that perspectively warped images have extra bands of frequency components compared to those of just affine warped images. For this reason, the two frequency domain images aren't comparable on any level. The Fourier transforms are not directly related through any planar transformation, and so transformation information cannot be extracted from the Fourier transforms of two views related by a perspective transformation.

One solution to this problem was also proposed by Lucchese. He used the initially estimated affine matrix as an approximation to the full perspective case, and solved for the remaining two perspective elements of the homography using a dense matching technique in the spatial domain. This used all of the information available in the images in both the frequency and space domains, keeping with the fundamental idea of this area.

These methods are prone to the same lighting variation and background frequency artifacts that exist for the DLT frequency based methods. A difference in scale between views is still problematic too, since a magnification of an image in the spatial domain will cause a shift of the magnitude peaks towards the origin in the frequency domain. Thus for radial projections, if the scale is too different, not all of the peaks that were included in one image will be included in the other. These along with the problems associated with perspective images make it difficult to use real data.

### Perspective Theorem of the Fourier Transform?

In order to capture all of the degrees of freedom that exist with planar transformations, relative scene depth is required because without it, only in-plane, affine transformations may be estimated. This typically arises through the



(a) Original un-warped image

(b) Perspectively-warped image

(c) Fourier magnitude of un-warped image

(d) Fourier magnitude of warped image

Fig. 4.2: The Fourier magnitude plots of two related images

use of homogeneous coordinates in the spatial domain. However, perspective image relations in the frequency domain provide a more complex challenge in their derivation, since spatially varying depth is lost because of the global Fourier transform that is employed. Because points with different depths in the spatial domain all combine to produce each Fourier magnitude, no information about individual scene depth can be inferred from the magnitude of the frequency domain plot. As such, perspective relations can not be derived in the frequency domain.

Even assuming that perspective relations were derivable in the frequency domain, no correspondences with which to estimate transformations between views could be found. This problem is very evident in the two magnitude plots of perspectively related images in figure 4.2. Extra Fourier magnitude peaks appear in directions other than the two principal directions for the perspectively transformed image's Fourier transform. A set of point correspondences

between the two images' peaks doesn't exist, and so the homography can not be estimated.

Because perspective relations can not be extracted from the frequency domain, affine relation approximations will need to be used. For real scenes there are a couple of additional issues that arise due to this affine approximation assumption. Using two real images of the same object taken from different view points, there are two main problems caused by perspective effects. Firstly, the Fourier transformed image magnitudes are no longer related by the determinant of the affine-warping matrix. Without the ability to normalise each magnitude plot and correctly find peaks of matching amplitude, no correspondences may be found. And secondly, unless a weak perspective assumption can be reasonably made, i.e. for planes imaged from a large distance, the images are no longer related by a simple affine warping, and again frequency correspondence can not be found.

### 4.2.3 Summary of Existing Techniques

Even though it is possible to estimate perspective transformations between images using techniques in the spatial domain, the methods are difficult to apply in certain circumstances. With images, within which features are difficult to accurately extract, such as in images of densely textured scenes or noisy images of scenes, feature matching is a difficult task that can result in erroneous matches. These erroneous matches along with poorly localised features due to noise make estimation algorithms inaccurate. To overcome these issues, featureless pose estimation techniques may be applied, but only to a limited extent.

Featureless rectification techniques are available through the use of the frequency domain. These techniques answered some of the issues that the spatial domain methods showed, bypassing the feature segmentation and matching problems, to make affine planar transformation estimation possible. However, perspective transformations need to be considered, since in general, a weak perspective assumption can only be made when the imaged object is a very large distance from the imaging device. Without this constraint, affine approximations no longer hold true, and the full perspective matrix needs to be determined. This is not directly available through the use of the Fourier

transform.

Ideally, the featureless aspects of the frequency domain and the perspective transformation estimation abilities of the spatial domain would be combined to achieve featureless perspective transformation estimation. However as it is not possible to examine how frequency components are locally changing using the global Fourier transform, a localised equivalent of the Fourier transform would give the information that is required. This would suggest examining the problem in a new domain.

## 4.3 Space-Frequency Domain Planar Transformation Estimation

Because of the limitations of both the spatial domain and frequency domain techniques, an alternative domain is sought. With the spatial domain methods, feature detection and matching across views may be difficult, and with the frequency domain techniques, because of their global transform nature, important spatial information is lost. Ideally, the frequency components at each position in the image, from which pose may be determined, will be found and used. There are a number of different methods that facilitate this, overcoming the hindrances of the other two domains.

To start off with, an obvious choice for obtaining localised frequency information from images is to compute Fourier transforms for smaller regions of the images in question. These smaller Fourier transforms can be calculated at each position in each image yielding localised frequency information. These are known as short-time Fourier transforms (STFTs), a name obtained from their corresponding use in time varying signal analysis. There is one significant issue that exists with using STFTs. Due to the global nature of Fourier transforms, over the image patch, a trade-off between spatial resolution and frequency resolution needs to be made. For a high frequency resolution, the Fourier transform has to be computed over a larger region of the image, yielding a common frequency response for that entire region of the image. If a smaller region is selected, the frequency response is common only to that smaller region resulting in higher spatial resolution, but the frequency resolution suffers as a consequence. Prior information about the image type and structure allows for

the intelligent selection of image region size to optimise the resolutions in both domains for a given task. However, this is too restrictive for general purpose use.

Similar approaches can be used with variations on the global transform that is used, such as sine and cosine transforms. Each of these methods are limited by the same requirement that a trade-off between spatial and frequency resolutions has to be made, as with the short time Fourier transforms. An additional constraint on the approach to be taken can be added because of this. Because the technique involved should be capable of handling perspective pose, in a featureless framework, ideally an infinite range of frequency and spatial resolutions would be available simultaneously. This will guarantee that for every position in the image, any frequency component can be examined.

One representation that provides this limitless resolution in the spatial and frequency domains is through the use of the Continuous Wavelet Transform. Previously, wavelets have been successfully used for estimating shape from texture, and in particular in Clerc and Mallat (1999), it was shown that wavelets can be used for finding surface normals for curved surfaces. There is not yet any formulated approach to using wavelet transforms for rectifying planar images.

### 4.3.1 Continuous Wavelet Transform

Wavelets capture both the textural information, and the geometric feature information of images through a series of "frequency transforms" across an image. These frequency transforms come about through convolving a wave shape with the image. The one-dimensional wavelet transform was first developed by Grossmann and Morlet (1984) as a means of overcoming the shortcomings of the short-time Fourier transform (STFT), with particular application in their case to seismic analysis. Wavelets have subsequently found application in areas such as image and speech compression, and to a large extent in areas such as face recognition, texture classification, and image segmentation. They haven't yet been exploited in rectification problems. The next sections will show how wavelet transformations may be used for the purposes of pose normalisation.

Equation 4.2 gives the basic representation of the one-dimensional continuous

wavelet transform, which will form the basis of the rectification techniques discussed in this chapter.

$$W(a,b) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{|a|}} \psi^* \left( \frac{x-b}{a} \right) dx \tag{4.2}$$

where $W(a,b)$ is the wavelet decomposition of $1D$ function $f(x)$ at scale $a$ and position $b$, and the $*$ operator indicates the complex conjugate of $\psi\left(\frac{x-b}{a}\right)$, the mother wavelet function scaled by $a$ and translated to position $b$. From this expression it can be noted that the representation is very similar to a one dimensional Fourier transform. Unlike the Fourier transform, the wavelet transform takes integrals for each frequency *and* for different positions. Thus a two dimensional output is achieved, the frequency decomposition of the signal at different positions along the object function.

To simplify the expression, the following substitution for the wavelet function is used:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}} \psi \left( \frac{x-b}{a} \right) \tag{4.3}$$

which results in the expression:

$$W(a,b) = \int_{-\infty}^{\infty} f(x) \psi^*_{a,b}(x)\, dx \tag{4.4}$$

Expressions for the two-dimensional Continuous Wavelet Transformation of a 2D signal exist which are of the same form as that of the 1D wavelet decomposition. In the 2D case, the transformation is parameterised by three or four values, three values when two position and one scale coefficient are required, and four values when two position and two scale values are required. The use of two scale values allows the shape of the wavelet to be changed to highlight certain aspects in the decomposition. For the purposes of this thesis, only standard wavelet shapes with a single scale parameter are used, the expression for which is given in 4.5.

$$W(a,b,c) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \frac{1}{\sqrt{|a|}} \psi^* \left( \frac{x-b}{a}, \frac{y-c}{a} \right) dx\, dy \tag{4.5}$$

To show the resulting effects of a perspective transformation of an input image to the wavelet decomposition, one dimensional wavelet decompositions shall be used to highlight the mathematics.

### 4.3.2 Wavelet Transform of Perspective Images

This section will show the effect of perspectively transforming an image prior to wavelet transformation on the frequency coefficients of the resulting wavelet transformation. For ease of notation, the perspective effect on a one dimensional signal will be examined.

#### Perspective Transform of a planar image

From Equation B.6 a projective transform was expressed in terms of a $3 \times 3$ matrix, shown again below for clarity. The first row determines the amount of transformation of the $x$ ordinate, the second row, the $y$ ordinate and the third row accounts for the perspective warping.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4.6}$$

In inhomogeneous image coordinates, the $x$ and $y$ coordinates are normalised by the $z$ ordinate.

$$x' = \frac{ax + by + c}{gx + hy + i} \tag{4.7}$$

$$y' = \frac{dx + ey + f}{gx + hy + i} \tag{4.8}$$

Operating under a reduced perspective assumption where approximations to only yaw and pitch rotations are considered, an estimate of the perspective portion of the full homography is required. To facilitate this approximation, the affine parameters are chosen to be $I_{2\times2}$, and since the planes are assumed to be infinite no translation parameters are required. This leaves a reduced representation for the inhomogeneous coordinates as follows:

$$x' = \frac{x}{gx + hy + i} \tag{4.9}$$

$$y' = \frac{y}{gx + hy + i} \tag{4.10}$$

The inhomogeneous coordinates are still dependent on the other coordinate in the projective part of the transformation, but the effects of just the $x$ perspective parameter alone may be seen by examining a slice of the image where $y = 0$. The image will only expand and contract in one direction, along the

$x$ axis. Based on this assumption a more simple algorithm can be utilised for the purposes of image rectification.

$$x' = \frac{x}{gx + i} \tag{4.11}$$

Inserting the expression for the transformed $x$ into the wavelet transform equation, we get the following:

$$W(a, b) = \int_{-\infty}^{\infty} f(x')\psi^*_{a,b}(x)\, dx \tag{4.12}$$

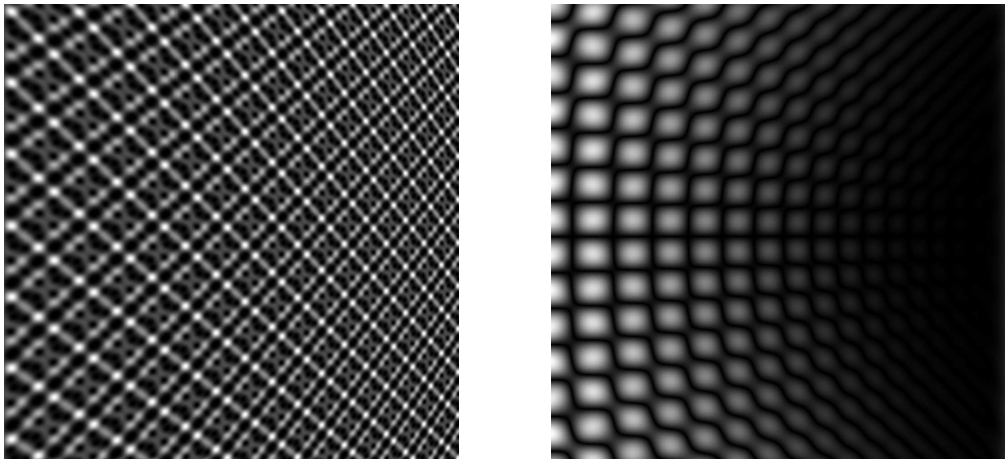
Inserting the expression for $x'$ yields:

$$W(a, b) = \int_{-\infty}^{\infty} f\left(\frac{x}{gx + 1}\right)\psi^*_{a,b}(x)\, dx \tag{4.13}$$

Typically wavelet decompositions are computed using filtering techniques copied from signal analysis research. As such, depending on the function to be examined and the wavelet function being used, mathematical derivations may be difficult to analytically compute or are unavailable.

To highlight the application benefits of the CWT a simple example is described. A high response to a large scale wavelet would indicate the presence of low frequency information, and if this response changed across the image of a plane, it could be said that the left of the plane was closer to the camera than the right, see Fig 4.3. The opposite effect will be noted when examining the high frequency data, one side of the image will have a lower response to a high frequency filter than the other. From this simple description alone, the usefulness of the wavelet transform is quite clear, it encapsulates the change in frequency components across an image.

This is a property that can be easily used to rectify images. In the simplest case, the energies at different frequencies within an image can be balanced such that all of the frequency component magnitudes are equal on each side of the image. This will require the symmetry measure, defined in Section 2.2, to be maximised, in an iterative scheme. Of course, the relative orientations of two views can also be registered through the optimisation of a cost function that will ensure that frequency components are equal at the same positions in each of the two images. The transformation that achieves this can then be saved, and is the homography that relates the two views.



(a) Perspectively transformed textured plane (b) The wavelet coefficients

Fig. 4.3: The wavelet decomposition of a perspectively transformed image

### 4.3.3 Estimating Homographies

We will use the CWT of perspectively imaged planes to find a solution for the $x$ and $y$ perspective warping effects. To do this we need to find the inverse warping. So we are required to find the $g'$ and $h'$ that will map the $x'$s and $y'$s back onto their corresponding $x$s and $y$s.

$$x = \frac{x'}{g'x' + h'y' + 1} \tag{4.14}$$

$$y = \frac{y'}{g'x' + h'y' + 1} \tag{4.15}$$

The WBSC cost function of Section 2.2 was adapted slightly so that the optimum symmetry value of the wavelet transformed image is achieved when central symmetry is restored. The theory behind this is that the frequency coefficient values across the imaged textured plane should be similar in a fronto-parallel view, and so the symmetry coefficient can be re-designed to ensure this.

Once again, the symmetric and antisymmetric components of the image are examined and the proportion of the signal that is symmetric is used as the continuous symmetry measure. The symmetric and antisymmetric components are calculated based on central symmetry, the expressions for which are given in Equations 4.16 and 4.17, where $W(s, x, y)$ is the wavelet decomposition of

the image at position $(x, y)$, using a wavelet of scale $s$.

$$W_{cen\_sym}(s, x, y) = (W(s, x, y) + W(s, -x, -y))/2 \tag{4.16}$$

$$W_{cen\_asym}(s, x, y) = (W(s, x, y) - W(s, -x, -y))/2 \tag{4.17}$$

The same asymmetry cost function is minimised based on these new definitions of symmetric and antisymmetric components, which directly equates to obtaining the optimum symmetry level, hence restoring the pose of the imaged plane to a fronto-parallel view. The equation that describes the asymmetry value at each position in the image, the total of which must be minimised is given in Equation 4.19

$$AS\{W(s, x, y)\} = \frac{\|W_{cen\_asym}(s, x, y)\|^2}{\|W_{cen\_sym}(s, x, y)\|^2 + \|W_{cen\_asym}(s, x, y)\|^2} \tag{4.18}$$

The actual cost function to be minimised is the sum of the asymmetry values over the image within the windowed region, and over all of the wavelet scales being used.

$$Cost = \sum_{s=m}^{s=n} \sum_{x} \sum_{y} AS\{W(s, x, y)\} \tag{4.19}$$

$$= \sum_{s=m}^{s=n} \sum_{x} \sum_{y} \left( \frac{\|W_{cen\_asym}(s, x, y)\|^2}{\|W_{cen\_sym}(s, x, y)\|^2 + \|W_{cen\_asym}(s, x, y)\|^2} \right) \tag{4.20}$$

### The Algorithm

The algorithm involved is similar to that used for removing pose from face images in Chapter 3. In an optimisation process, the following steps will be iteratively carried out, varying the $(g', h')$ perspective transformation coefficient pair, until a minimum cost is converged upon. An initial estimate for $(g', h')$ is selected as $(0, 0)$.

(i) The input image will be perspectively transformed by some $(g', h')$ coefficient pair placed in a planar transformation matrix.

(ii) The wavelet transformation of this newly generated image will be calculated at a number of different scales.

(iii) The symmetry cost to be minimised will be calculated on this wavelet transformed image.

(iv) Small increments are applied to $g'$ and $h'$ terms individually.

(v) Steps (i)-(iii) are performed with the new values of $g'$ and $h'$.

(vi) The slope of the cost function with respect to $g'$ and $h'$ will be internally calculated.

(vii) Steps (iv)-(vi) are continually carried out and the values of $g'$ and $h'$ will be continually updated such that the cost is decreasing using a gradient descent method.

(viii) When the cost can no longer decrease, the final values of $g'$ and $h'$ are stored.

The symmetry cost function is calculated for a windowed region of the image. This is to prevent any introduced edges, resulting from the perspective transformation, from being included in the frequency component analysis. Using the iterative optimisation algorithm, the values of $g'$ and $h'$ that ensure the optimum central symmetry in the wavelet transformed images are found, providing a rectified image. To increase efficiency, only a small number of wavelet scales are chosen for the range $m$ to $n$. The Energy Balancing Planar Rectification (EBPR) algorithm can be summed up with the expression given in Equation 4.21.

$$(g', h') = \arg \min_{(g',h')} \sum_{s=m}^{s=n} \sum_{x} \sum_{y} AS\{W(s, x, y)'\} \tag{4.21}$$

where $W(s, x, y)'$ is the wavelet decomposition of the input image, perspectively unwarped by the $(g', h')$ perspective coefficients.

## 4.4 Experiments

Three experiments are presented. The first two experiments compare the proposed rectification method with the technique proposed in Lucchese (2001b). These two experiments operate on a reduced perspective matrix where only yaw rotations are allowed. The first experiment operates on the ideal case of a set of simulated symmetric images. The second experiment uses real non-symmetric images of planar textures captured from varying degrees of rotation in the comparison. The third experiment operates on simulated homogeneous

Table 4.1: Ratio of g/h for varying roll and pitch, with yaw = 20°

| Angle ° | Roll | Pitch | Roll = Pitch | Pitch = 0.5° Roll varying |
|---|---|---|---|---|
| 0 | 7.63e+14 | 7.63e+14 | 7.63e+14 | 3.91e+01 |
| 1 | 5.72e+01 | 1.95e+01 | 1.45e+01 | 1.24e+02 |
| 2 | 2.86e+01 | 9.79e+00 | 7.27e+00 | 1.06e+02 |
| 3 | 1.90e+01 | 6.52e+00 | 4.82e+00 | 3.72e+01 |
| 4 | 1.43e+01 | 4.89e+00 | 3.59e+00 | 2.25e+01 |
| 5 | 1.14e+01 | 3.90e+00 | 2.84e+00 | 1.61e+01 |

textured planes. The series of textured planes, that are synthetically rotated about their yaw and pitch axes, are rectified using the proposed technique. This involves solving for the two perspective transformation parameters.

### 4.4.1 1 Degree of Freedom - Yaw Rotations

In Lucchese (2001b) it is shown that for small out of plane rotations, an affine approximation will suffice for image rectification. However in cases where out of plane rotations are larger, a perspective only approximation is more accurate. These experiments will demonstrate the principle. Furthermore, it will be shown that for pure single-axis, out of plane rotations, a reduced transformation model may be used.

Camera motions that are pure rotations about the yaw axis, generate homographies where the coefficient $h$ will always be exactly zero. Also, as illustrated in Table 4.1, typically the coefficient $g$ is an order of magnitude greater than $h$ for small in-plane rotations. This indicates that for cases where the reduced perspective assumption is not precisely met, values for $h$ are very small. Using this assumption the transformation that will be solved for inhomogeneous coordinates may be reduced to:

$$x' = \frac{x}{gx + i} \tag{4.22}$$

$$y' = \frac{y}{gx + i} \tag{4.23}$$

### Simulated Images

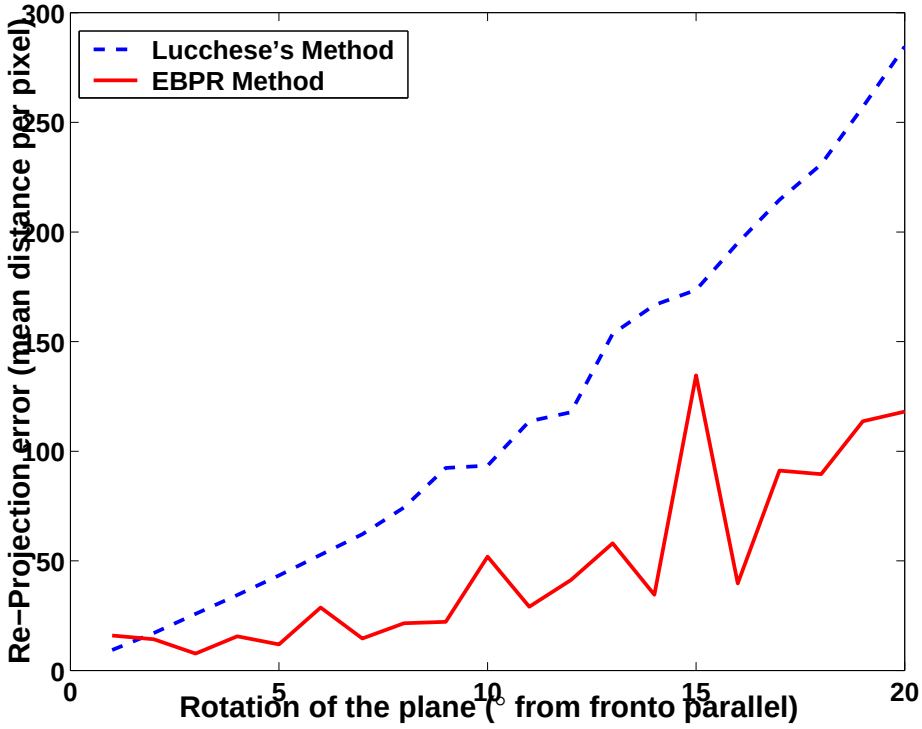
The algorithm is tested using simulated rotated planar images at varying degrees of rotation. Images of various scenes are reflected along their vertical axes to create symmetric images, the ideal scenario for single image rectification. Gaussian noise with a standard deviation of 3 grayscales is added to each image for each rectification to simulate real imaging conditions. These images are synthetically transformed into rotated positions using homographies with only the $g$ coefficient of perspective warping in the $x$ direction deviating from the identity matrix.

Each image is then rectified using both the comparison method and the proposed method. In the case of the comparison method, an image of the fronto-parallel plane is used as a target to transform the test images to, and each of the images is windowed using a Kaiser window to prevent edge effects showing up in the Fourier transform as a frequency leakage (Stearns and Hush, 1990). At each angle of rotation, from 1 to 20 degrees, 20 trials of each rectification are carried out to obtain a fair average error for each orientation. The error is computed as the angle between the normal to the rectified plane and the normal to the same plane in canonical position. The average error at each angle is computed. These results are graphed in Figure 4.4(a). Some examples of rectified images are shown in Figure 4.5.

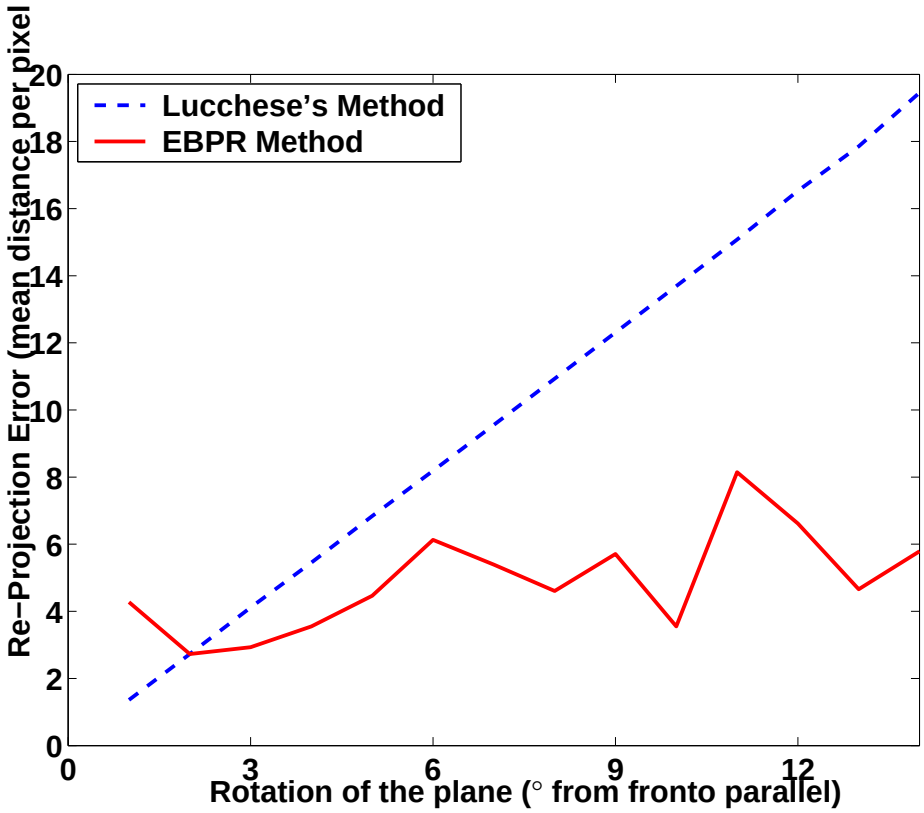
As it can be seen from Fig. 4.4(a), the EBPR algorithm performs better than the comparison method for the majority of the presented angles. The mean distance error per pixel is small relative to the comparison technique. As expected, the comparison method outperformed the EBPR method for angles where a weak perspective assumption holds. For the larger angles of rotation, the proposed wavelet-based method demonstrate superior performance compared to the comparison technique. From the example images in Fig. 4.5, it can be seen that very accurate rectifications are accomplished with the proposed technique.

### Real Images

This test uses real images of almost-homogeneous textured planar surfaces captured at various rotations from the fronto-parallel position. The images




(a)




(b)

Fig. 4.4: The error in re-projection using the two approximations to the homography operating on simulated and real images. (a) Errors for simulated data. (b) Errors for real data. For larger out of plane rotations, the EBPR method performs better.

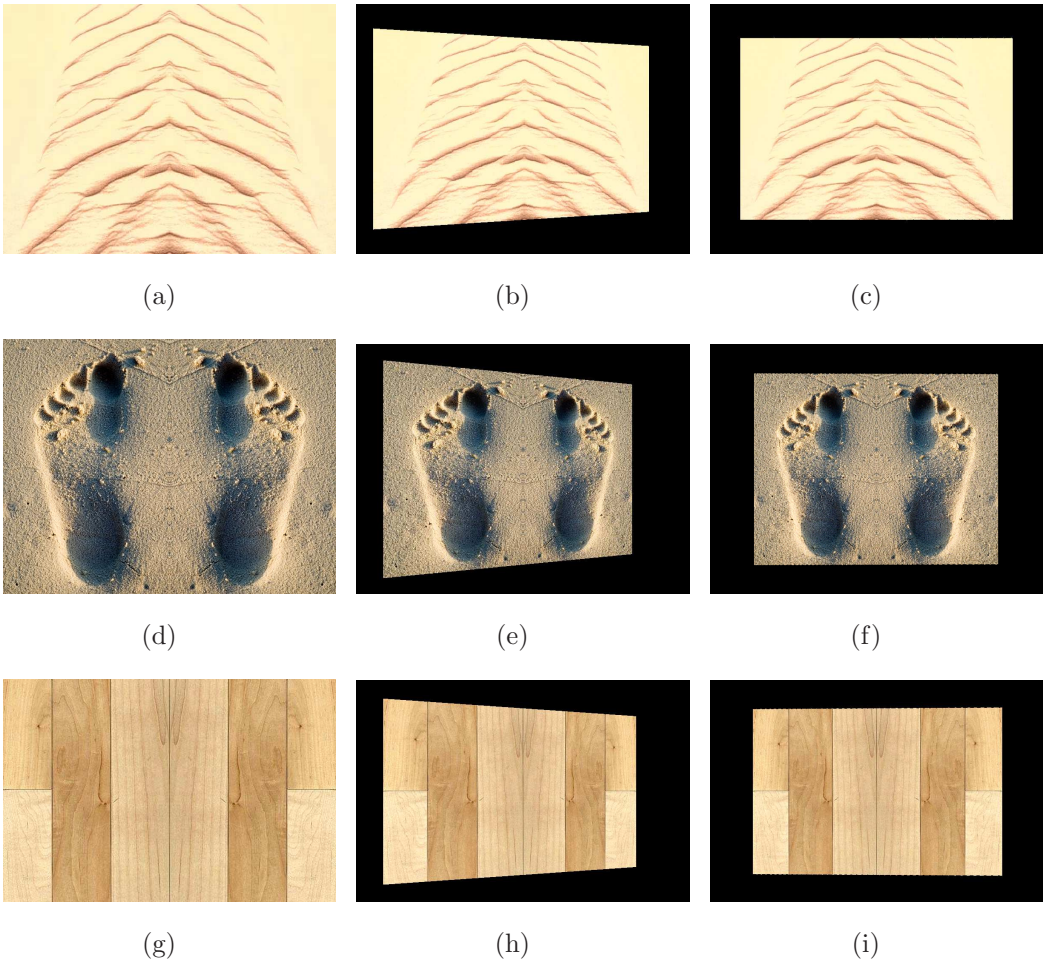
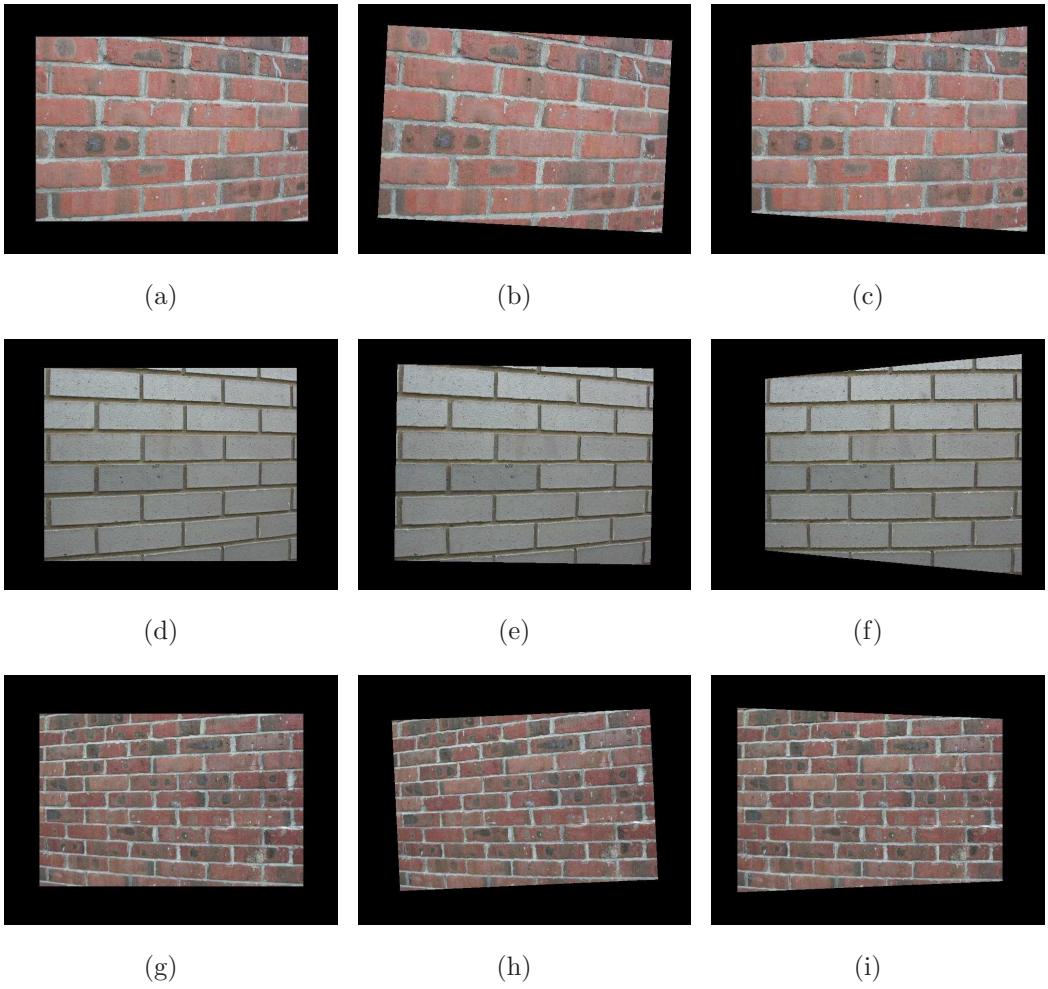


Fig. 4.5: First column shows the synthetically generated symmetric images, second column shows the synthetic images rotated along their vertical axes with noise added, and third column shows the rectified images using the proposed method

are rectified using both the EBPR method and that proposed in Lucchese (2001c), and an error is computed as the distance between the back-projected points and the same points in a fronto-parallel view. The images are again windowed using a Kaiser window to enable the Fourier transform method to perform optimally. Examples of the rectified planar images obtained using the two methods are given in figure 4.6. A graph of the reprojection errors is given in Fig. 4.4(b).

As it can be seen from Fig. 4.4(b), the EBPR algorithm performs better than the comparison method for the majority of the presented angles. Once again, at lower rotation angles, the weak perspective approximation holds and the comparison method performs better. For larger angles of rotation up to 15



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4.6: The first column shows images taken with rotations of 14, 13 and -16 degrees respectively. The second column shows rectified images with the comparison method, and the third column shows rectified images using the proposed method.

degrees, the EBPR method is superior. With rotations that are larger than 15 degrees, the EBPR algorithm breaks down and yields errors of similar magnitude to the comparison technique's errors.

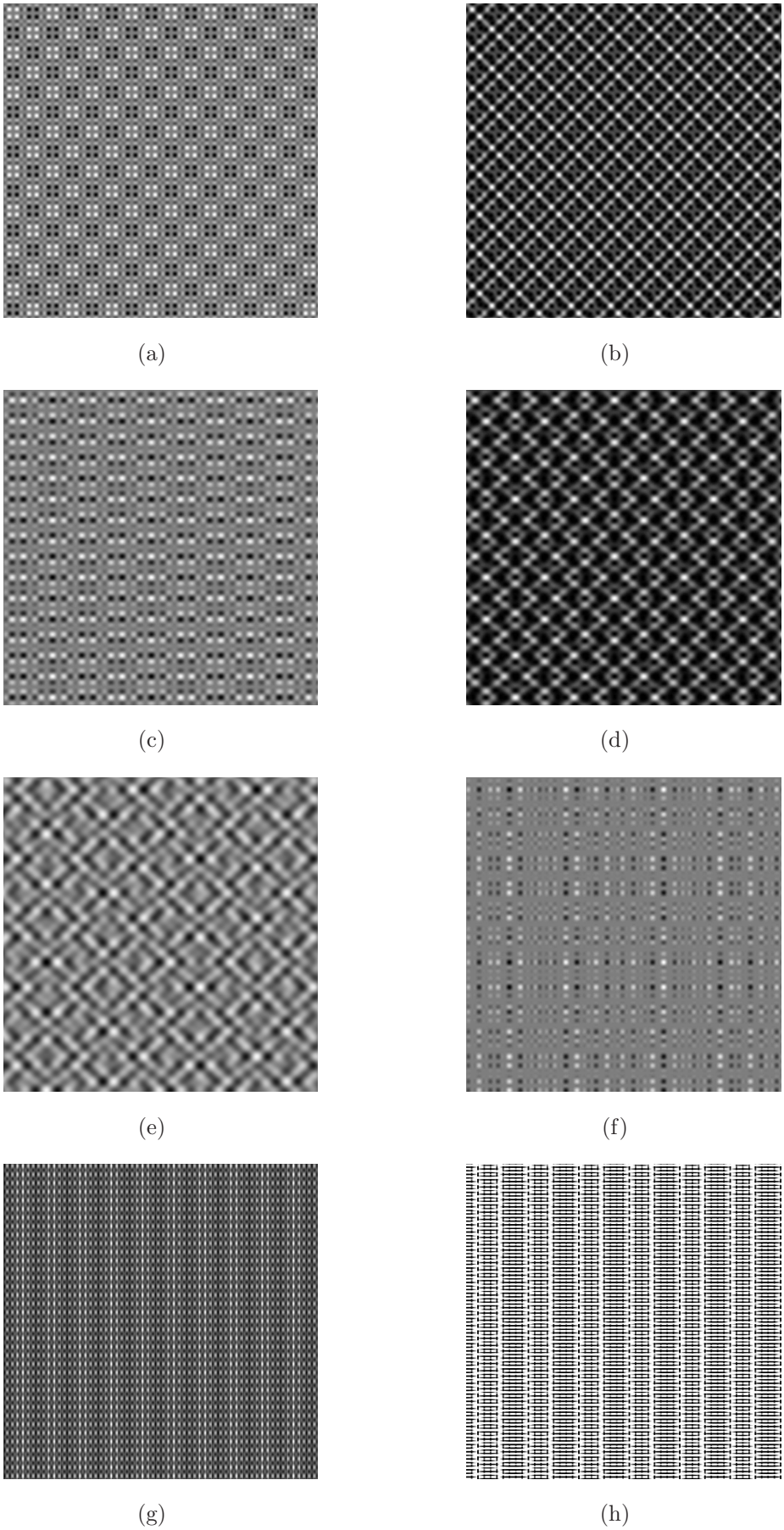
It should be noted that for angles up to 14 degrees, the re-projection error is less than 8 pixels, resulting in very realistic rectifications. This is evident in the provided sample images in Fig. 4.6. The images obtained with the proposed rectification appear to have the pose removed, while the images obtained using the comparison method show additional warping, resulting in little improvement or a degradation to the image's appearance. These factors indicate that the EBPR method is more suited to removing perspective pose from imaged planes.

### 4.4.2 Synthetic Images - 2 Degrees of Freedom

In order to fully demonstrate the abilities of the proposed rectification algorithm, perspective transformations about two axes, namely the yaw and pitch axes, will have to be simultaneously removed. Some initial assumptions on the structure of the images will have to be made. Firstly, each image is assumed to contain a plane, homogeneously textured with a pattern. Each textured plane is also assumed to be infinite, this is to remove any introduced edge effects that may occur through perspective transformations. As such, only the two perspective parameters in each homography matrix will need to be estimated. A database of 25 images portraying homogeneous textures of different sorts was created, some examples of which are shown in Fig. 4.7. Each texture is treated as a plane, with planar transformations being applied to them to simulate rotation about the yaw and pitch axes. Each textured plane is transformed to 100 different orientations, all combinations of 10 values for each of the yaw and pitch perspective parameters. Each transformed plane is rectified using the EBPR method. The error in each estimated perspective parameter is easily calculated as the sum of the known forward transformation parameters and the estimated backward transformation parameters. If an exact match is estimated, the sum of the coefficients will be zero. The mean absolute error for each position is calculated across all 25 images. This error is calculated at each of the 100 possible positions. These mean absolute errors for each estimated transformation are given in Fig.4.8-Fig.4.17. The green dashed line in each graph represents the fundamental limit of the presented algorithm. This limit exists due to the algorithm being unable to resolve differences between wavelet energies at different positions in the image for almost fronto-parallel images. No rectification that could achieve an error below this level was accomplished.

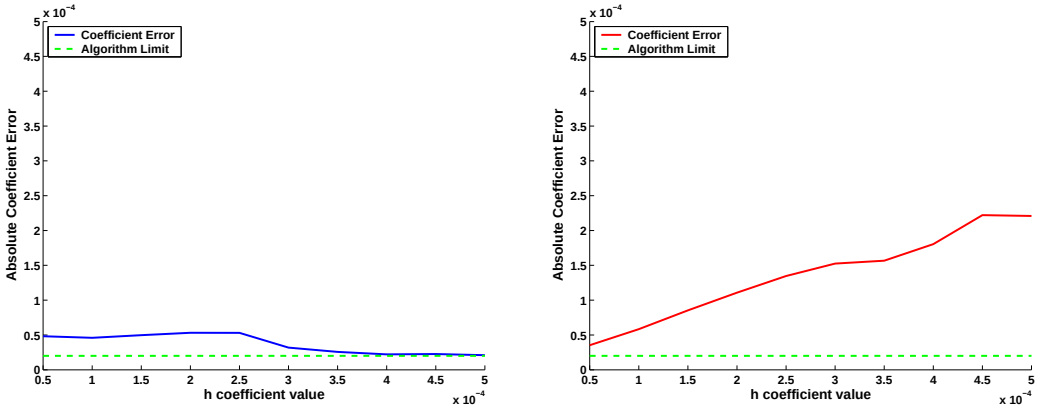
The results are presented in terms of the mean absolute error in the estimated coefficients across the image set. The overall mean errors in the estimated coefficients are $0.8\times10^{-4}$ and $1.3\times10^{-4}$, for the $g$ and $h$ coefficients respectively. These are of a similar magnitude to the smallest of the input transformations to be removed. This indicates that a good rectification is possible with the EBPR technique.

The experimental results highlight the ability of the EBPR algorithm to remove pose from perspectively imaged textured planes, where both the spatial



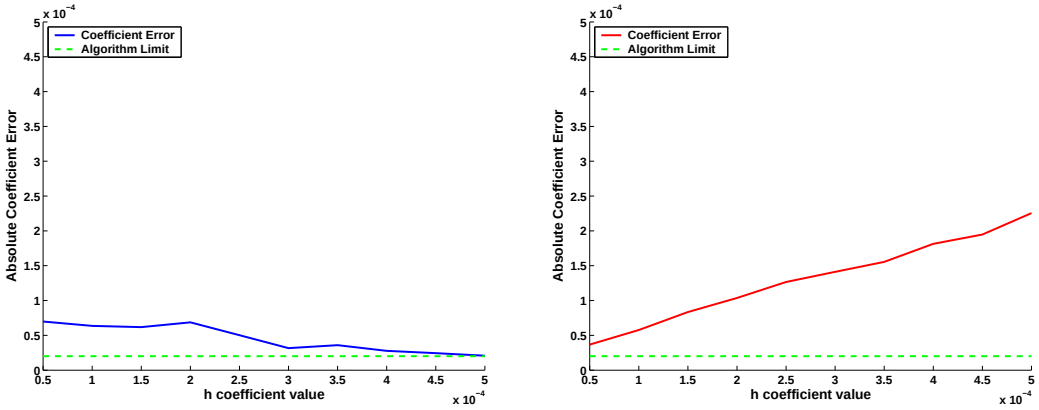
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4.7: A selection of synthetic textures used for experimental validation.



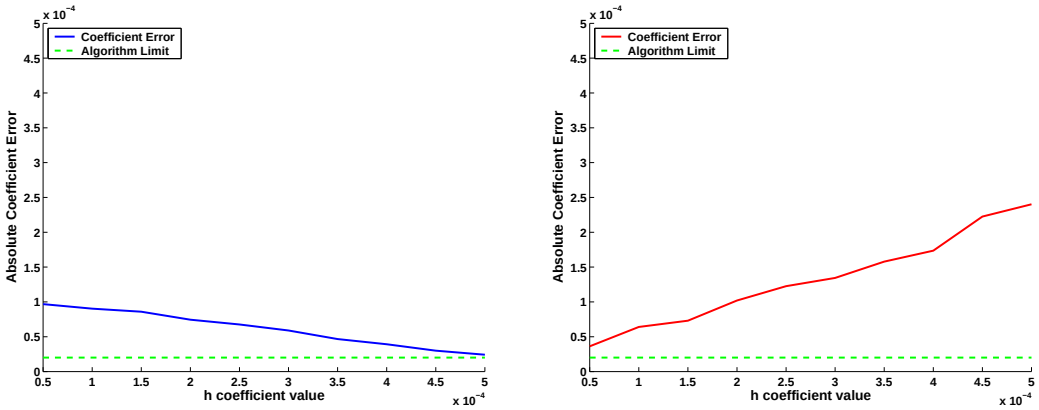

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.8: Error in estimated $g'$ and $h'$ terms with g coefficient at $0.5e-4$



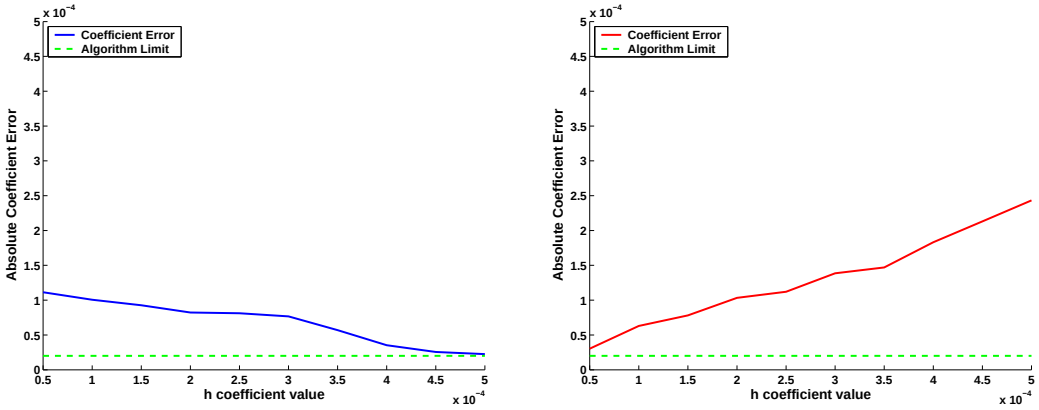

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.9: Error in estimated $g'$ and $h'$ terms with g coefficient at $1e-4$



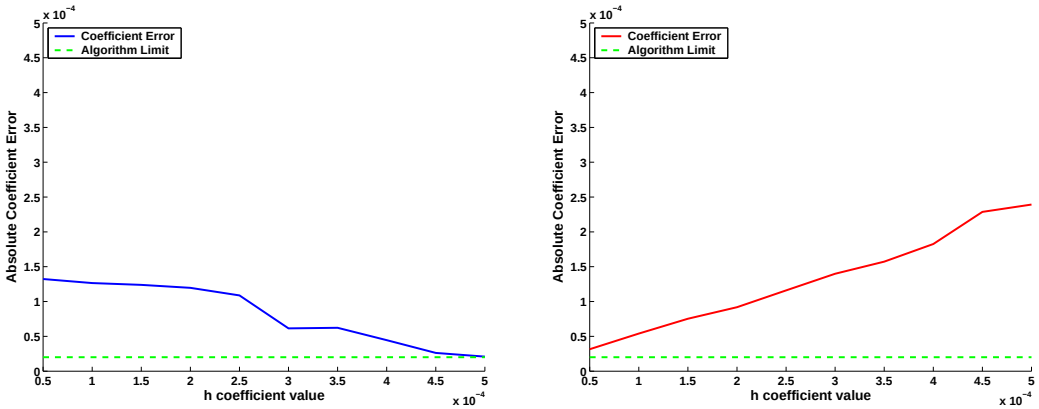

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.10: Error in estimated $g'$ and $h'$ terms with g coefficient at $1.5e-4$



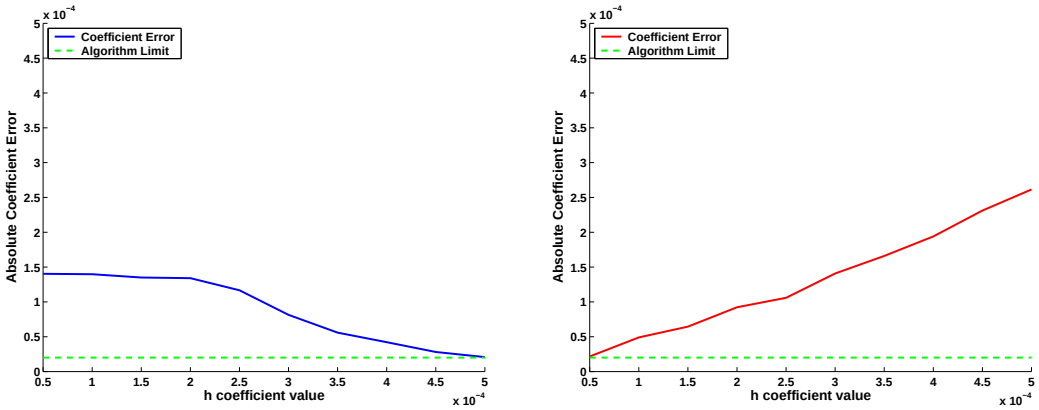

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.11: Error in estimated $g'$ and $h'$ terms with g coefficient at $2e-4$



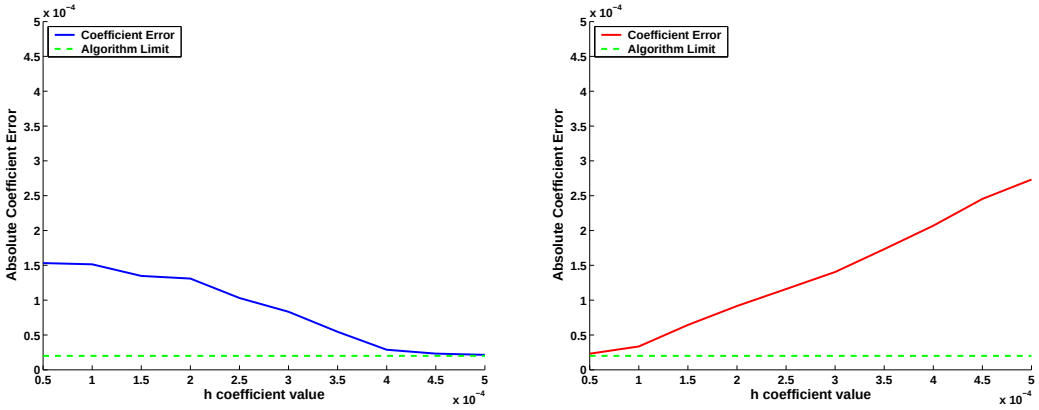

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.12: Error in estimated $g'$ and $h'$ terms with g coefficient at $2.5e-4$



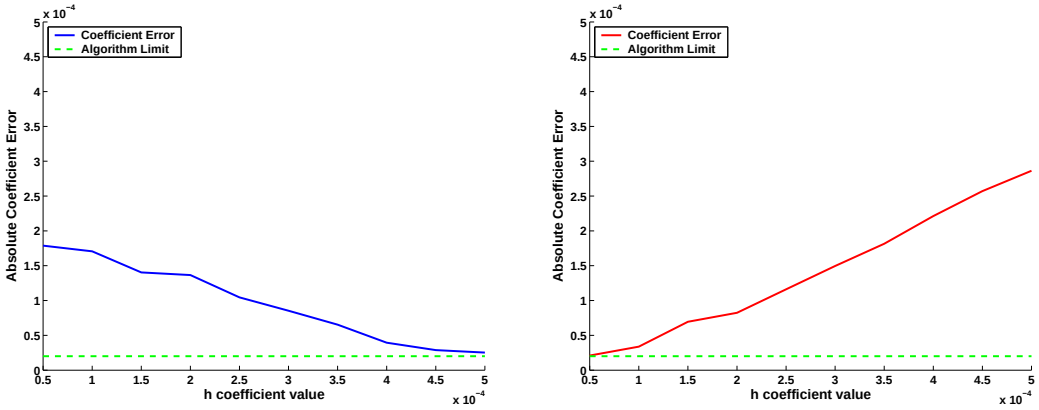

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.13: Error in estimated $g'$ and $h'$ terms with g coefficient at $3e-4$



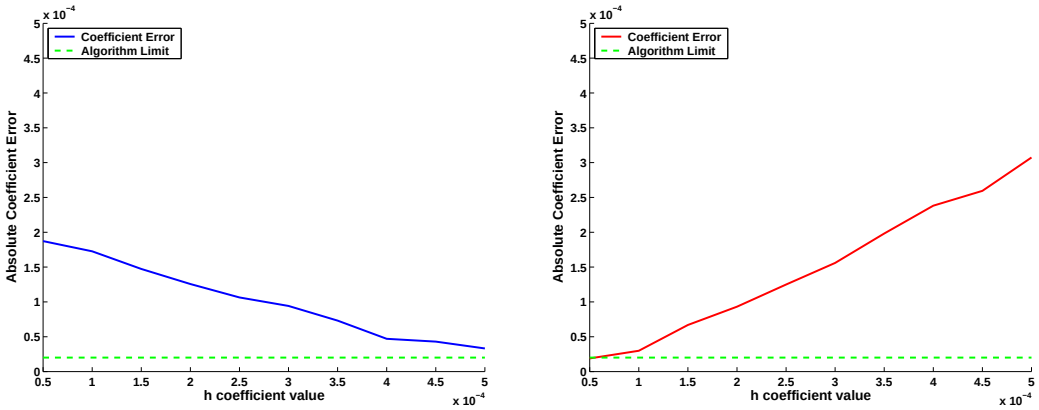

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.14: Error in estimated $g'$ and $h'$ terms with g coefficient at $3.5e-4$



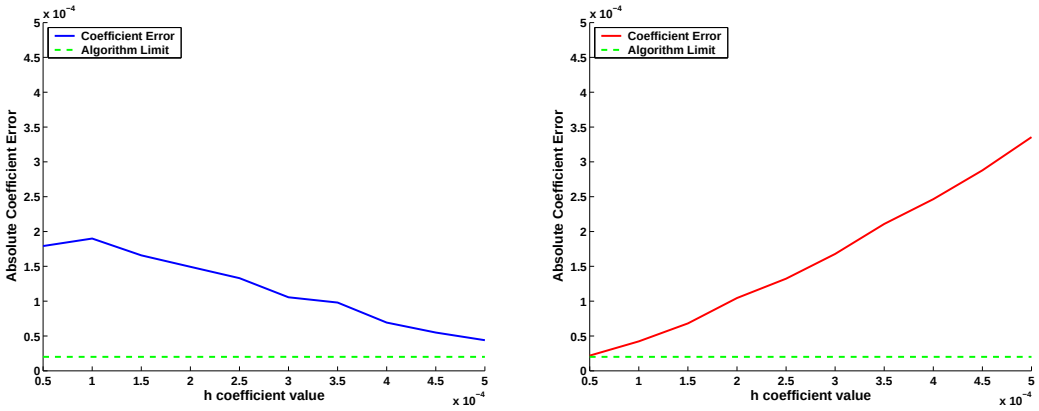

(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.15: Error in estimated $g'$ and $h'$ terms with g coefficient at $4e-4$




(a) Error in estimated $g'$ (b) Error in estimated $h'$s

Fig. 4.16: Error in estimated $g'$ and $h'$ terms with g coefficient at $4.5e-4$




(a) Error in estimated $g'$

(b) Error in estimated $h'$s

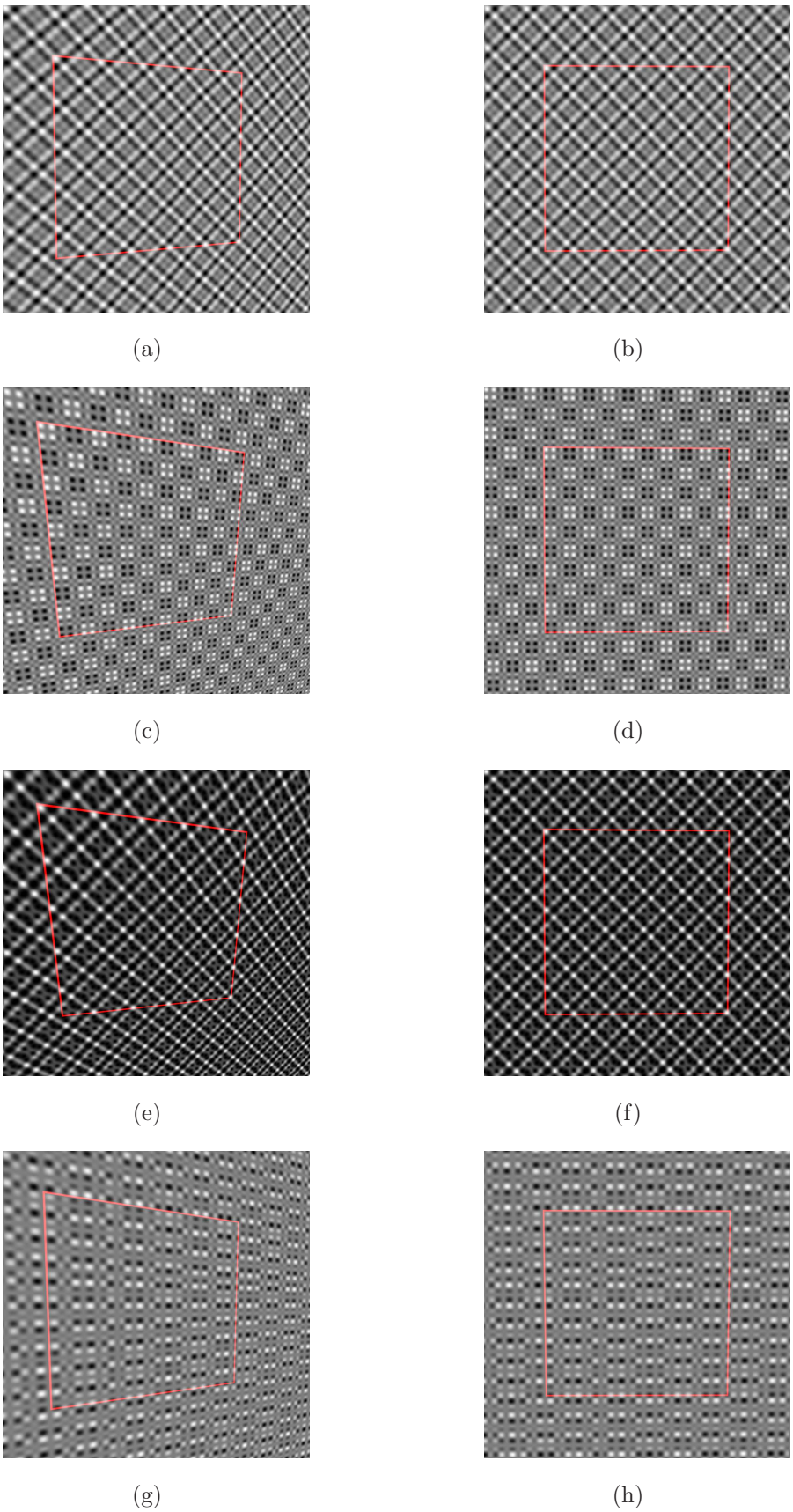
Fig. 4.17: Error in estimated $g'$ and $h'$ terms with g coefficient at $5e-4$

domain techniques and the frequency-based, featureless techniques would fail. The use of the space-frequency domain allowed the removal of pose about two axes, yaw and pitch, in a featureless framework, and it is shown that the resulting estimated transformations accurately remove the majority of the overall transformation. Some example input and output images are displayed in Fig. 4.18 and Fig. 4.19. In each of the images displayed, to aid with visualising the transformation of the plane before and after rectification, a red square was superimposed on the plane. This square was not present in the images during the estimation of the transformations.

## 4.5 Conclusion

This chapter presented a novel approach to planar transformation estimation. An introduction into the estimation of planar transformations in both the spatial and frequency domains was given. The limitations of each were identified through mathematical derivations and discussions of their usage. The need for a new domain within which to rectify images in a featureless framework was identified. This domain is the space-frequency domain, which in the author's case was obtained through a Continuous Wavelet Transformation. This domain incorporates the better aspects of both the spatial domain and the frequency domain in that perspective transformations may be removed using featureless methods.

An algorithm was designed such that the frequency components at positions



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4.18: Results obtained using the proposed rectification algorithm. The first column shows the input images. The second column shows the output images.

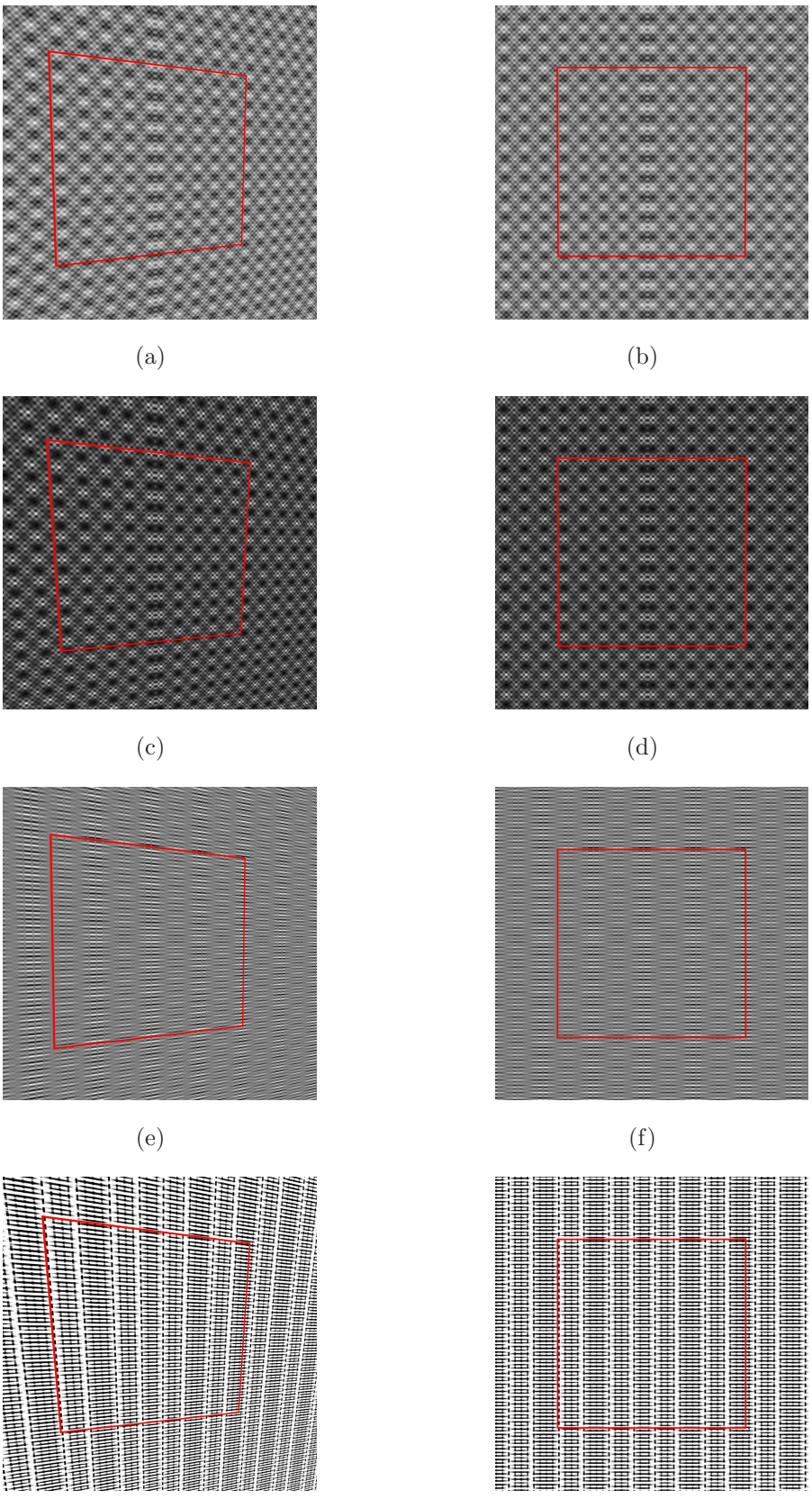


Fig. 4.19: Results obtained using the proposed rectification algorithm. The first column shows the input images. The second column shows the output images.

centrally symmetric to each other would be balanced to restore overall central symmetry in the wavelet transformed image. This came about through an iterative optimisation process using the wavelet based symmetry cost that was developed in Section 2.2. A modification to this cost ensured that central symmetry rather than axial symmetry was optimised.

Perspective transformations were then estimated in a featureless framework. At present, no other technique operates in a featureless regime to estimate perspective planar transformations. The author's EBPR technique was shown to be capable of estimating perspective transformations. The experiments, demonstrate that the proposed approximation and its associated rectification algorithm provide a more realistic approximation to a full perspective transformation than the comparison method. This was highlighted through constrained, out-of-plane motion of textured planar scenes, where weak perspective assumptions break down and the comparison method yields less accurate rectifications.

It was also shown that for two axes rotations, the error in rectification is very small. The resulting errors in the rectification matrix are small relative to the input transformations. Certain input transformations with small homography coefficients, cause the algorithm to fail resulting in errors similar in magnitude to the input coefficients. This is because, for small motions, the algorithm is unable to resolve differences in wavelet responses at different positions in the image because the planes are almost in fronto-parallel positions. For all other motions, the transformations can be removed up to this fundamental limit which was demonstrated in the sample of images shown.

The space frequency domain was shown to be accurate and efficient in removing perspective pose from planar textures where other methods may fail. Spatial domain techniques would not be capable of finding and matching strong features in the images used, and the frequency domain was shown to be unable to cope with perspective distortions, save for a few small values of rotation where weak perspective may still be assumed. In all other cases, the proposed EBPR procedure was shown to be superior, confirming its valid use.

# Chapter 5

# Conclusions and Future Work

## 5.1 Introduction

This thesis directly dealt with the problem of removing pose from face images in order to facilitate improved results in facial expression removal, currently being researched by the Computer Vision and Imaging Laboratory at NUI Maynooth. Face pose was removed using featureless rectification techniques in the space-frequency domain. To accomplish this featureless rectification, certain aspects of image structure were examined. Two aspects of symmetry in images, namely axes of symmetry and continuous symmetry measures, were examined to obtain structural information from the images. Also, the information contained in the frequency components at every location in images was examined as a basis for restoring perspective pose. This came about through the examination of the space-frequency domain, obtained though a continuous wavelet transform. A wavelet-based symmetry coefficient (WBSC) was developed to enable featureless rectification. Based on the WBSC, a system that removed pose from imaged faces was designed. To further investigate the benefits of using the space-frequency domain in pose removal applications, an examination of pose removal from planar scenes was carried out. A featureless, planar rectification algorithm was developed and experimentally compared to similar existing methods in the literature.

## 5.2 Conclusions

This section summarises the methods used and the conclusions drawn from results of this body of research. The conclusions are labelled according to their corresponding chapters.

### 5.2.1 Symmetry in Images

Two aspects of symmetry were examined as a pre-requisite to determining pose from images. Firstly, the problem of estimating the location and orientation of axes of symmetry in digital images in an efficient manner was examined. And secondly, despite the inherent appreciation of most people that symmetry is either present or not, levels of how symmetric objects appear may be derived. An investigation into improving upon existing methods of symmetry quantisation was also carried out.

#### Symmetry Axis

Methods for determining the axis of symmetry in images were examined and evaluated. The method proposed in Prasad and Yegnanarayana (2004) was deemed the most appropriate algorithm for our application and was thus selected. Modifications were made to the algorithm to improve both the efficiency and the accuracy of the algorithm. This came about through restricting the point pairs that were allowed to vote for their axis of symmetry. The initial feature points were filtered though a threshold on curvature of the GVF field at that point's position. Each point must achieve at least this threshold to be considered for point correspondence matching and subsequent voting for their respective axis of symmetry.

This significantly reduced both the number of features to be matched and the search space within which their correspondences must be found. This simple restriction on the allowable voting pairs yielded a vast improvement in efficiency. Results were shown that demonstrate removing background pixels from the voting scheme reduced sources of error, resulting in improved accuracy. The largest source of error was accounted for in the quantisation of the angles in the histogram. The location and orientation of symmetry axes for

symmetric objects with axes of symmetry not running though the image centre were also detected to a high accuracy. The algorithm was also shown to be very robust to noise due to the filtering effect of calculating the GVF field.

#### Symmetry Measure

Current methods in the literature that quantify the level of symmetry an object portrays in images were critiqued. A symmetry coefficient, developed in Gofman and Kiryati (1996), was selected for use in this research. A modification to the calculation of the coefficient was made to account for uneven lighting and noise. A wavelet transformation of the image domain, yielding localised frequency information, achieved these goals. This new wavelet-based symmetry coefficient (WBSC) was used in the remainder of the research. Since it was the frequency content and not the direct grayscale values that were being used in the calculation of the symmetry coefficient, no lighting gradient would affect the symmetry measure.

Results were presented that demonstrate improved smoothness in the WBSC cost manifold for images that are deviating from perfect symmetry. The experiments that were carried out show that the WBSC cost function improves upon the comparison cost functions, in that no incorrect local minima are found, and the single minimum that is found is at the correct location. The WBSC cost function is also shown to be robust to lighting variations and noise. The alterations to the symmetry coefficient make it suitable to be used in an iterative optimisation process to remove pose from images.

### 5.2.2 Removal of Pose from Face Images

Two systems were created to cope with varying pose in images. with each system operating on each presented image individually, with no requirement for a pre-learned 3D face model. The first system removed facial pose from images after the image capture process. An optimisation of the wavelet-based symmetry coefficient (WBSC), proposed in Section 2.2.2, measured on re-rendered views of the subject accomplished this task. The second system was designed to filter out non-frontal images in an online process through a threshold on the calculated WBSC values. Images that deviated too far from ideal values could

be rejected, which reduces the overall computational load required in obtaining pose removed face images with the aid of the pose removal techniques.

Experiments were carried out on a database of real face images to remove pose. For the pose removal algorithm, the results demonstrate improved rectification over the comparison methods. It was shown with subjective and quantitative results that the ellipsoid rectification algorithm provides the most realistic results. These improved results were achieved with few restrictions on the input data. The results obtained for the image filtering technique demonstrate that differentiability between fronto-parallel and otherwise posed images is easily achievable with the author's selection method. This selection algorithm could easily be used to pre-process images from video sequences to ensure the storage of high quality face images.

### 5.2.3 Removal of Pose from Imaged Planar Textures

Through the examination of pose removal from face images, the continuous wavelet transform was used as a tool. Knowing how the frequency response changes across an image provides valuable information about the orientation of imaged planes. A more detailed examination of planar pose estimation was carried out. The two most prominent streams of homography estimation were evaluated and their weaknesses highlighted and discussed. The space-frequency domain was proposed as an alternative domain within which pose estimation may be carried out in a featureless framework. The space-frequency domain incorporates the better aspects of both the spatial domain and the frequency domain techniques.

An algorithm was designed such that the frequency components at positions centrally symmetric to each other would be balanced to restore overall central symmetry in the wavelet transformed image. This came about through an iterative optimisation process using a modified WBSC cost that ensured central symmetry rather than axial symmetry was optimised. This second WBSC, based on central symmetry, provided a mechanism that allowed the estimation of perspective transformations in a featureless framework, a feat which at present no other technique is capable of accomplishing.

The experiments demonstrated that for constrained single axis out of plane

motions the proposed rectification algorithm provided a more realistic approximation to a full perspective transformation than the comparison method which relied on a weak-perspective assumption holding true. It was also shown that for two axes rotations, the error in rectification using the proposed algorithm was very small. Certain motions where the imaged planes are almost in fronto-parallel positions were too small for the algorithm to resolve. In all other cases, perspective pose was removed to this same level. The space frequency domain was shown to be accurate and efficient in removing perspective pose from planar textures where other methods may fail.

## 5.3 Directions for Future Work

The image selection algorithm could be used as a pre-processing stage for face database acquisition from video sequences. An investigation into the improved facial recognition rates achieved using a refined database of images such as this, compared to an unfiltered database would be carried out to further highlight the benefits of such a system.

To further highlight the advantages of using the wavelet based facial rectification algorithm, it would be desirable to examine the accuracy of a recognition system before and after pose removal. Also, a comparison between the accuracy of all current recognition systems that operate on posed face images and the accuracy rate of a recognition system operating on the images obtained using the WFPR technique would be carried out. It would also be desirable to compare the accuracy of the WFPR against the comparison methods again, after the light imbalance has been removed from the images. This would be a fairer representation of a face verification system, where the environmental conditions are more strictly controlled.

The use of the Energy Balancing Planar Rectification techniques could be extended to the pose estimation of pre-configured featureless calibration grids. This would allow easier estimation of homographies with less constraints on the user for the purposes of camera calibration. Automatic camera calibration on robotic visual systems could also be carried out in this manner.

## 5.4 Summary

A face pose removal system was developed in a featureless framework. Symmetry was used as a cue for determining pose. Modifications to a symmetry axis estimation algorithm in the literature were carried out. These modifications achieved better efficiency and accuracy because the majority of unimportant background pixels were removed in a pre-filtering stage developed by the author. A continuous symmetry measure was also developed and shown to be more accurate than existing measures in the literature. The WBSC was shown to be robust to noise and lighting variation.

The facial pose removal system was built upon these symmetry properties. A number of different techniques were used to remove pose from face images. The presented results demonstrated that the WFPR system outperforms similar techniques in the literature when measured against the ground truth. A peer reviewed paper on the topic has been accepted for publication. The WBSC was also applied in an image filtering role. Results demonstrate that the most suitable images for face recognition are extracted from the video sequences using the author's WBSC.

The proposed featureless pose removal technique was applied to out of plane pose estimation of textured planar surfaces. It was shown how perspective pose may be removed in a featureless framework. The presented results demonstrate that pose can be recovered to a reasonable degree of accuracy with the author's EBPR method.

### Publications Arising

The first publication is directly associated with the methods and techniques discussed in the thesis. The second publication deals with the implementation of a high performance computing (HPC) cluster to assist in the execution of the large experiments required for the thesis.

- *Removing Pose from Face Images.*
  Seán Begley, John Mallon and Paul F. Whelan.
  Fourth International Symposium on Visual Computing (ISVC08)
  Las Vegas, Nevada, USA, 1st-3rd December 2008.

- *Cost-Effective HPC Clustering For Computer Vision Applications.*
  Julia Dietlmeier, Seán Begley and Paul F. Whelan.
  International Machine Vision & Image Processing Conference 2008 (IMVIP08)
  Portrush, Northern Ireland, 3rd-5th September 2008.

# Bibliography


Bartlett, M. S., Viola, P. A., Sejnowski, T. J., Golomb, B. A., Hager, J. C. and Ekman, P. (1999). Classifying facial actions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**: 974–989.

Blanz, V. and Vetter, T. (2003). Face recognition based on fitting a 3d morphable model, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**(9): 1063–1074.

Bracewell, R., Chang, K., Jha, A. and Wang, Y. (1993). Affine theorem for two-dimensional fourier transform, *Electronic Letters* **29**(3): 304.

Clerc, M. and Mallat, S. (1999). Shape from texture through deformations, *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Vol. 1, pp. 405–410.

Collins, R. and Beveridge, J. R. (1993). Matching perspective views of coplanar structures using projective unwarping and similarity matching, *IEEE Computer Vision and Pattern Recognition*, pp. 240–245.

Criminisi, A., Reid, I. and Zisserman, A. (1997). A plane measuring device, *Proceedings of the 8th British Machine Vision Conference, Colchester*, UK.
**URL:** *http://www.robots.ox.ac.uk/ vgg*

Edwards, G., Lanitis, A., Taylor, C. and Cootes, T. (1996). Statistical models of face images: Improving specificity.
**URL:** *citeseer.ist.psu.edu/edwards96statistical.html*

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* **24**(6): 381–395.

Gao, Y. and Leung, M. (2002). Face recognition using line edge map, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(6): 764–779.

Garding, J. (1992). Shape from texture for smooth curved surfaces in perspective projection, *Journal of Mathematical Imaging and Vision*, Vol. 2, pp. 329–352.

Ghent, J. and McDonald, J. (2003). Generating a mapping function from one expression to another using a statistical model of facial shape, *Proceedings of the 7th Irish machine vision and image processing conference.*
**URL:** *citeseer.ist.psu.edu/ghent03generating.html*

Gofman, Y. and Kiryati, N. (1996). Detecting symmetry in grey level images: the global optimization approach, *Pattern Recognition, 1996., Proceedings of the 13th International Conference on* **1**: 889–894 vol.1.

Goldstein, A. J., Harmon, L. D. and Lesk, A. B. (1971). Identification of human faces., *Proceedings of the IEEE* **59**(5): 748–760.

Grossmann, A. and Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape, *SIAM J. of Math. An.* **15**: 723–736.

Haralick, R. M., Lee, C., Ottenberg, K. and Nolle, M. (1991). Analysis and solutions of the three point perspective pose estimation problem, *Technical report*, Universitaet Hamburg, Hamburg, Germany.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*, second edn, Cambridge University Press, Cambridge, UK.

Heikkilä, J. (2002). Multi-scale autoconvolution for affine invariant pattern recognition, *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1*, IEEE Computer Society, Washington, DC, USA, p. 10119.

Huang, F. J., Zhou, Z., Zhang, H.-J. and Chen, T. (2000). Pose invariant face recognition, *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on* **1**: 245–250.

Jiang, G., Tsui, H.-T., Quan, L. and Zisserman, A. (2002). Single axis geometry by fitting conics, *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, Springer-Verlag, London, UK, pp. 537–550.

Kadyrov, A. and Petrou, M. (2006). Affine parameter estimation from the trace transform, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(10): 1631–1645.

Kahl, F. and Heyden, A. (1998). Using conic correspondences in two images to estimate the epipolar geometry, *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, p. 761.

Kannala, J., Rahtu, E. and Heikkilä, J. (2005). Affine registration with multi-scale autoconvolution, *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, Vol. 3, pp. III–1064–7.

Kovesi, P. (2005). Shapelets correlated with surface normals produce surfaces, *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, pp. 994–1001.

Lanitis, A., Taylor, C. J. and Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **19**(7): 743–756.

Lee, H.-S. and Kim, D. (2006). Generating frontal view face image for pose invariant face recognition, *Pattern Recogn. Lett.* **27**(7): 747–754.

Lehmann, S., Bradley, A. P., Clarkson, I. V. L., Williams, J. and Kootsookos, P. J. (2007). Correspondence-free determination of the affine fundamental matrix, *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(1): 82–97.

Loh, A. and Hartley, R. (2005). Shape from non-homogeneous, non-stationary, anisotropic, perspective texture, *Proc. British Machine Vision Conference*, Springer, Berlin, pp. 69–78.

Lucchese, L. (2000). A new method for perspective view registration, *Image Processing, 2000. Proceedings. 2000 International Conference on* **2**: 776–779.

Lucchese, L. (2001a). Estimating affine transformations in the frequency domain, *ICIP (2)*, pp. 909–912.

Lucchese, L. (2001b). A frequency domain technique based on energy radial projections for robust estimation of global 2d affine transformations, *Comput. Vis. Image Underst.* **81**(1): 72–116.

Lucchese, L. (2001c). A hybrid frequency-space domain algorithm for estimating projective transformations of color images, *ICIP (2)*, pp. 913–916.

Lucchese, L. and Cortelazzo, G. (1997). Noise-robust estimation of planar roto-translations with high precision, *ICIP '97: Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 1*, IEEE Computer Society, Washington, DC, USA, p. 699.

Mallon, J. and Whelan, P. F. (2007). Which pattern? biasing aspects of planar calibration patterns and detection methods, *Pattern Recogn. Lett.* **28**(8): 921–930.

Marola, G. (1989). On the detection of the axes of symmetry of symmetric and almost symmetric planar images, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **11**(1): 104–108.

OpenCV (2001). Opencv, *www.intel.com/technology/computing/opencv* .

Petrou, M. and Kadyrov, A. (2004). Affine invariant features from the trace transform, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(1): 30–44.

Prasad, V. and Yegnanarayana, B. (2004). Finding axes of symmetry from potential fields, *Image Processing, IEEE Transactions on* **13**(12): 1559–1566.

Quan, L. and Lan, Z. (1998). Linear n>=4-point pose determination., *ICCV*, pp. 778–783.

Rosin, P. L. and Zunic, J. (2005). Measuring rectilinearity, *Computer Vision and Image Understanding* **99**(2): 175–188.

Saint-Marc, P., Rom, H. and Medioni, G. (1993). B-spline contour representation and symmetry detection, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **15**(11): 1191–1197.

Stearns, S. D. and Hush, D. R. (1990). *Digital signal analysis*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Sung, J. and Kim, D. (2008). Pose-robust facial expression recognition using view-based 2d + 3d aam, *Systems, Man and Cybernetics, Part A, IEEE Transactions on* **38**(4): 852–866.

Sung, J.-W. and Kim, D. (2004). Real-time facial pose identification with hierarchically structured ml pose classifier., *IJPRAI* **18**(2): 127–142.

Tsukamoto, A., Lee, C. and Tsuji, S. (1994). Detection and pose estimation of human face with synthesized image models, *ICPR-A* **94**: 754–757.

Weng, J., Ahuja, N. and Huang, T. (1988). Closed-form solution and maximum likelihood: a robust approach to motion and structure estimation, *Computer Vision and Pattern Recognition, 1988. Proceedings CVPR '88., Computer Society Conference on*, pp. 381–386.

Weng, J., Huang, T. and Ahuja, N. (1988). Motion and structure from point correspondences: A robust algorithm for planar case with error estimation, *International Conference Pattern Recognition ICPR88*, pp. 247–251.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture., *Artif. Intell.* **17**(1-3): 17–45.

Xu, C. and Prince, J. (1998). Snakes, shapes, and gradient vector flow, *Image Processing, IEEE Transactions on* **7**(3): 359–369.

Zabrodsky, H., Peleg, S. and Avnir, D. (1995). Symmetry as a continuous feature, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **17**(12): 1154–1166.

Zhang, C. and Cohen, F. S. (2002). 3-d face structure extraction and recognition from images using 3-d morphing and distance mapping., *IEEE Transactions on Image Processing* **11**(11): 1249–1259.

# Appendix A

# Selective Retrieval of Faces from Video - Additional Results

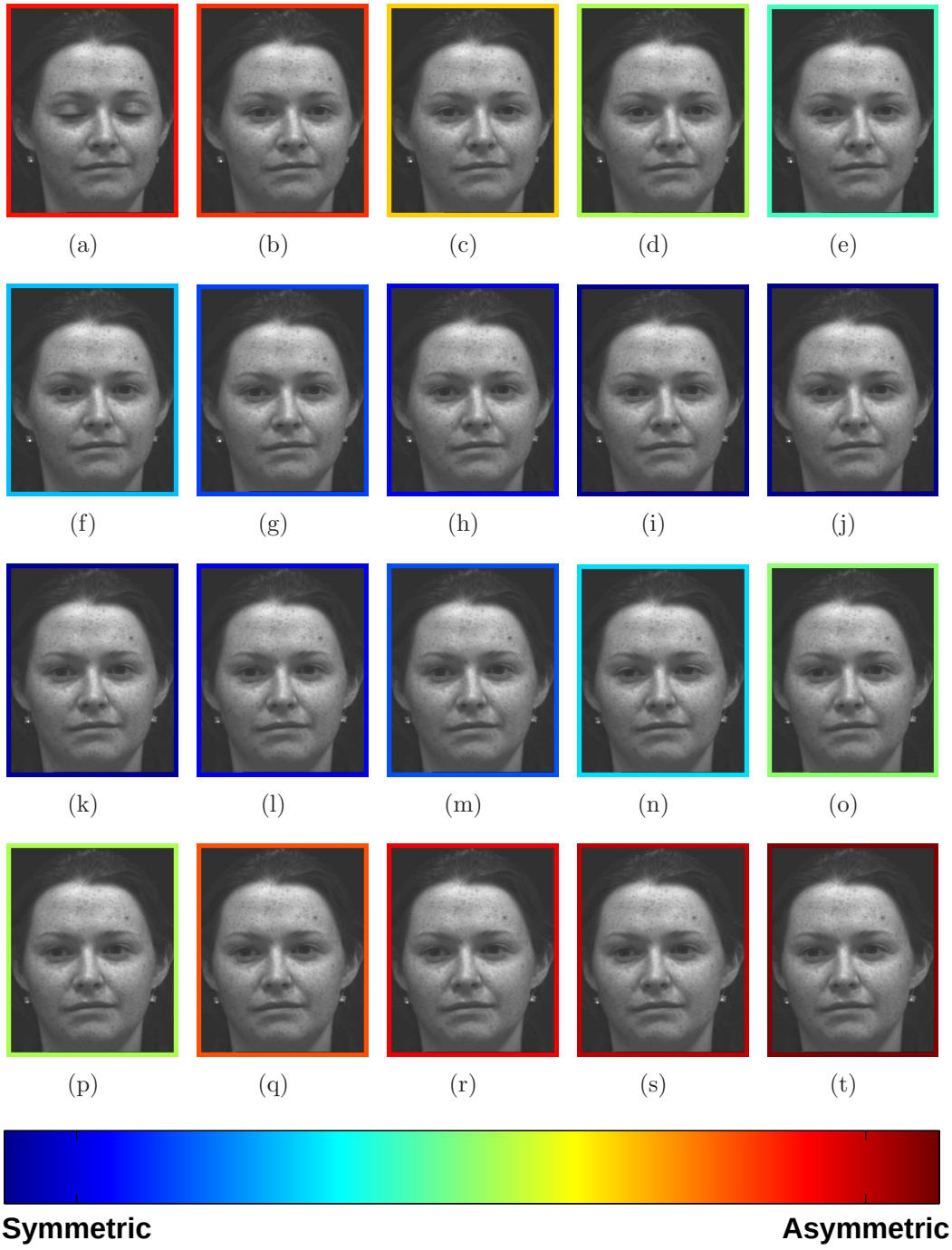



Fig. A.1: Sequence #2 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

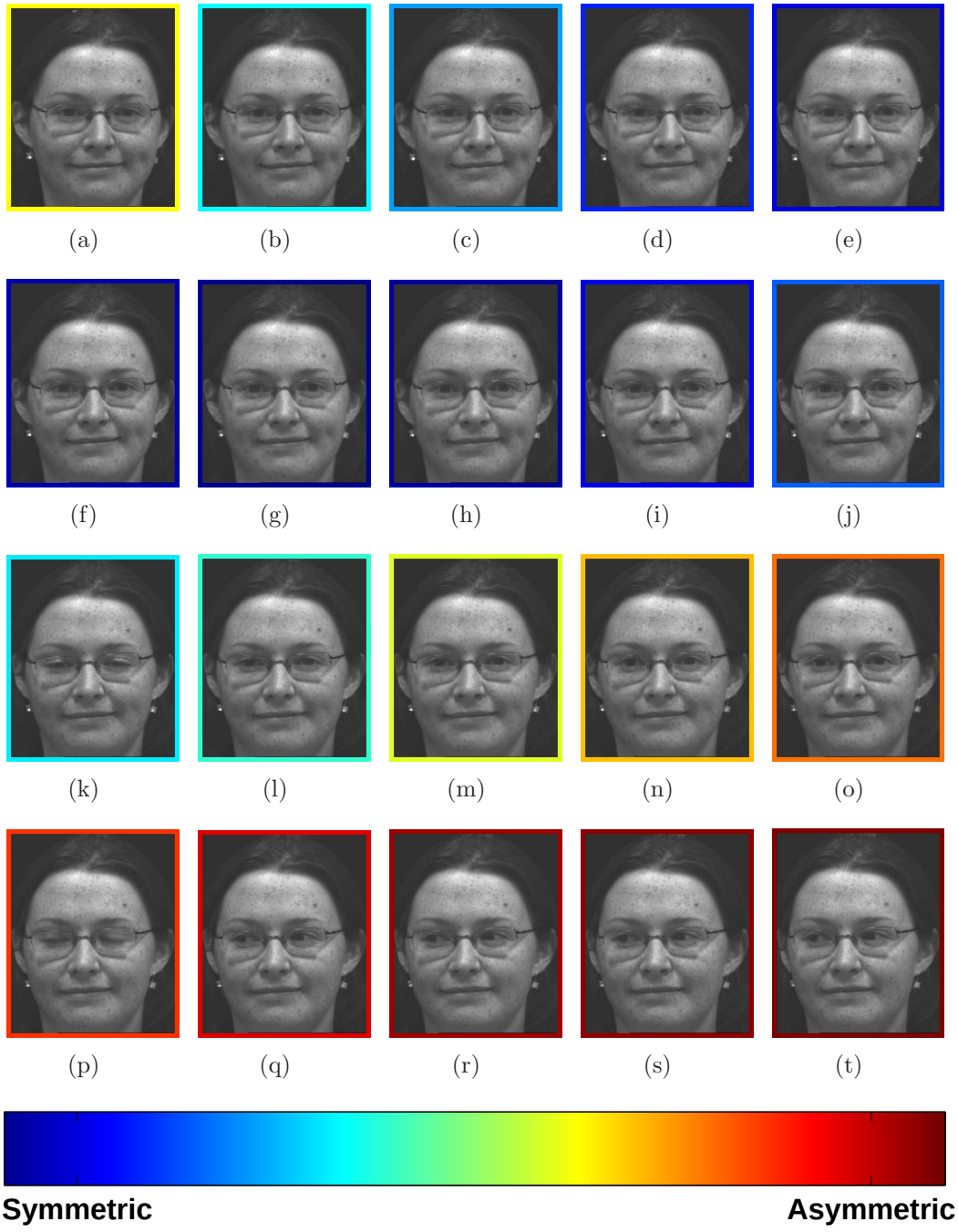



Fig. A.2: Sequence #3 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.


Fig. A.3: Sequence #4 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

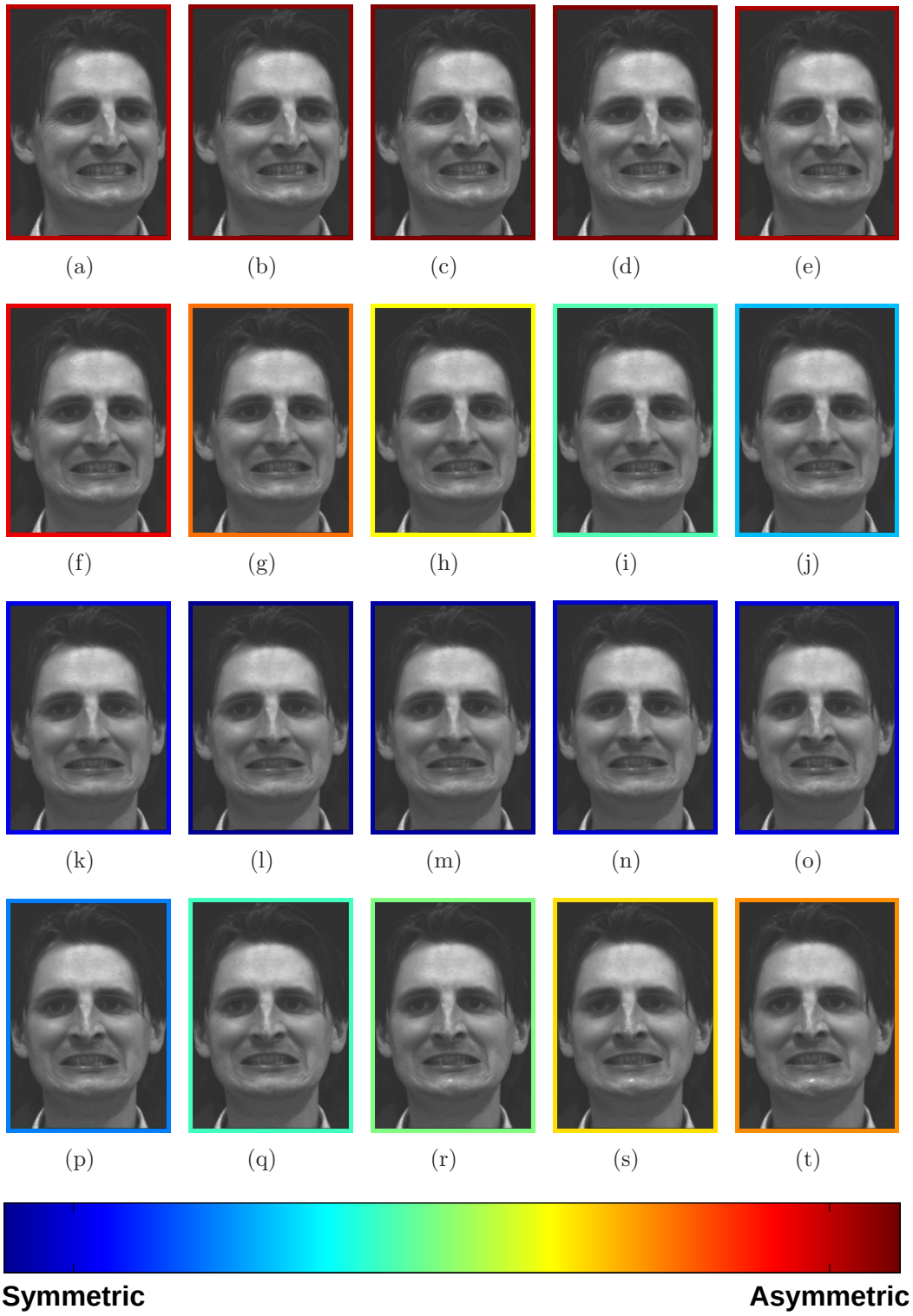



Fig. A.4: Sequence #5 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

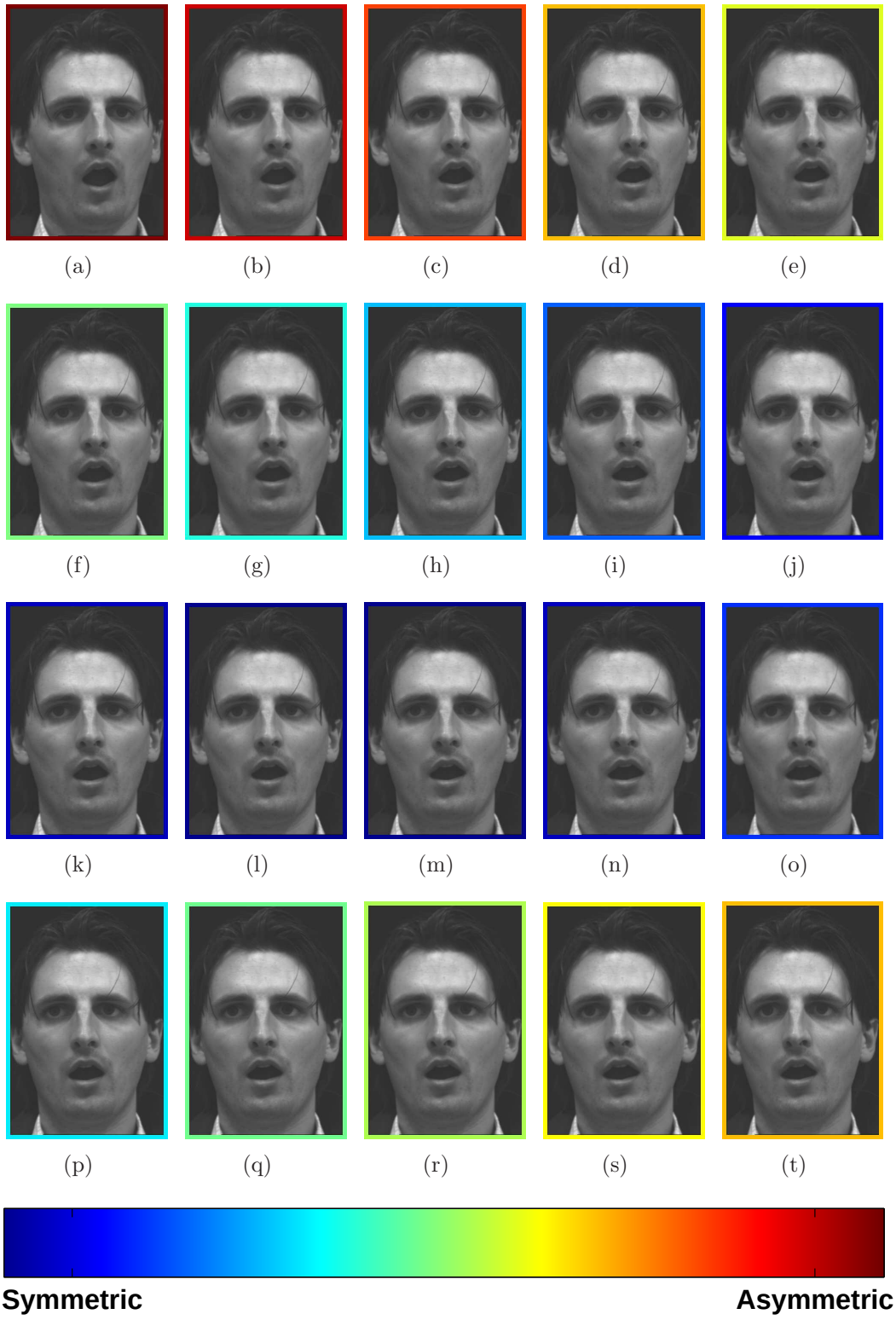



Fig. A.5: Sequence #6 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

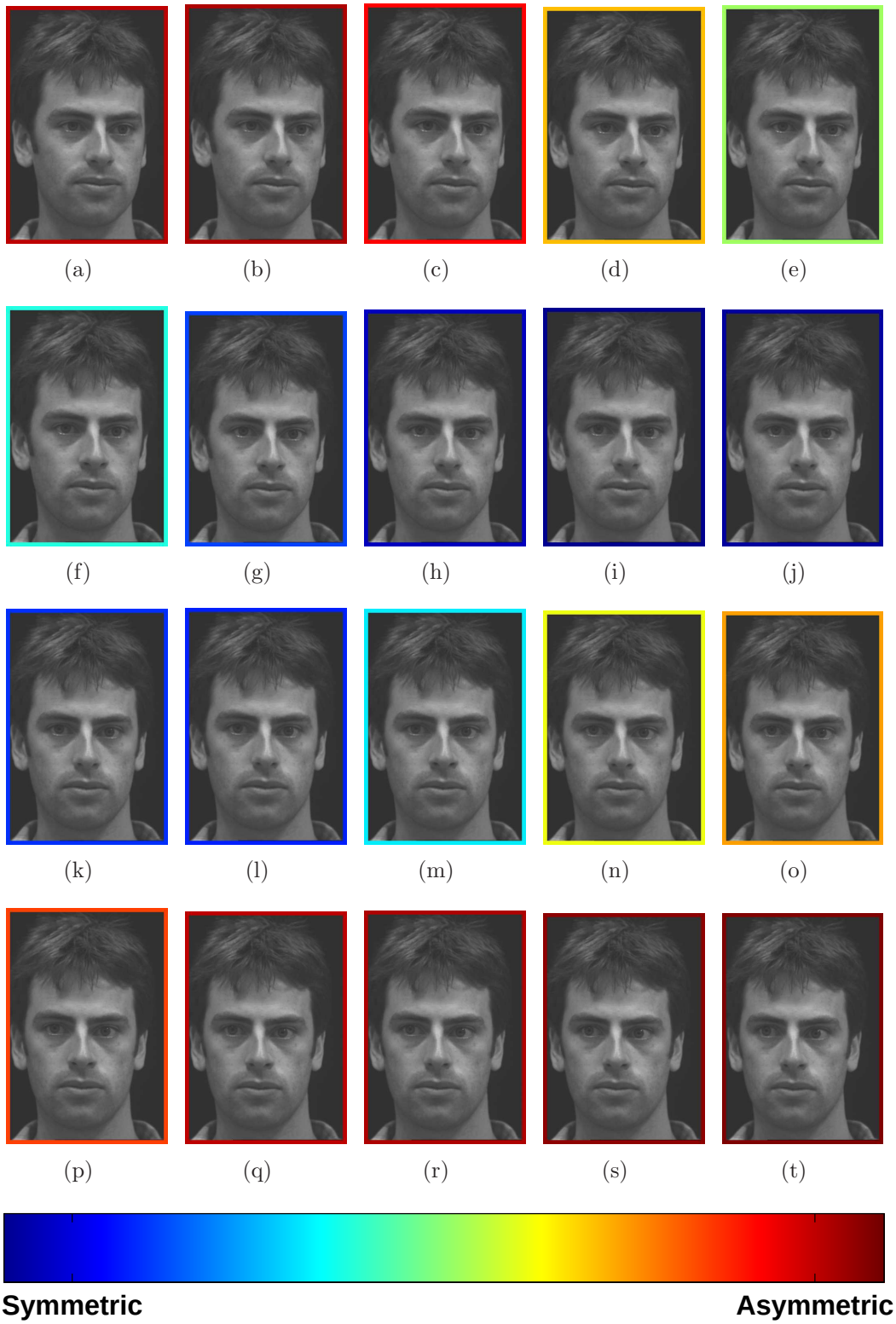



Fig. A.6: Sequence #7 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

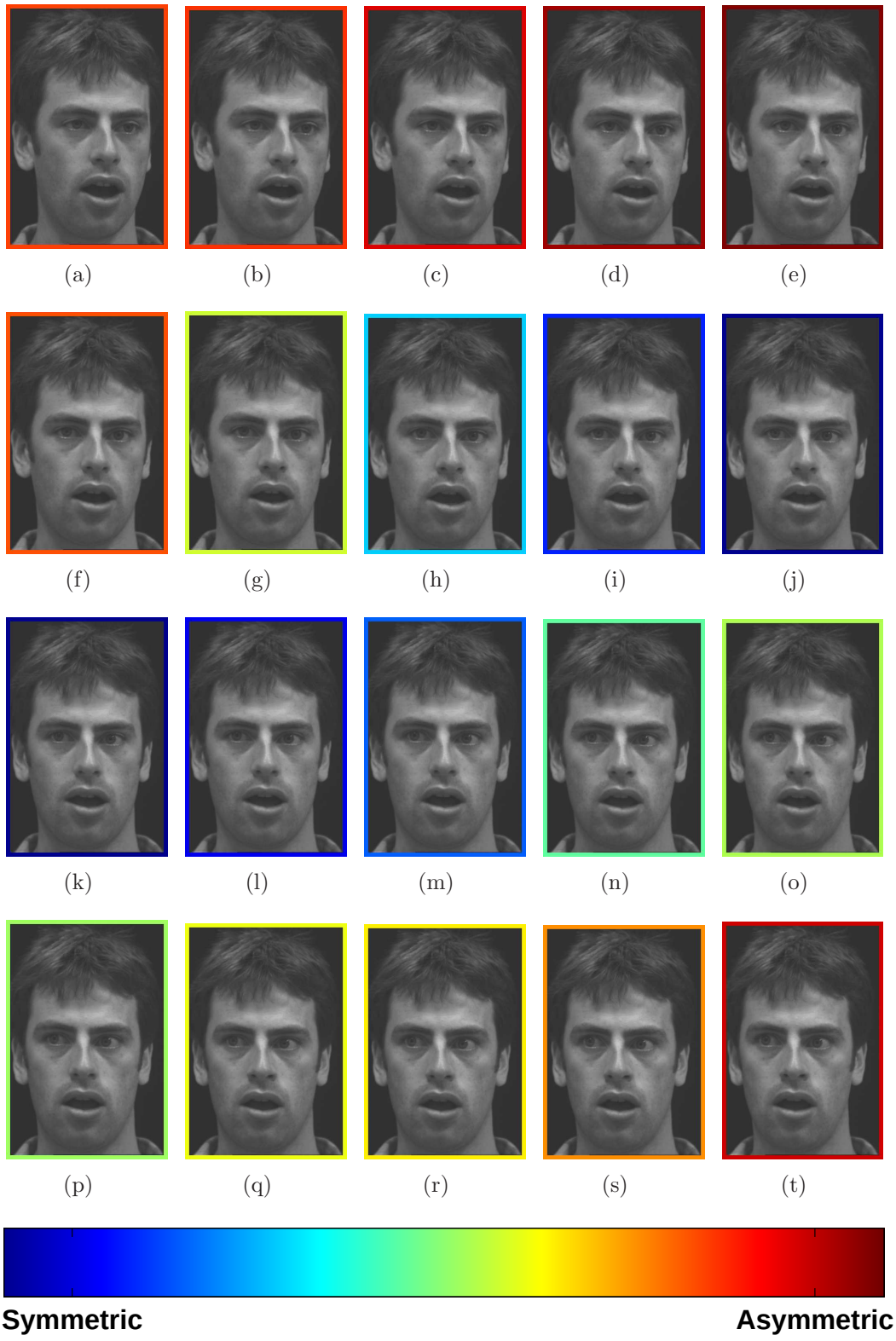



Fig. A.7: Sequence #8 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

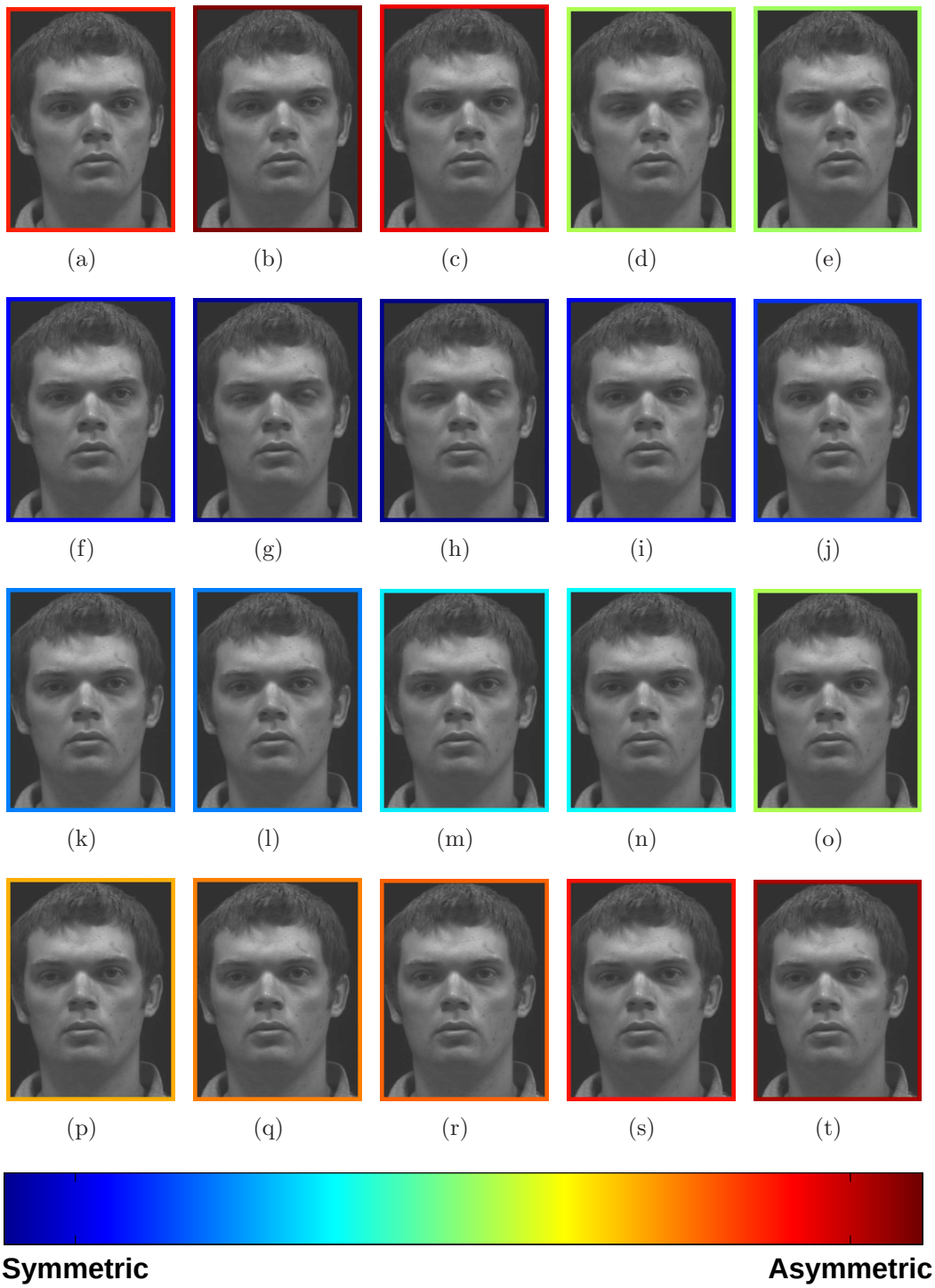



Fig. A.8: Sequence #9 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

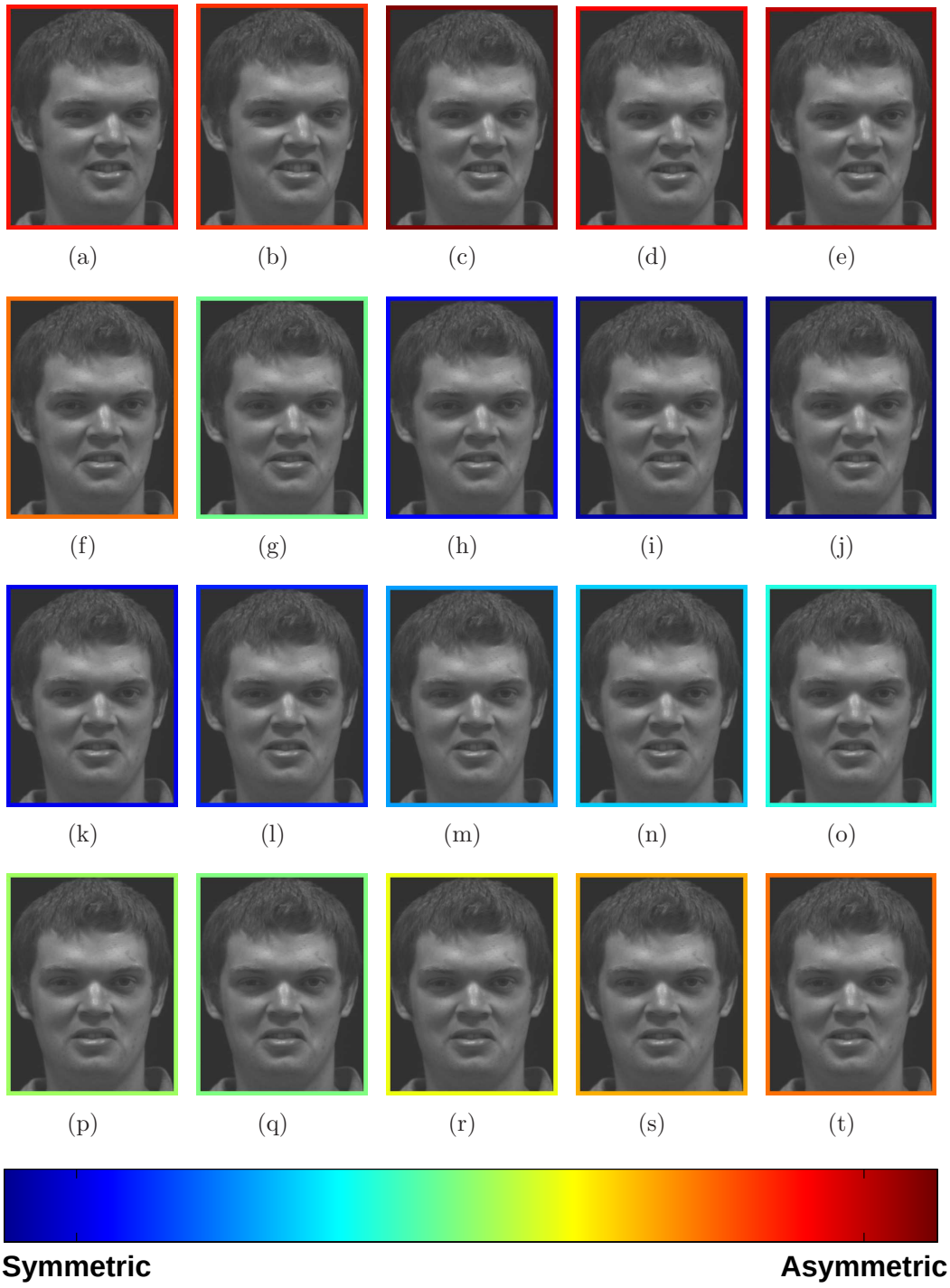



Fig. A.9: Sequence #10 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.


Fig. A.10: Sequence #11 of video frames. The borders around each video frame indicate the value of the WBSC for that frame. The dark blue frames indicate the most fronto-parallel subjects.

# Appendix B

# Planar Image Transformations

A thorough examination of the topics covered in this chapter are available in Hartley and Zisserman (2003), but will be given a brief description here to aid in the understanding of the main theories in this thesis.

## B.1 Homogeneous Coordinates

Two vectors related by an overall scaling factor my be considered to be equivalent in certain cases. For example, two lines $l_1$ and $l_2$, of the form

$$ax + by + c = 0 \tag{B.1}$$

form the same line if the vector of line coefficients for $l_2$, $(a_2, b_2, c_2)^T$, is some multiple of the vector of coefficients for $l_1$. That is, $l_1$ and $l_2$ are equivalent if

$$(a_2, b_2, c_2)^T = k(a_1, b_1, c_1)^T \tag{B.2}$$

Under this equivalence relationship, the two vectors of coefficients are said to be homogeneous, that is they are equal up to some arbitrary scale factor.

A similar representation exists to describe points in a manner that will make applying transformations more efficient, and in fact the representation will allow linear transformations of points under any transformation. Imaged points are generally given coordinates of the form $(x, y)^T$, which for simple rotation and shearing transformations allows a linear transformation of the points to be carried out.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{B.3}$$

However, if a rotation or shearing transformation has to be combined with a translation, the overall transformation can no longer be represented with a linear transformation.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{B.4}$$

To overcome this limitation, we can re-write the coordinate vector as $(x, y, w)^T$, which we will call an homogeneous coordinate, and create a new $3 \times 3$ transformation matrix that combines the rotation or shearing transformation with the translation, resulting in a single matrix multiplication representing the whole transformation.

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & t_1 \\ a_{2,1} & a_{2,2} & t_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{B.5}$$

Typically the $w$ coefficient is 1, and in the simple case of an in-plane or affine transformation like that described above, the output $w'$ coefficient will also be 1. However, when we introduce non-zero elements into the third row of the transformation matrix, we may end up with a $w'$ value that is not equal to 1. Because we said that $(x, y, w)^T$ and $(x', y', w')^T$ were homogeneous vectors, we can find the location of $(x', y')$ in the image by normalising the output vector by $w'$. But how can we be sure that the vectors are homogeneous.

For the vectors to be homogeneous, the un-normalised and the normalised vectors must represent the same image point. To demonstrate this, if you consider an image plane at $w = 1$, then the vector $(x, y, 1)^T$ is a vector from the origin to the point $(x, y)^T$ on the image plane. All multiples of $(x, y, 1)^T$ will have the same direction as $(x, y, 1)^T$ but will have different magnitudes and will pass through $(x, y)^T$ on the image plane. Thus all image coordinates of the form $(x, y, w)^T$ represent the point $(x/w, y/w)^T$ on the image plane and are homogeneous.

## B.2 Projections

A projection is the formation of a 2D image from 3D real world coordinates. Two types of projections will be described here, (i) perspective projections, and (ii) parallel projections. A discussion on the planar transformations required to describe the relation between two views formed using each projection technique will also be given. Additionally, a discussion of how perspective projections may be approximated as parallel projections in certain very specific situations, will be given. This situation arises when objects are image from large distances and is known as the weak perspective situation.

Perspective projections are projections onto an image plane that are formed when all light rays reflected from an object are projected through a single point known as the centre of projection, $C$ in the Fig. B.1. This is the most common situation that exists, and is how the human visual system and cameras form 2D images. With this type of projection, as the name suggests, perspective distortion will be evident meaning that objects further from the imaging device will appear smaller in the image. To fully capture the relationship between two perspective projections of a plane onto the image plane, a full planar transformation matrix is required.

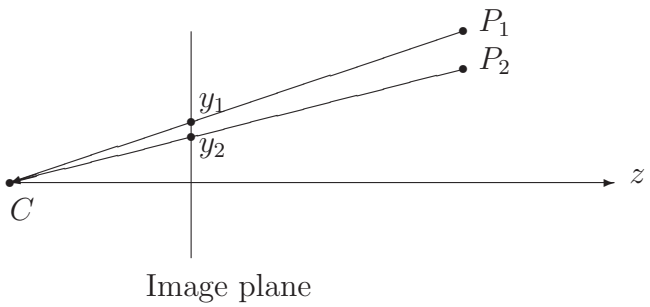



Fig. B.1: Perspective Projection.

Parallel projections are projections onto an image plane that are formed when all light reflected from an object travels in parallel rays onto the image plane. With this type of projection, no perspective effects are present, which results in objects of the same type imaged at different distances appearing as the same size in the image. Each real world coordinate of the form $(x, y, w)^T$ will get projected onto the image plane by simply setting the $w$ coordinate to 1. Because $w$ and $w'$ will always be 1, to describe the relationship between two parallel projection views of a plane, the third row of the planar transformation matrix will always be $(0, 0, 1)$, or in other words, the two view are related by an affine transformation.

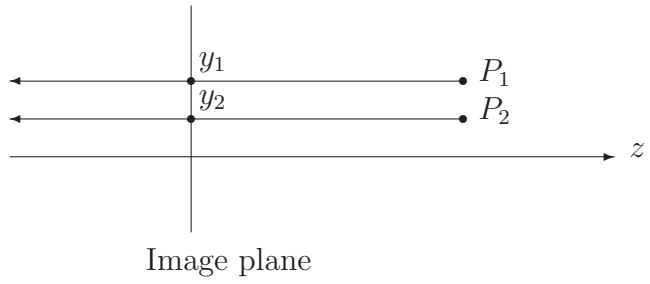



Fig. B.2: Parallel Projection.

Weak perspective projections are formed when an object is imaged at a large distance from the imaging device. If the distance from the imaging device to the object is far greater than the distance from the centre of projection to the imaging plane, then the light rays reflected from the object are seen to be almost parallel. If two identical objects are imaged from this large distance, with one object slightly further away, they will appear to be almost the same size in the image. The further away from the imaging device the objects are placed, the more parallel the reflected light rays are, and less perspective is noticeable. This is known as weak perspective projection. Because weak perspective projections are similar to parallel projections, and can be approximated by parallel projections, so too can two views of a plane imaged under weak perspective projection assumptions be related by an affine transformation.

## B.3 Planar Image Transformations

To fully capture the relative orientation between two planes, a total of eight quantities are required, three for the three rotation angles, three for the three translations, and two to describe affine skewing. In terms of imaged planar scenes, a linear transformation matrix that describes the motion of image points from one view to the other is used. This matrix has eight free parameters, which encode all of the degrees of freedom in the motion between views. To capture changes in scene depth between views, homogeneous coordinates are used. These homogeneous coordinates are effectively a vector of the $x$, $y$, and $z$ coordinates. The image coordinates are simply the $x$ and $y$ values normalised by the $z$ value. Because the $z$ value is not an actual real-world scene measurement and only describes a change in scene depth, it is typically denoted with a $w$ rather than a $z$.

In homogeneous coordinates the linear mapping of points from one view to another may be expressed as the product of the $3 \times 3$ homography matrix $H$, with the homogeneous 3-vector representation of each coordinate. An array of $m$ such coordinates may be built up, with each of the coordinates being one column of the matrix. The transformed points are typically denoted as $\mathbf{x}'$, another homogeneous 3-vector. Mathematically, the transformation is written as:

$$\mathbf{x}'_i = H\mathbf{x}_i \tag{B.6}$$

$$\begin{bmatrix} x'_1 & x'_2 & \dots & x'_m \\ y'_1 & y'_2 & \dots & y'_m \\ w'_1 & w'_2 & \dots & w'_m \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \dots & x_m \\ y_1 & y_2 & \dots & y_m \\ w_1 & w_2 & \dots & w_m \end{bmatrix} \tag{B.7}$$

The transformed homogeneous vectors and their corresponding normalised output vectors in the above equations are not equal in magnitude, but are equal in direction. This is a vital piece of information that provides a mechanism for a least squares solution for the homography matrix $H$ to be found. It is more convenient to re-write the matrix equation in terms of the rows of the homography ($\mathbf{h}_1$, $\mathbf{h}_2$ and $\mathbf{h}_3$) as a vector as this will allow a more concise way of writing the equations for determining the linear spatial solutions.

$$H\mathbf{x}_i = \begin{bmatrix} \mathbf{h}_1\mathbf{x}_i \\ \mathbf{h}_2\mathbf{x}_i \\ \mathbf{h}_3\mathbf{x}_i \end{bmatrix} \tag{B.8}$$

# Appendix C

# The Direct Linear Transform

Given an array of imaged points on a plane in one view and their corresponding imaged points on that plane in a second view, a linear transformation may be computed to map one set of points onto the other. Using homogeneous coordinates, the transformation is represented as a $3 \times 3$ matrix.

It was noted in Appendix B that the transformed homogeneous vectors for one set of points will be equivalent to the homogeneous vector representations of the second set of imaged points up to a scale factor. This means that if the cross product of any two corresponding vectors is found, it should be zero. It is this that allows us to get an algebraic least squares solution for the homography matrix H using 4 or more points. The cross product may be written as:

$$\mathbf{x}'_i \times H\mathbf{x}_i = \begin{bmatrix} x'_i \\ y'_i \\ w'_i \end{bmatrix} \times \begin{bmatrix} \mathbf{h}_1\mathbf{x}_i \\ \mathbf{h}_2\mathbf{x}_i \\ \mathbf{h}_3\mathbf{x}_i \end{bmatrix} \tag{C.1}$$

$$= \begin{bmatrix} y'_i\mathbf{h}_3\mathbf{x}_i - w'_i\mathbf{h}_2\mathbf{x}_i \\ w'_i\mathbf{h}_1\mathbf{x}_i - x'_i\mathbf{h}_3\mathbf{x}_i \\ x'_i\mathbf{h}_2\mathbf{x}_i - y'_i\mathbf{h}_1\mathbf{x}_i \end{bmatrix} \tag{C.2}$$

At this point it can be noted that the matrix representation for the cross product can be expanded and written with all nine elements of the homography appearing in each row. This will allow us to factor out a vector of the homography elements and thus solve the equation using standard linear algebra methods.

$$\mathbf{x}'_i \times H\mathbf{x}_i = \begin{bmatrix} 0\mathbf{h}_1\mathbf{x}_i - w'_i\mathbf{h}_2\mathbf{x}_i + y'_i\mathbf{h}_3\mathbf{x}_i \\ w'_i\mathbf{h}_1\mathbf{x}_i + 0\mathbf{h}_2\mathbf{x}_i - x'_i\mathbf{h}_3\mathbf{x}_i \\ -y'_i\mathbf{h}_1\mathbf{x}_i + x'_i\mathbf{h}_2\mathbf{x}_i + 0\mathbf{h}_3\mathbf{x}_i \end{bmatrix} \tag{C.3}$$

$$= \begin{bmatrix} \mathbf{0}^T & -w'_i\mathbf{x}_i^T & y'_i\mathbf{x}_i^T \\ w'_i\mathbf{x}_i^T & \mathbf{0}^T & -x'_i\mathbf{x}_i^T \\ -y'_i\mathbf{x}_i^T & x'_i\mathbf{x}_i^T & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \mathbf{h}_3^T \end{bmatrix} \tag{C.4}$$

This is of the form $A_i\mathbf{h} = 0$, and may now be solved if enough constraints on the vector $\mathbf{h}$ are found. The third row of the above matrix may be disregarded as it

is a linear combination of the first two. Therefore, each point correspondence gives rise to a pair of equations (the top two rows of the $A_i$ matrix), and if enough of these point correspondences are found, then a linear solution for the vector $\mathbf{h}$ may be found. Since there are eight degrees of freedom in the elements of the vector $\mathbf{h}$ and each point correspondence gives rise to two equations or two constraints, at least four point correspondences are needed to solve for $\mathbf{h}$.

$$\begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix} (\mathbf{h}) = \mathbf{0} \tag{C.5}$$

$$A \quad \mathbf{h} = \mathbf{0} \tag{C.6}$$

So if four linearly independent points are used to construct the matrix $A$, then the matrix will have a rank of eight. This means that a single null vector will exist to allow the equation $A\mathbf{h} = 0$ to be true. This null vector is the exact solution for the vector $\mathbf{h}$, which in turn is the homography matrix $H$.

If there are more than four points, the matrix is over-determined and the solution must be found using the singular value decomposition of the matrix $A$. The solution is the singular vector corresponding to the smallest singular value, and in turn this vector gives back the homography matrix. For the over-determined case, effectively we are finding the minimum norm of the error vector $A\mathbf{h}$ which is the same as the algebraic distance between the points of interest in one image and the same interest points in the second image that have been transformed by $H$.

$$d_{alg}\left(\mathbf{x}_i', H\mathbf{x}_i\right) \tag{C.7}$$

# Appendix D

# Affine Theorem of the Fourier Transform

The affine theorem of the Fourier transform in two dimensions was formalised in Bracewell et al. (1993). It was shown that affine warping in the spatial domain is the inverse of the affine warping of the magnitude of the Fourier transform in the frequency domain. Also, the translation between two views can be easily derived from the two-dimensional shift in the phase plots. Using this theory, a number of different methods may be employed to solve for the affine and translation parameters. These methods include a direct linear transform method that uses the positions of corresponding Fourier magnitude peaks in two images to compute the affine transform, and an optimisation method, developed in Lucchese (2001a), that iteratively estimates the transform. A brief derivation of the theorem will clarify the methods and their modes of operation.

**Affine Transformation of Images**

The affine point transformation of inhomogeneous co-ordinates, is described by the pair of equations:

$$x' = ax + by + c \tag{D.1}$$

$$y' = dx + ey + f \tag{D.2}$$

which in matrix form may be written as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ d & e \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c \\ f \end{bmatrix} \tag{D.3}$$

This equation can be manipulated in order to get an expression for $\begin{bmatrix} x \\ y \end{bmatrix}$ in terms of the other elements.

$$\begin{bmatrix} a & b \\ d & e \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} - \begin{bmatrix} c \\ f \end{bmatrix} \tag{D.4}$$

$$\begin{bmatrix} a & b \\ d & e \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' - c \\ y' - f \end{bmatrix} \tag{D.5}$$

This is in the form of $Ax = b$, and so $x$ can be directly written in terms of $A$ and $b$ by getting the inverse of $A$.

$$\mathbf{x} = A^{-1}\mathbf{b} \tag{D.6}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} e & -b \\ -d & a \end{bmatrix}\begin{bmatrix} x'-c \\ y'-f \end{bmatrix} \tag{D.7}$$

**The two-dimensional Fourier Transform**

This is a standard transform that will be written in a more convenient form. In this form, it's properties will be easier utilised in order to solve for affine transformations using all of data in an image.
The standard $2D$ Fourier Transform:

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(ux+vy)}\,dx\,dy \tag{D.8}$$

If we extract just the phase exponent of the above equation, we can rewrite it in a more convenient manner:

$$(ux+vy) = \begin{bmatrix} u & v \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} \tag{D.9}$$

$$= \begin{bmatrix} u & v \end{bmatrix}\frac{1}{\Delta}\begin{bmatrix} e & -b \\ -d & a \end{bmatrix}\begin{bmatrix} x'-c \\ y'-f \end{bmatrix} \tag{D.10}$$

If we consider $\begin{bmatrix} u & v \end{bmatrix}$ to be the transposed vector $k$, representing the position on the Fourier transformed plot of the image, then we may write the above equation more conveniently in symbolic form as:

$$(ux+vy) = \mathbf{k^T x} \tag{D.11}$$

$$= \mathbf{k^T A^{-1} b} \tag{D.12}$$

Vector $\mathbf{b}$ can be broken into two separate vectors, which is what makes it possible to isolate the affine warping from the translation and analyse them separately.

$$\mathbf{b} = \mathbf{x'} - \mathbf{t} \tag{D.13}$$

$$\begin{bmatrix} x'-c \\ y'-f \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} - \begin{bmatrix} c \\ f \end{bmatrix} \tag{D.14}$$

Thus a concise representation of the Fourier transform's exponent is obtained.

$$(ux+vy) = \mathbf{k^T A^{-1} x'} - \mathbf{k^T A^{-1} t} \tag{D.15}$$

The two-dimensional Fourier Transform may now be written as:

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi\mathbf{k^T x}}\,dx\,dy \tag{D.16}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi\mathbf{k^T A^{-1} b}}\,dx\,dy \tag{D.17}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(\mathbf{k^T A^{-1} x'} - \mathbf{k^T A^{-1} t})}\,dx\,dy \tag{D.18}$$

Instead of having just one exponent, we may split the exponent into two since there's a linear combination of the elements in the exponent.

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(\mathbf{k^T A^{-1}x'})}e^{j2\pi(\mathbf{k^T A^{-1}t})}\,dx\,dy \quad \text{(D.19)}$$

This may be further simplified to give a more convenient representation. The $\mathbf{t}$ vector doesn't have any effect on the integration since it only contains the translation information, so it may be moved outside the integral entirely.

$$F(u,v) = e^{j2\pi(\mathbf{k^T A^{-1}t})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(\mathbf{k^T A^{-1}x'})}\,dx\,dy \quad \text{(D.20)}$$

Just as $F(u,v)$ is the $2D$ Fourier transform of $f(x,y)$, so $G(u,v)$ is the $2D$ Fourier transform of $g(x,y)$, where $g(x,y)$ is the affine transformed image.

$$g(x,y) = f(ax+by+c, dx+ey+f) \quad \text{(D.21)}$$

so we may now write an expression for $G(u,v)$

$$\begin{aligned} G(u,v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(ax+by+c,dx+ey+f)e^{-j2\pi\mathbf{k^T A^{-1}b}}\,dx\,dy \\ &= e^{j2\pi\mathbf{k^T A^{-1}t}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(ax+by+c,dx+ey+f)e^{-j2\pi\mathbf{k^T A^{-1}x'}}\,dx\,dy \end{aligned}$$

Using the Jacobian relation $dx'\,dy' = |\Delta|\,dx\,dy$ we can change the variables of integration in the Fourier transform of the affine warped image.

$$G(u,v) = e^{j2\pi\mathbf{k^T A^{-1}t}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')e^{-j2\pi\mathbf{k^T A^{-1}x'}}\,dx'\,dy'\frac{1}{|\Delta|} \quad \text{(D.22)}$$

where $\Delta$ is the determinant of the $2\times 2$ affine warping matrix $A$. To simplify the expression, a coordinate frame for the affine warped Fourier transform image is required. this may simply be written as.

$$\mathbf{k'^T} = \mathbf{k^T A^{-1}} \quad \text{(D.23)}$$

$$\begin{bmatrix} u' & v' \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} u & v \end{bmatrix}\begin{bmatrix} e & -b \\ -d & a \end{bmatrix} \quad \text{(D.24)}$$

or

$$\mathbf{k^T} = \mathbf{k'^T A} \quad \text{(D.25)}$$

$$\begin{bmatrix} u & v \end{bmatrix} = \begin{bmatrix} u' & v' \end{bmatrix}\begin{bmatrix} a & b \\ d & e \end{bmatrix} \quad \text{(D.26)}$$

This gives a nice tidy equation for the Fourier transform of the affine warped image.

$$G(u,v) = \frac{1}{|\Delta|}e^{j2\pi\mathbf{k'^T t}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')e^{-j2\pi\mathbf{k'^T x'}}\,dx'\,dy' \quad \text{(D.27)}$$

Now if we were to take the Fourier transform of the affine-warped image in the affine-warped frequency coordinates, we'd get the following:

$$F(u', v') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x', y') e^{-j2\pi \mathbf{k'^T x'}} \, dx' \, dy' \tag{D.28}$$

So now we can directly relate the Fourier transform of the affine-warped coordinate system to that of the un-warped system.

$$G(u, v) = \frac{1}{|\Delta|} e^{j2\pi \mathbf{k'^T t}} F(u'v') \tag{D.29}$$