

# - 3-D Face Recognition -

Anders Eriksson

December 1999



Thesis presented in partial fulfilment  
of the requirements for the degree of  
**Master of Electronic Engineering**  
at the  
**University of Stellenbosch**

*Study Leader: David Weber*

## *Declaration*

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and has not previously in its entirety or in part been submitted at any university for a degree.

Signature :

Date : February 7, 2000

# *Abstract*

In recent years face recognition has been a focus of intensive research but has still not achieved its full potential, mainly due to the limited abilities of existing systems to cope with varying pose and illumination. The most popular techniques to overcome this problem are the use of 3-D models or stereo information as this provides a system with the necessary information about the human face to ensure good recognition performance on faces with largely varying poses.

In this thesis we present a novel approach to view-invariant face recognition that utilizes stereo information extracted from calibrated stereo image pairs. The method is invariant of scaling, rotation and variations in illumination. For each of the training image pairs a number of facial feature points are located in both images using Gabor wavelets. From this, along with the camera calibration information, a sparse 3-D mesh of the face can be constructed. This mesh is then stored along with the Gabor wavelet coefficients at each feature point, resulting in a model that contains both the geometric information of the face as well as its texture, described by the wavelet coefficients. The recognition is then conducted by filtering the test image pair with a Gabor filter bank, projecting the stored models feature points onto the image pairs and comparing the Gabor coefficients from the filtered image pairs with the ones stored in the model. The fit is optimised by rotating and translating the 3-D mesh. With this method reliable recognition results were obtained on a database with large variations in pose and illumination.

# *Opsomming*

Alhoewel gesigsherkenning die afgelope paar jaar intensief ondersoek is, het dit nog nie sy volle potensiaal bereik nie. Dit kan hoofsaaklik toegeskryf word aan die feit dat huidige stelsels nie aanpasbaar is om verskillende beligting en posisie van die onderwerp te hanteer nie. Die bekendste tegniek om hiervoor te kompenseer is die gebruik van 3-D modelle of stereo inligting. Dit stel die stelsel in staat om akkurate gesigsherkenning te doen op gesigte met groot posisionele variansie.

Hierdie werk beskryf 'n nuwe metode om posisie-onafhanklike gesigsherkenning te doen deur gebruik te maak van stereo beeldpare. Die metode is invariant vir skalering, rotasie en veranderinge in beligting. 'n Aantal gesigspatrone word gevind in elke beeldpaar van die oplei-data deur gebruik te maak van Gabor filters. Hierdie patrone en kamera kalibrasie inligting word gebruik om 'n 3-D raamwerk van die gesig te konstrueer. Die gesigmodel wat gebruik word om toetsbeelde te klassifiseer bestaan uit die gesigraamwerk en die Gabor filter koëffisiënte by elke patroonpunt.

Klassifisering van 'n toetsbeeldpaar word gedoen deur die toetsbeelde te filter met 'n Gabor filterbank. Die gestoorde modelpatroonpunte word dan geprojekteer op die beeldpaar en die Gabor koëffisiënte van die gefilterde beelde word dan vergelyk met die koëffisiënte wat gestoor is in die model. Die passing word geoptimeer deur rotasie en translasie van die 3-D raamwerk. Die studie het getoon dat hierdie metode akkurate resultate verskaf vir 'n databasis met 'n groot variansie in posisie en beligting.



# *Acknowledgements*

I would like to thank the following people, without who this thesis would never have been finished :

My supervisor, David Weber (even though he is a Penguin fan), for his great support and encouragement.

Chris Venter for the many hours he put into helping me collect images for the facial databases.

My fellow lab members for putting up with me for two long years.

... and anyone else that made my stay in South Africa a very memorable one.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Statement . . . . .	1
1.2	Prior Work . . . . .	2
1.2.1	Feature Based Recognition . . . . .	2
1.2.2	Template Matching . . . . .	3
1.2.3	View Tolerance . . . . .	4
1.3	Algorithm Overview . . . . .	5
1.3.1	Training Procedure . . . . .	5
1.3.2	Testing Procedure . . . . .	7
1.4	Case Studies . . . . .	7
1.4.1	Facial Stereo Image Database - FASIM . . . . .	9
1.4.2	Pose Image Database - POSIM . . . . .	10
1.5	Outline of thesis . . . . .	10
<b>2</b>	<b>Stereo Vision Systems</b>	<b>12</b>
2.1	Stereopsis . . . . .	12
2.2	Camera Calibration . . . . .	16
2.2.1	Locating the Calibration Points. . . . .	18
2.2.2	Camera Parameter Estimation . . . . .	20
<b>3</b>	<b>Face Recognition</b>	<b>25</b>
3.1	Gabor Features . . . . .	25
3.2	Similarity Functions . . . . .	27
3.3	Training . . . . .	33
3.3.1	Facial Feature Points . . . . .	33

---

3.3.2	Stereo Matching . . . . .	35
3.3.3	Model Extraction . . . . .	39
3.3.4	Weighting of Features . . . . .	41
3.4	Testing . . . . .	42
3.4.1	Optimal Mapping of the Model . . . . .	43
3.4.2	Scale and In-plane Rotation Compensation . . . . .	48
3.4.3	Recognition & Verification Analysis . . . . .	50
<b>4</b>	<b>Experimental Evaluation</b>	<b>51</b>
4.1	Investigation of the Performance of the Pose Compensating Algorithms. .	51
4.2	Recognition & Verification . . . . .	60
4.2.1	The Testing Procedure . . . . .	62
4.2.2	Stereo Image Mapping Results . . . . .	62
4.2.3	Single Image Mapping Results . . . . .	64
4.2.4	Glasses . . . . .	66
4.3	Analysis & Discussion . . . . .	68
<b>5</b>	<b>Conclusions</b>	<b>71</b>
5.1	Summary of Results . . . . .	71
5.2	Future Work . . . . .	72
5.3	Summary . . . . .	73

# List of Figures

1.1	A block diagram of our face recognition system. . . . .	6
2.1	A pinhole, or perspective, camera. . . . .	13
2.2	The pinhole camera model. . . . .	13
2.3	The setup of a simple binocular stereo system. . . . .	14
2.4	The epipolar geometry of a stereo system. . . . .	15
2.5	Reconstruction on non-intersecting back-traced rays. . . . .	16
2.6	A demonstration of the discretisation of the reconstructable space. . . . .	17
2.7	Illustration of the camera geometry with the different coordinate system used. . . . .	17
2.8	The calibration object used. . . . .	18
2.9	An image of a calibration object where the calibration points have been located and sorted. . . . .	18
2.10	An illustration of the different steps of the calibration point sorting algorithm. . . . .	19
2.11	The calibration object and its projection onto the image plane. . . . .	21
2.12	The finished camera calibration GUI for single images. . . . .	24
3.1	The real (left) and imaginary (right) parts of a Gabor filter. . . . .	26
3.2	The absolute value of two example Gabor filters in the spatial domain, left, and in the frequency domain, right. Observe how the different orientation and scale of the filters determine the frequency domain positioning. . . . .	27
3.3	The filtering of a face with two different Gabor kernels. Note how the two filters light up different parts of the face, they are both sensitive to features with an orientation and scale similar to their own. . . . .	28
3.4	The position in the frequency domain of the 32 Gabor filters used by our system. . . . .	29
3.5	A jet containing the N complex outputs from the filtering procedure. . . . .	29

3.6	Illustration of the positions of the extracted jets. The eye jet $J_{eye}$ , marked by a circle, is compared from left to right with all the jets positioned on the horizontal line, $J'_l$ . . . . .	32
3.7	The outputs of the three different similarity functions when comparing jets extracted from Figure 3.6. Note the local maxima at $x \approx 160$ , the position of the right eye. . . . .	32
3.8	The resulting similarities of a jet taken from the left eye of the same subject, but from another image, compared with the jets along a line at the same horizontal position using the three methods described. The results are compared with the similarity values from Figure 3.7, (dashed), for (a) the absolute based-, (b) the displacement based and (c) the phase based similarity function. In this figure the robustness of the different approaches can clearly be seen. . . . .	33
3.9	A more detailed block diagram of the training procedure. . . . .	34
3.10	A face marked with the location of the facial landmarks used by the system. . . . .	35
3.11	A closer look at a feature point located close to the top of the left ear. . . . .	36
3.12	Point $A$ in the left image is matched to point $B$ in the left image. $B$ is then matched back into the right image on to point $A'$ . The match $A \mapsto B$ is deemed left-to-right consistent if $A - A'$ is smaller than a threshold $\epsilon$ . . . . .	37
3.13	The desired result of the stereo correspondence matching procedure. . . . .	39
3.14	The desired result of the stereo correspondence matching procedure. . . . .	39
3.15	The reconstructed 3-D facial mesh. . . . .	40
3.16	The extracted model, the sparse 3-D facial mesh with the corresponding jets of each landmark node. . . . .	40
3.17	The resulting weighting of the facial landmarks. The higher significance of the eye regions indicates that an important part of the discriminatory information is located there. . . . .	43
3.18	A more detailed block diagram of the testing procedure. . . . .	44
3.19	The mapping of a model onto an image pair with and without the rigid transformation that constitutes the pose compensation. . . . .	46
3.20	Examples of scaled and in-plane rotated heads with the corresponding filter compensations. . . . .	49



4.1	A posed scaling versus a manually constructed one, both images are scaled down by 30% compared to the training image. The posed scaling is very difficult to control and is therefore carried out by first capturing a number of images with different, unknown scales and then manually measuring scaling afterwards. When constructing scaled images by manipulation of an original image issues such as how to deal with image edges will arise. In the left image above an attempt to correct such an edge effect by stretching the image borders can be observed. . . . .	52
4.2	Two versions of in-plane rotated faces. The same problems as in Figure 4.1 will occur in this instance as well. . . . .	52
4.3	The images in the pose database used to examine scale compensation, scaling ranges from +15% to -45% compared to the training image. . . .	53
4.4	The images in the rotated test set. . . . .	54
4.5	The resulting model mapping onto the scaled image set. . . . .	56
4.6	The manually measured scaling along with the system's estimated scaling referenced to the training image of the 8 image pairs in this test set. . .	57
4.7	The facial similarity values for different compensation methods as a function of scale. . . . .	57
4.8	This figure illustrates how two initially well aligned faces, the training image, top left, and test image no. 7, with the same scale as the training face, in the top right need further in-plane translation compensation. Comparing the un-optimized mapping of the extracted mesh, bottom left, and in-plane translation optimized mapping, bottom right, the facial similarity value almost trippled when the $x$ and $y$ parameters are optimized. This again points out the importance of accurately locating the face to be recognized, the optimal mapping has a projection only 20 pixels away from the initial one. . . . .	58
4.9	The resulting model mapping onto the rotated image set. . . . .	59
4.10	The manually measured and the, by the system, estimated in-plane rotation of the 11 faces in the test set. . . . .	60
4.11	The facial similarity values for different compensation methods as a function of rotation angle. The symmetry of the setup was utilized to extend the figure to included rotation in both directions. . . . .	60
4.12	The recognition results on the FASIM database. Cumulative match score as a function of rank, 78% of the subjects were correctly identified and in 92% of the cases the correct match was found within the top three ranks.	63
4.13	The averaged equal error rate of a 100 test runs as a function of verification set size. It can here clearly be seen how performance improves as the verification set grows. . . . .	64



4.14	The average false rejection and false acceptance rate from 30 test runs for different acceptance thresholds, $t_{acc}$ . The equal error rate is 7.5% for $t_{acc} \approx 0$ , this should be compared with the lowest obtainable EER, 4.5%, see Figure 4.13. . . . .	65
4.15	Single image mapping recognition results on the FASIM database compared with the stereo image mapping results from Figure 4.12. A first rank recognition rate of 66% and a third rank recognition rate of 85% constitutes a performance reduction of about 10% from the stereo image mapping method. . . . .	66
4.16	The acceptance/rejection curve, averaged from 30 test runs, for single image mapping compared to stereo image mapping. The equal error rate increased from 7.5% to 12%. . . . .	67
4.17	The recognition rate when testing subjects wearing glasses against the models from the FASIM database. Same session first rank recognition rate is 75% and different session recognition rate 60%. . . . .	68
4.18	The average acceptance/rejection curve for subjects with and without glasses. Note the differing equal error rates for same session ( $\approx 10\%$ ) and different session images ( $\approx 15\%$ ). . . . .	69
4.19	Comparison of recognition rates for test images acquired during the same session as the training image and images acquired in different photo session. Note that these results can not be compared directly with the previous ones as the size of the training set is different, see Section 1.4. . .	70

# List of Tables

3.1	The different wavelengths and orientations of the Gabor filters used by our system . . . . .	26
-----	--	----

# Chapter 1

## Introduction

In recent years face recognition has been a focus of intensive research but has still not achieved its full potential, mainly due to the limited abilities of existing systems to cope with varying pose and illumination. In this thesis we present a novel approach to view-invariant face recognition by employing 3-D face models extracted from calibrated stereo image pairs. The use of stereo imaging provides robustness against pose and illumination variations.

### 1.1 Problem Statement

The formal objective of this thesis was to investigate how stereo information from binocular image pairs could be used in face recognition and compare the performance of such an approach against results from existing, well known, single-cue systems. A fully working system, was to be implemented with the following criteria:

- The system was to be able to perform face recognition as well as face verification from one training image pair per person. By recognition we mean the *determination* of a persons identity and by verification the *confirmation* of a persons identity.
- The underlying advantages of using stereo information, such as pose and scale invariance, should be exploited and the performance gains quantified.

- The whole system should be made fully automatic.

## 1.2 Prior Work

There has been much work done in the field of face recognition but the task of comparing the various methods is a difficult one since the results reported originated from the use of different databases of varying size and degree of difficulty. The lack of a common terminology for describing different approaches also adds to the dilemma.

In 1994 the U.S. Army Research Laboratory tried to address this problem by starting the *Face Recognition Technology* (FERET) program [1]. Its main purpose is to supply the image processing community with a large database of images as well as a reliable testing procedure. Since the launch of this project there has been three supervised face recognition tests administered, with participants from all over the world. The results from these tests are the most reliable face recognition benchmarks available today. The FERET test also showed how an algorithm's performance can change drastically between databases. Direct comparisons of face recognition systems can therefore not be carried out, the testing procedure and the particulars of the training and testing set must be taken into account.

The algorithms that have participated in the FERET test, as well as most others, can be separated into two categories, geometric- or feature based approach and template based matching. There also exists a number of combined, hybrid, versions of these two.

### 1.2.1 Feature Based Recognition

One of the first face recognition system developed by Kanade in 1977 [2] had a feature based approach. He extracted a small number of geometrical features from each face for classification. A recognition rate of 75%, on a database of 20 subjects, was reported. A similar but more recent algorithm, by Brunelli and Poggio [3], where a larger number of geometrical measurements were used for recognition has also shown promising results. On a database of 47 people this system managed to correctly identify 90%. Jai and Nixon [4],



extended the feature vector by taking the profile projection, the moments of the eye areas and the face contours into account. These features were proven to increase performance significantly when used in combination with the geometric facial measurements.

One of the most successful feature based method today is the Elastic Bunch Graph matching algorithm by Wiskott *et al.* [5]. They use elastic grids and Gabor wavelet coefficients to represent individual faces. Matching is performed by examining how much a model grid has to be stretched to fit a face. This algorithm was one of the contenders in the FERET phase III test in September 1996 [6] where it, in most categories, finished in the top three. One important property of this system is its inherent ability to automatically locate faces.

## 1.2.2 Template Matching

The second method of performing face recognition is template matching, where the images pixel values are compared instead of their respective distance as was the case in the previous method. The perhaps most famous face recognition approach, the eigenface algorithm by Turk and Pentland [7], belongs to this category. They treat each image as a row vector and then performs a Karhunen-Loeve transform on all the face vectors in the training set. The basis vectors in this new, reduced coordinate system are called eigenfaces and it is in this space recognition is performed by examining the distance from the reduced training vectors with the projection of the test images vector onto this space. This has on numerous occasions been proven to be a very accurate method but since this algorithm works on the whole image and not only the face, it is very sensitive to variations in pose, scale, facial expression, illumination and background clutter. The eigenface algorithm was implemented and tested in the FERET phase III competition for comparison only but still produced good results. The most successful system in this test was a further development of the eigenface method, Moghaddam [8], where a Bayesian analysis of the class distributions is used to enhance accuracy.

Brunelli and Poggio [3], performed recognition using normalized grey-scale correlation with some success. An accuracy of 90% on a database containing 47 subjects was achieved.

Hidden Markov models (HMM) have proven to be successful in this field as well. Samaria [9] managed to recognize 85% out of 200 people by using HMM's to model the spatial configuration of human faces.

Neither of these two took part in the FERET III test however, but the results still seem promising.

### 1.2.3 View Tolerance

A few words must be said about the robustness of the above mentioned algorithms when it comes to changes in scaling and pose. Not many face recognition systems today that can deal with these variations. This was proved in the FERET III test where only two out of the ten finalists could perform recognition without being provided with the eye coordinates.

Due to the grey level matching nature of template matching this category is generally more sensitive to pose and scale than the feature based approach. The latter does struggle with this problem as well as the task of accurately locating facial features becomes more difficult with changes in pose.

The most popular techniques to overcome this problem are the use of 3-D models or stereo information as this provides a system with the necessary information about the human face to ensure good recognition performance on faces with largely varying poses.

Fromherz [10] used image sequences to extract dense depth maps of human faces. These depths maps were then used, in combination with the texture map, to perform face recognition with convincing results. The use of range scanners in face recognition has been proposed by Achermann *et al* [11]. A recognition rate of 100%, using an eigenface approach, on a database containing 24 subjects was reported. Miruyama *et al.* suggested the use of bidirectional synthesis on off-frontal images to generate virtual views of the subjects in different poses for the purpose of face recognition. Preliminary results have been reported to be promising.



## 1.3 Algorithm Overview

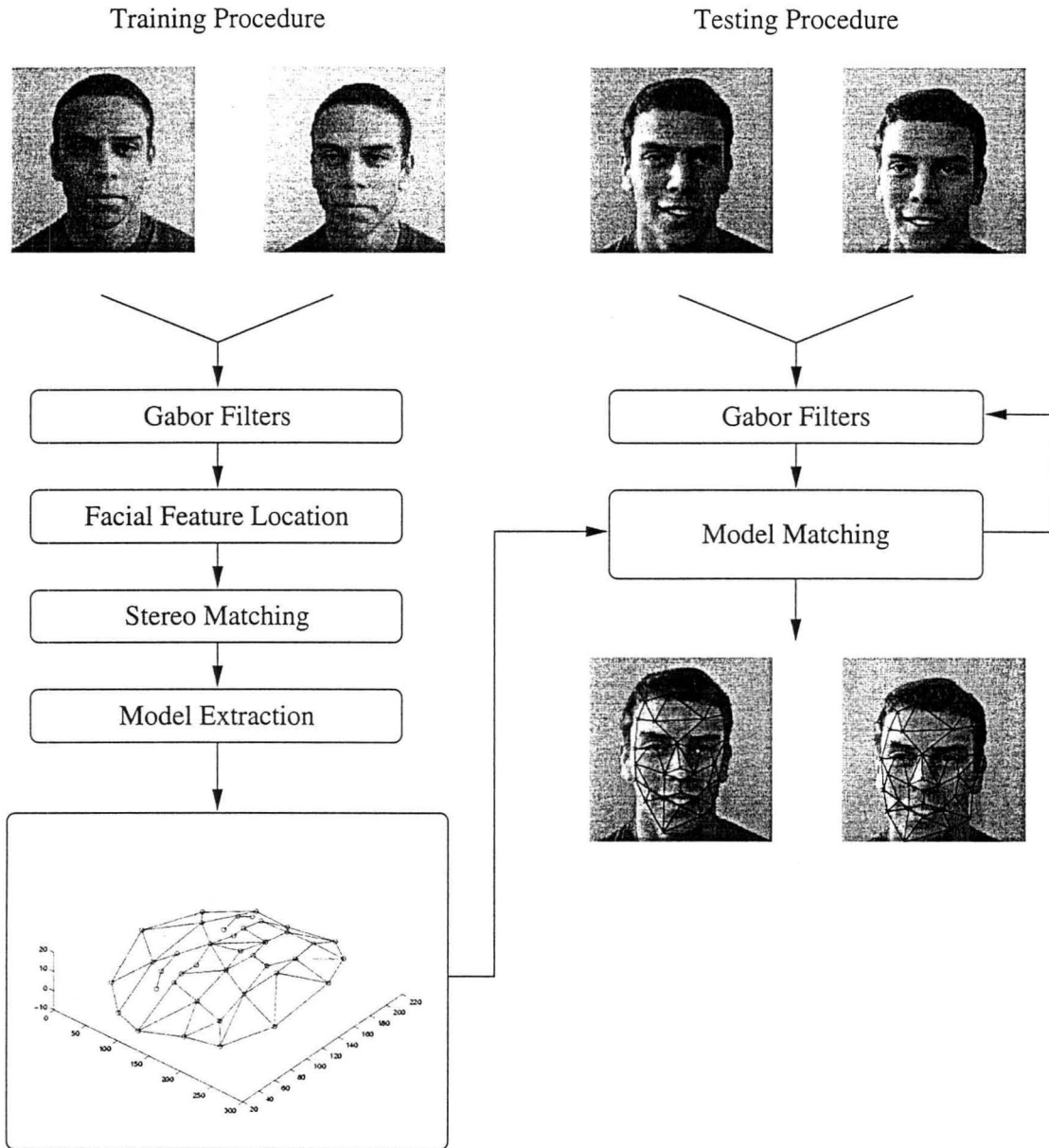
Since one of the main criteria for our system was pose- and scale invariance we chose a feature based approach, loosely based on the Elastic Bunch Graph (EBG) algorithm by Wiskott *et al.* [5], a choice based on its demonstrated performance but also on account of the intuitive potential this algorithm show for use in stereo modeling.

A block diagram of our system can be seen in Figure 1.1. Similar to EBG our system is also made up of two distinctly separate procedures. One for training and one for testing.

### 1.3.1 Training Procedure

The training procedure extracts 3-D models from calibrated stereo image pairs of subjects to be inserted into the database. This is a stand alone procedure that only requires the camera calibration data as additional information. No re-training of the database has to be performed after new models have been added, as is the case for, for example, the eigenface method. The training procedure is made up of the following parts:

- **Preprocessing with Gabor Filters** In this part the images are filtered through a large filterbank of Gabor filters. These Gabor filters are tuned to detect image features at different scales and orientations. The output coefficients from these convolutions are used for facial feature location, stereo matching and finally recognition.
- **Facial feature location** The first step in constructing a 3-D model from a face is to locate a number of facial features in one of the images, in our case this is done in the left image. The feature finding is performed by employing standard pattern recognition techniques, along with manually trained examples of the facial features in question, on the image. The processing is performed on the Gabor filter coefficients from the previous stage.
- **Stereo matching** The second step in constructing the 3-D face model is to locate the facial features in the right image as well, by performing stereo matching. This



**Figure 1.1:** A block diagram of our face recognition system.

is accomplished by searching along the epipole in the right image for the Gabor coefficients that matches best between the two images. The performance of this matching is improved by first locating coarse features and then gradually zooming in to the smaller ones.

- **Model extraction** The positions of the facial features in both images and the camera calibration information, makes a 3-D construction possible. This 3-D model or *mesh* of the face is stored along with the Gabor filter outputs to make up the template.

### 1.3.2 Testing Procedure

In the testing procedure the system compares an image pair with a stored template and returns the probability of a correct match.

- **Preprocessing with Gabor Filters** The preprocessing stage is the same as in the training procedure.
- **Model matching** The template matching is performed by mapping the template mesh onto the test image pair and rotating and translating the mesh so that it fits the face optimally. By examining the required shift in position of the face we can easily determine the scaling and in-plane rotation of the face. This is compensated for by refiltering the image pairs with a Gabor filterbank that has been scaled and rotated the same amount. For increased accuracy the mapping is also repeated. It should be noted that this mapping can be performed on only one image, with a slight reduction in performance. This will have the advantage, however, of allowing a test station of only needing one camera.

## 1.4 Case Studies

As previously mentioned it is very difficult to compare the results from different face recognition algorithms as no other 3-D databases were available. As shown in the FERET



phase III test [1] the recognition rate can vary from close to a 100% down to 30% for the same algorithm, depending on the database.

Some of the aspects that must be taken into consideration when evaluating the difficulty of a facial database are listed below:

- **Size :** The number of subjects in the database and the number of training images per person. A small number of subjects in the training set will make recognition easier as there are fewer options to choose from. As this number increases the recognition rate will typically decrease. More than one training image per subject will improve performance as more examples of what the subject looks like will be available. The size of the database will also determine the significance of the results, a larger test set will ensure that the resulting recognition rate was not achieved by chance but that it really reflects the performance of the system.
- **Lighting :** Have efforts been made to control the lighting conditions? Even small changes in how the face is illuminated will affect performance, especially for gray-scale based template matching algorithms such as the eigenface method.
- **Demographics :** Is age, gender and ethnic origin evenly represented? The narrower the demographics the harder the problem gets. It is more difficult to distinguish between people that resemble each other.
- **Time :** What was the time span between photo sessions. People change with time which makes identification harder. If the time span is very small the same camera setup might have been used resulting in very similar lighting and background. Subjects also have a tendency to pose in a very similar way during one and the same photo session, all of this resulting in increased similarity.
- **Background :** Were the photos taken with a uniform or a cluttered background? Unless the system has ways of separating the face from the background, different backgrounds will affect performance negatively.

- **Appearance** : Does the appearance of the subjects change significantly between sessions? Are glasses, beards, changes in hairstyle and varying facial expressions allowed?
- **Pose & Scaling** : Is the pose and scaling of the face controlled?
- **Face location** : Were the faces automatically or manually located? Automatic location along with pose and scale invariance is a feature most systems lack today, the eye coordinates must be supplied by some type of preprocessing stage. Using this information the images are rescaled, shifted and resized before they are passed on to the face recognition stage.

Since these questions can not be answered quantitatively, it is not possible to make a direct comparison of the results from algorithms tested on different databases. They do help to make any such comparison less biased. We were not able to acquire a publicly available facial stereo image database, one had to be constructed.

Two different databases were constructed, the first was used to test how well the algorithm could distinguish between a large number of individuals. The second was used to evaluate the system's ability to cope with large variations in pose and scale.

### 1.4.1 Facial Stereo Image Database - FASIM

The purpose of this database was to test our system's recognition abilities. It consists of calibrated image pairs of 219 different subjects, 108 of whom only appeared for one session and were therefore used for the training of the facial feature location stage and as a garbage class in validation tests. Out of the other 111 individuals, 26 were used in a validation set, leaving 85 people to make up the testing set. The subjects were all undergraduate student in engineering, resulting in a very narrow demographics consisting of mainly white male between 21 and 23 years old. No real efforts were made to control lighting, pose, scaling or facial expression. Therefore scaling varies between +13% and -27%, the pose varies

mostly in tilting between  $\pm 7^\circ$ <sup>1</sup>. Persons wearing glasses were photographed twice, once with glasses and once without, resulting in an additional set of 58 image pairs, 13 of which were used in the validation set. This spectacle database is used to evaluate the system's robustness to these kind of variations in appearance. The time between the photosessions for the training set and testing set were between one and two months.

It also should be added that the resulting database is very difficult, due to the narrow demographics and because of the time difference between training and testing images. Many of the facial recognition results that are being reported is on same session recognition, when training and testing images are taken only minutes apart.

### 1.4.2 Pose Image Database - POSIM

The second database was constructed in order to investigate the system's ability to deal with and compensate for pose and scale variances. This was carried out by keeping the other variables, such as facial expression and lighting, as constant as possible and controlling the pose and position of the subjects closely. Twenty image pairs were acquired of one subject at a number of different poses, 11 with the in-plane rotation of the subjects face varying from  $0^\circ$  to  $32^\circ$  in even steps and 8 images with different scales, from +14% to -45%. The last image pair was used to train the system and as it is to this image that scale and rotation is referenced.

## 1.5 Outline of thesis

Chapter 2 describes the basic concepts of stereo vision and disparity, followed by a discussion on the camera setup and the calibration procedure.

Chapter 3 contains a more detailed explanation of the face recognition system. All the components of both the training- and the testing procedure from the preprocessing with Gabor filters and the facial feature location to the stereo matching and the template

---

<sup>1</sup>These scale and rotation variations were estimated by our system during testing



matching are thoroughly explained.

The experimental results from the tests carried out on the different databases are presented in Chapter 4.

The thesis is finally rounded off in Chapter 5 with a summary and evaluation of the work and the results and also a discussion on possible directions for future work.

## Chapter 2

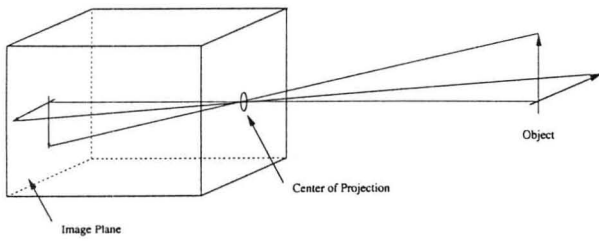
# Stereo Vision Systems

### 2.1 Stereopsis

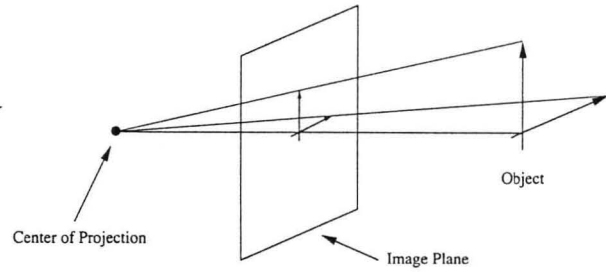
Stereo vision refers to the ability to extract depth information of objects from two or more images taken from different positions. In this system, image pairs are acquired from a binocular stereo camera setup to determine the 3-D structure of human faces that are to be used for recognition purposes. The fundamental principles behind these kinds of stereo systems are fairly straight-forward and can easily be visualized using pinhole camera models.

In a pinhole, or perspective, camera, (Figure 2.1), the lens aperture is made extremely small, pinhole sized, so that only single rays from each visible point in space can enter. The simplest way to model such an optical system is with an image plane and a center of projection  $O$  at a distance  $f$  from that plane, (Figure 2.2). More intricate models include lens distortion and will be discussed in section 2.2.2.

Figure 2.3 shows the top view of a simple stereo system consisting of two pinhole cameras and, since the system described in this thesis utilizes binocular stereo, the following discussion will be focused on such cases. The cameras are placed so that their two image planes  $I_l$  and  $I_r$  are co-planar and the two centers of projection  $O_l$  and  $O_r$  are at a distance  $D$  from each other. A point  $P$ , at a distance  $Y$ , is projected onto two the images planes at



**Figure 2.1:** A pinhole, or perspective, camera.



**Figure 2.2:** The pinhole camera model.

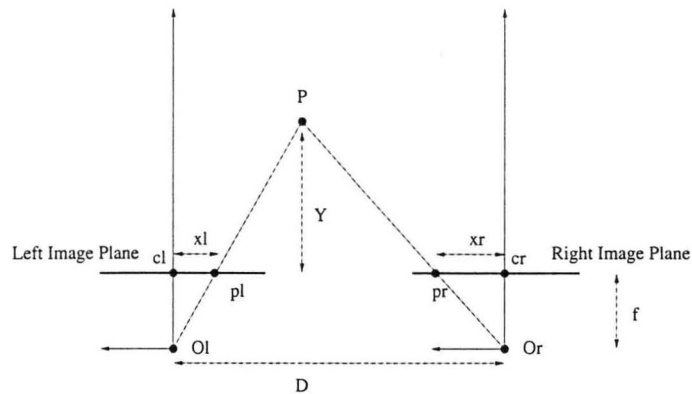
points  $p_l$  and  $p_r$  respectively. Given this geometry of a stereo system it is possible, for each location of  $P$ , to calculate the resulting, and unique, projection onto the image planes. Owing to the symmetrical nature of the problem it will also be possible to calculate the position of  $P$  if the two projection points  $p_l$  and  $p_r$  are known, using the basic geometrical theorem of similar triangles. It can be obtained that the distance  $D$  between  $P$  and the two image planes is given by

$$Y = f \frac{D}{x_l - x_r} = f \frac{D}{d} \quad (2.1)$$

where  $x_l$  and  $x_r$  are the coordinates of the projection points in the image planes. The difference between them is called *disparity*,  $d$ . Disparity is inversely related to distance so the closer an object is to the camera the more will it move when the viewpoint changes, a statement easily verified by everyday observations. It is this method of estimating depth by observing how different objects and image points move when the viewpoints change that make up the fundamental principle behind stereopsis. These theories can easily be extended to more than two cues if desired.

Even though the above mentioned example is extremely simplified, real systems can very rarely be modeled with coplanar image planes and no lens distortion, it does manage to adequately illustrate the two main issues of stereo; namely *correspondence* and *reconstruction*.

Correspondence is not only the procedure of determining the position of corresponding



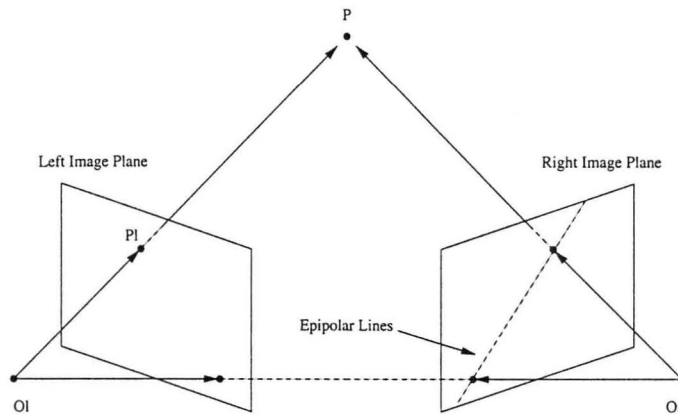
**Figure 2.3:** The setup of a simple binocular stereo system.

image points in both images but also the determination of what points, owing to occlusions, that must not be matched. This problem is a very difficult one and has been the focus of numerous researchers [12–14].

More formal the correspondence matching problem can be defined as *the task of for each point in the left image finding the corresponding point in the right image*. In order to reduce the search space for this matching the geometry of stereo systems, or *epipolar geometry*, is utilized even further. Instead of, for each point in the left image search the entire right image for the best match, the search space can be reduced by first determining all the potential positions in the right image of that point. As shown in Figure 2.4 all the potential positions of  $P_l$  in the right image plane are placed on a single line, *epipolar line*, running through the right image plane. Not only does this reduce the search space to one dimension, it also improves the performance of the matching since the number of potential false matches can be reduced.

The epipolar lines can be determined if both the *extrinsic* and the *intrinsic* parameters of the stereo system are known. The extrinsic parameters describe how the two image planes are located compared to each other. There are six extrinsic parameters, the three components for translation and the three rotation angles. These variables are mainly used to transform coordinates from 3-D world coordinates to coordinates with the image planes as references. The intrinsic parameters contains the physical information about the



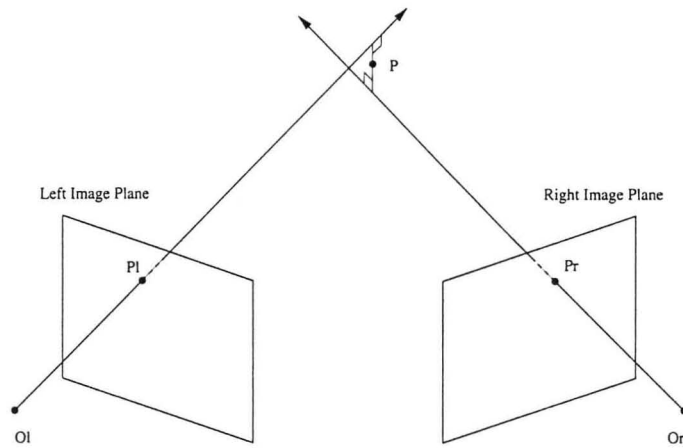


**Figure 2.4:** The epipolar geometry of a stereo system.

cameras, e.g. focal length, lens distortion and pixel width. Both these parameters are determined by performing a calibration of the stereo system. This is discussed in the next section. A description of how the correspondence matching is performed by this system can be found in section 3.3.2.

The task of the reconstruction stage is to construct a *disparity map* containing the disparity information of all or only selected points in the image. From this the true shape of the scene can be constructed. As both the extrinsic and intrinsic parameters are known it is a fairly straight forward exercise in triangulation. The laws of optics state, as previously mentioned, that the projections of any point  $P$  will be positioned where the lines drawn between  $P$  and the centers of projection intersect with respective image plane. Conversely, given the two projections and the centers of projection the location of  $P$  can be determined by tracing each of the two lines or rays back towards their point of origin and noting where they meet. This will obviously be the location of the projected point  $P$ .

There are however a few issues concerning reconstruction that need to be discussed. Firstly, errors in the calibration data as well as approximations made in the correspondence matching will result in non-intersecting rays, (Figure 2.5). This is usually solved with a least-squared-error approach placing the reconstructed point in between the two lines. This situation can also be caused by insufficient lens distortion modeling and quantization errors generated by pixel sampling. The quantization error will also have an effect on recon-



**Figure 2.5:** Reconstruction on non-intersecting back-traced rays.

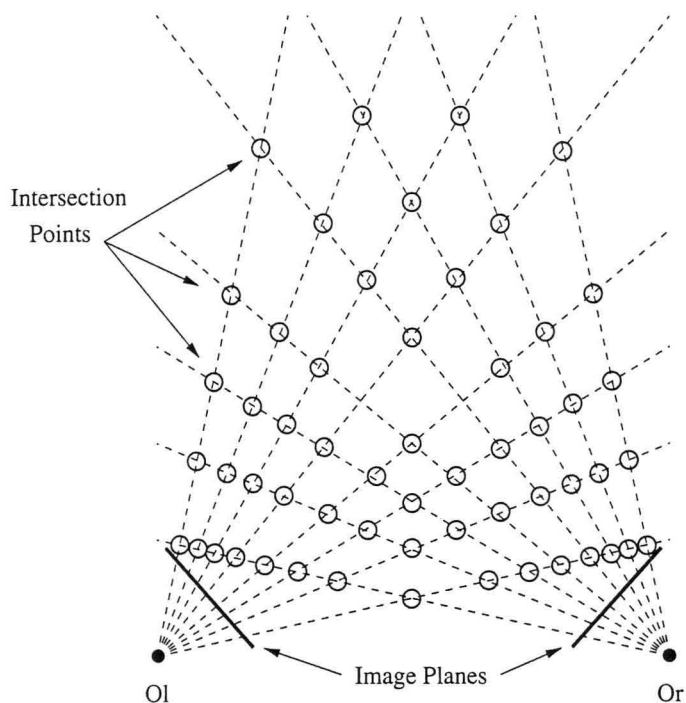
struction, because it causes the 3-D scene to be discretised by the sampling made by the cameras. The possible reconstruction points are located where the two lines, drawn from the center of projection  $O$  and the projection point in each camera model, intersect. Since the projection points are positioned at discrete locations (i.e. pixel locations) in the image, the intersections can only occur at discrete 3-D positions, Figure 2.6. To which degree scenes are discretised will obviously depend on the image resolution and size but also on the positioning of the cameras [15].

## 2.2 Camera Calibration

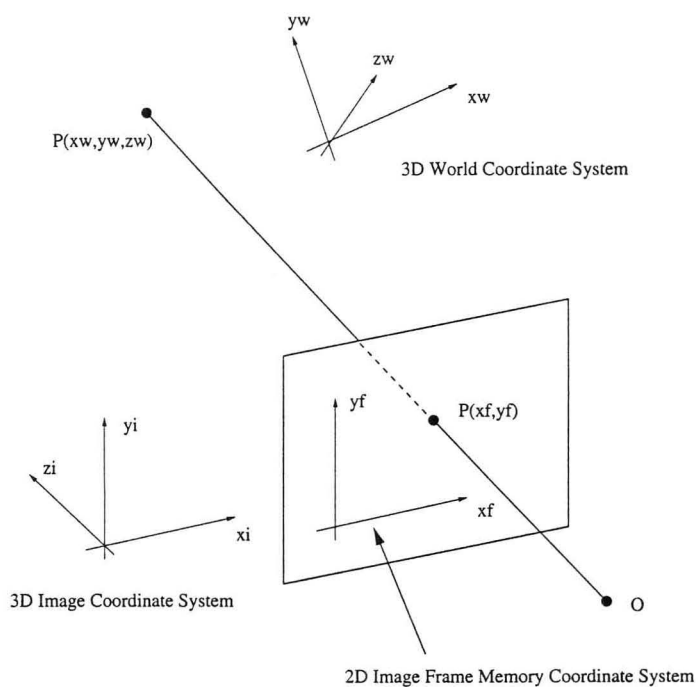
In order to make the representation of both three dimensional objects as well as two dimensional images pixels easier, additional coordinate systems are needed. The 2-D images are described in image frame coordinates,  $C_f$  and the 3-D scenes in real world coordinates  $C_w$ , Figure 2.7.

In order to be able to transform coordinates between the two systems, the cameras have to be calibrated. This is performed by making using a calibration object. Stereo images of this object are captured simultaneously by the cameras in the stereo setup. The calibration objects form is of little importance as long as it contains a number of points that can be

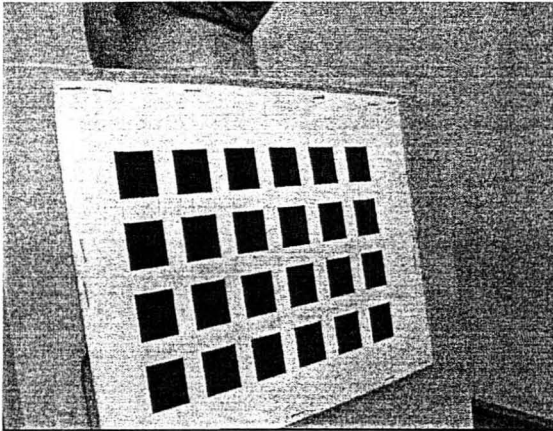




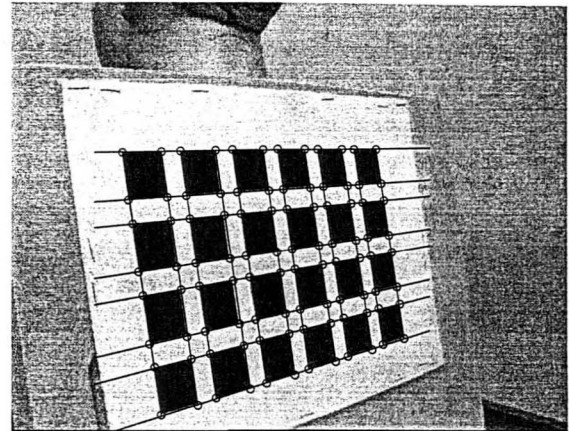
**Figure 2.6:** A demonstration of the discretisation of the reconstructable space.



**Figure 2.7:** Illustration of the camera geometry with the different coordinate system used.



**Figure 2.8:** The calibration object used.



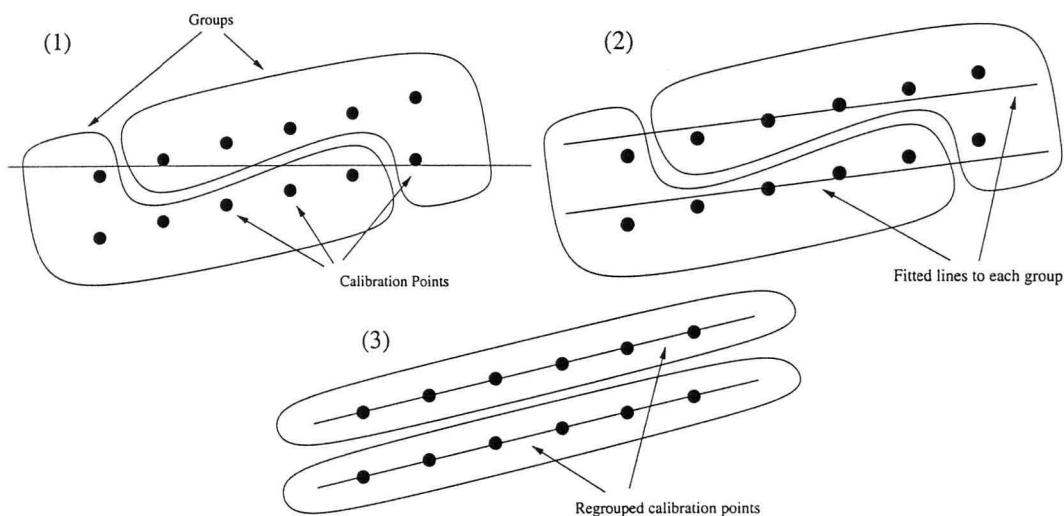
**Figure 2.9:** An image of a calibration object where the calibration points have been located and sorted.

easily located in the captured images, and whose position in real world coordinates have been accurately determined. The calibration object used here was a flat white surface with black boxes printed on it, (Figure 2.8). The corners of the boxes were used as calibration points. Using the camera calibration procedure suggested by Tsai [16], knowing only a few factory specific details of the cameras, both the extrinsic and intrinsic details can be determined with high accuracy.

### 2.2.1 Locating the Calibration Points.

The calibration points position in real world coordinates can easily be determined by physical measurement. If the position of the same points in the captured calibration image can be located Figure 2.11, the model that transforms the coordinates from one to the other with the smallest resulting error can be determined.

The calibration points, in this case the corners of a number of boxes, were located in the calibration image using basic image processing techniques of matched filters. The results of this corner location are however not sorted and it is crucial that the calibration points in the image are identified correctly. The real world position of the corners are known from



**Figure 2.10:** An illustration of the different steps of the calibration point sorting algorithm.

measuring the calibration object, this camera calibration is based on the assumption that the corresponding positions in the image of the same points are known.

Owing to the positioning of the calibration object as well as possible lens distortion it is not always the case that the calibration points are arranged neatly in horizontal rows oriented similarly to the pixel rows. The detected calibration points, corners, will therefore not always be arranged in horizontal rows oriented similarly to the pixel rows.

To ensure a correct indexing of the located points a simple sorting scheme was devised. Assuming that the calibration object consists of  $R$  rows and  $C$  columns of black squares. This will lead to  $M = 2R$  rows, each containing  $N = 2C$  points, It was furthermore assumed that these rows are roughly orientated horizontally.

The resulting sorting algorithm works in the following manner

1. First all the points are divided into  $R$  groups, each corresponding to a row, according to their position on the y-axis, Figure 2.10(1). This step will, based on the previous assumption on the orientation of the points, group most corners as belonging to the correct row.
2. To determine what points have been grouped wrong, a number of lines, one for each



row, are fitted to the points in each group Figure 2.10(2). Since the previous step was assumed to have sorted most points correctly, the line fitted to each group or row, will be orientated similar to the correct row orientation. Most points in each row will be on or close to their respective line, except for the wrongly sorted ones, they will lie closer to the line fitted to their correct group.

3. The points are then regrouped according to the line they are closest to, the  $R$  lines are refitted to these new groups, Figure 2.10(3). This step is repeated until the line fitting error is minimized.
4. To arrange the points in the correctly right-to-left, the algorithm is repeated from step 1, with the grouping being performed on the columns instead of the rows.

The output of this very robust calibration point locating procedure will accurately determine the location as well as the relative positions of all the calibration points, Figure 2.9.

## 2.2.2 Camera Parameter Estimation

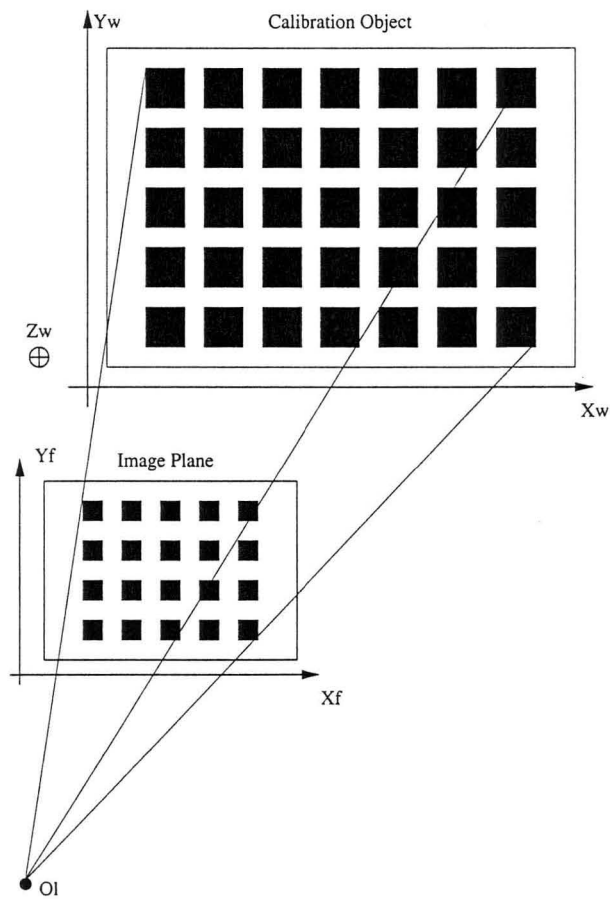
To be able to accurately calibrate a camera, a model must be designed to model how it maps 3-D coordinates onto 2-D coordinates. The complete camera model used by Tsai [16] is obtained by assuming that images are acquired by a sampling of a radially distorted perspective projection of an object. The transformation from 3-D world coordinates to 2-D image frame coordinates must be through such an optical system is carried out in four basic steps.

**Step 1 : 3-D world coordinates  $(x_w, y_w, z_w)$  to 3-D camera coordinates  $(x, y, z)$ .**

A rigid body transform given by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + T$$





**Figure 2.11:** The calibration object and its projection onto the image plane.

Parameters to be calibrated : the rotation matrix,  $R$ , and the translation vector  $T$ , both extrinsic parameters.

**Step 2 : Ideal undistorted 2-D image coordinates**  $(X_u, Y_u)$ . Perspective projection using pinhole camera model

$$\begin{aligned} X_u &= f \frac{x}{z} \\ Y_u &= f \frac{y}{z} \end{aligned}$$

Parameters to be calibrated : the focal length of the camera,  $f$ , intrinsic parameter.

**Step 3 : Distorted 2-D image coordinates**  $(X_d, Y_d)$ . Radial distortion modeling

$$\begin{aligned} X_d(1 + (\kappa_1 r^2 + \kappa_2 r^4 + \dots)) &= X_u \\ Y_d(1 + (\kappa_1 r^2 + \kappa_2 r^4 + \dots)) &= Y_u \\ r &= \sqrt{X_d^2 + Y_d^2} \end{aligned}$$

Parameters to be calibrated : The two intrinsic distortion parameters  $\kappa_1$  and  $\kappa_2$ .

**Step 4 : 2-D image frame coordinates**  $(X_f, Y_f)$  Sampling and acquisition.

$$\begin{aligned} X_f &= s_x d_x^{-1} X_d + C_x \\ Y_f &= d_y^{-1} Y_d + C_y \end{aligned}$$

where  $(X_f, Y_f)$  are the resulting pixel coordinates of the original point  $(x_w, y_w, z_w)$ .  $C_y$ ,  $C_x$ ,  $d_x$  and  $d_y$ , are camera specific, hardware related constants.

Parameters to be calibrated : the image uncertainty scale factor  $s_x$  (intrinsic)

To be able to separate the effects of the focal length  $f$  and the distance to the calibration points  $T_z$  simple geometry tells us that a perspective distortion is needed. All the calibration points must not be at an equal distance from the image plane. An in-depth rotation of the calibration object of  $30^\circ$  will provide adequate depth variation of the calibration points to ensure an accurate calibration.

With this knowledge of the dynamics of the optical system and with the locations of the calibration points in both the 3-D real world coordinates and 2-D image frame coordinates given, the parameters can be determined by implementing nonlinear optimization to minimize the error in transformation between the two coordinate systems <sup>1</sup>. This will result in a very reliable and accurate estimation of both the extrinsic and intrinsic camera parameters.

The 3-D real world coordinates will be given with the calibration object as reference. Since the relative position of the calibration object towards the cameras is not controlled the coordinate system references can vary considerably between calibration sessions. In an attempt to reduce these variations after every calibration the coordinate system reference is changed. The epipole and an average of the two image planes are used as references instead resulting in a 3-D real world coordinate system that does not depend on the positioning of the calibration object.

In an attempt to streamline the whole calibration procedure the calibration point location procedure was merged with the parameter estimation algorithm into a fully automatic stand-alone camera calibration application, Figure 2.12.

---

<sup>1</sup>A detailed description on the determination of these calibration parameters is given by Tsai in [16].



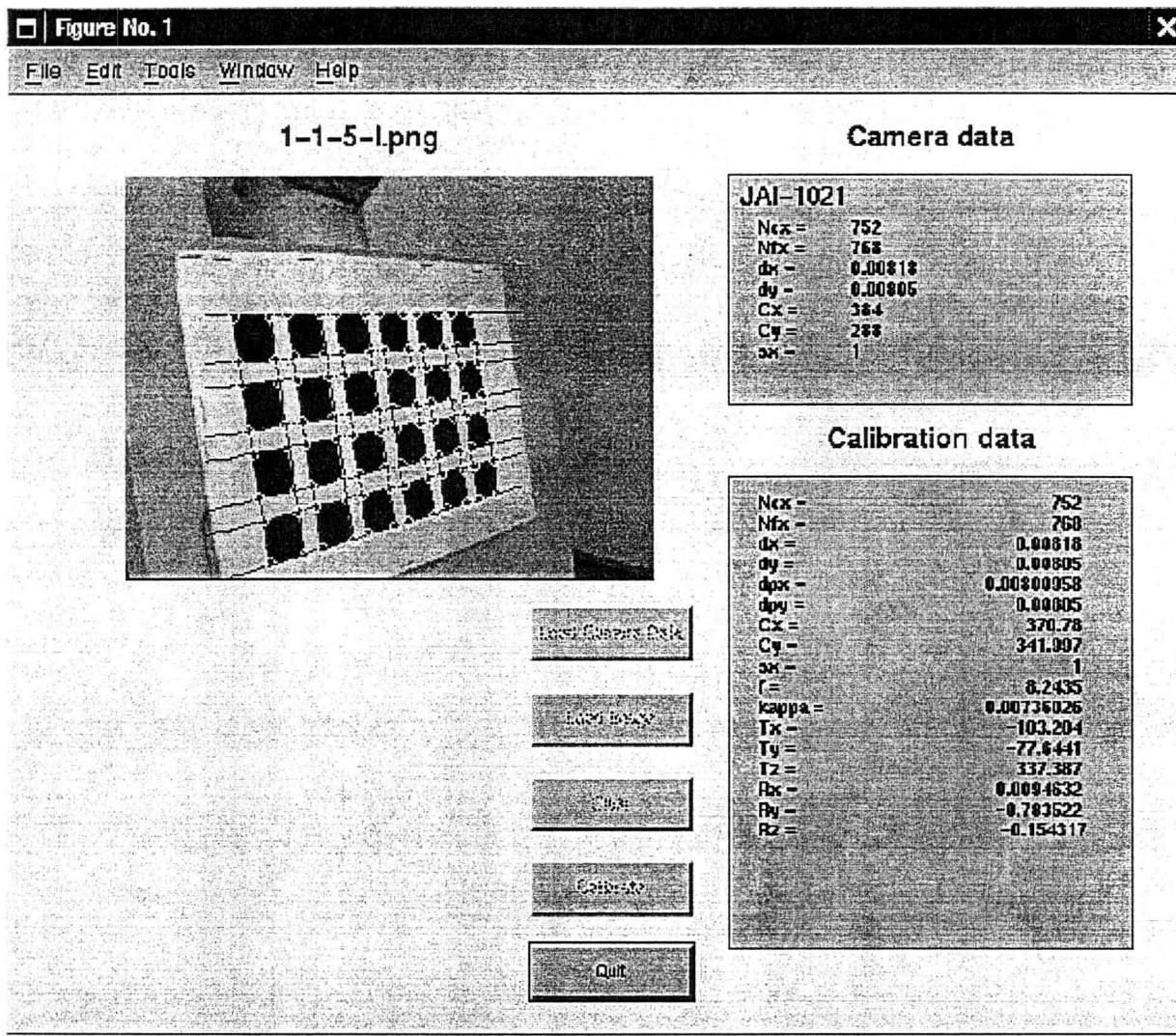


Figure 2.12: The finished camera calibration GUI for single images.



# Chapter 3

## Face Recognition

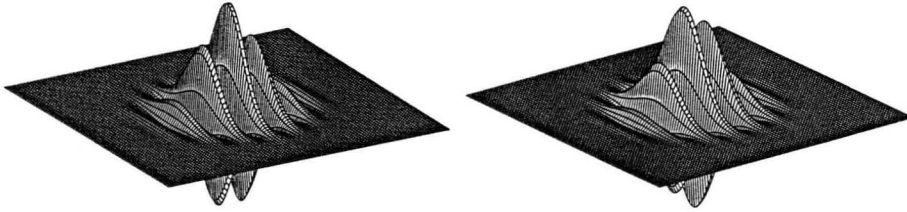
This chapter details the novel face recognition system presented in this thesis. The two main components of the system, the training and testing procedure will be discussed along with underlying theories that are necessary for their implementation. This includes the Gabor feature vectors and the related processing, facial landmark location, stereo matching and model mapping.

### 3.1 Gabor Features

A 2-D Gabor wavelet is a complex valued sine modulated Gaussian function, that can be tuned to specific frequencies and orientations, [17], Figure 3.1.

$$\begin{aligned} \psi_{\vec{k}}(\vec{x}) &= \frac{\vec{k}^2}{\sigma^2} \exp\left(-\frac{\vec{k}^2 \vec{x}^2}{2\sigma^2}\right) \left[ \exp(-i\vec{k}\vec{x}) \right. \\ &\quad \left. - \exp\left(-\frac{\sigma^{-2}}{2}\right) \right] \\ \vec{k}_j &= \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos(\varphi_\mu) \\ k_\nu \sin(\varphi_\mu) \end{pmatrix} \end{aligned} \quad (3.1)$$

Gabor filters can achieve simultaneous localization in space and in spatial frequency, which



**Figure 3.1:** The real (left) and imaginary (right) parts of a Gabor filter.

Wavelengths, $\lambda$	8, $8\sqrt{2}$ , 16, $16\sqrt{2}$
Orientations, $\varphi_\mu$	0, $\frac{\pi}{8}$ , $\frac{\pi}{4}$ , $\frac{3\pi}{8}$ , $\frac{\pi}{2}$ , $\frac{5\pi}{8}$ , $\frac{3\pi}{4}$ , $\frac{7\pi}{8}$

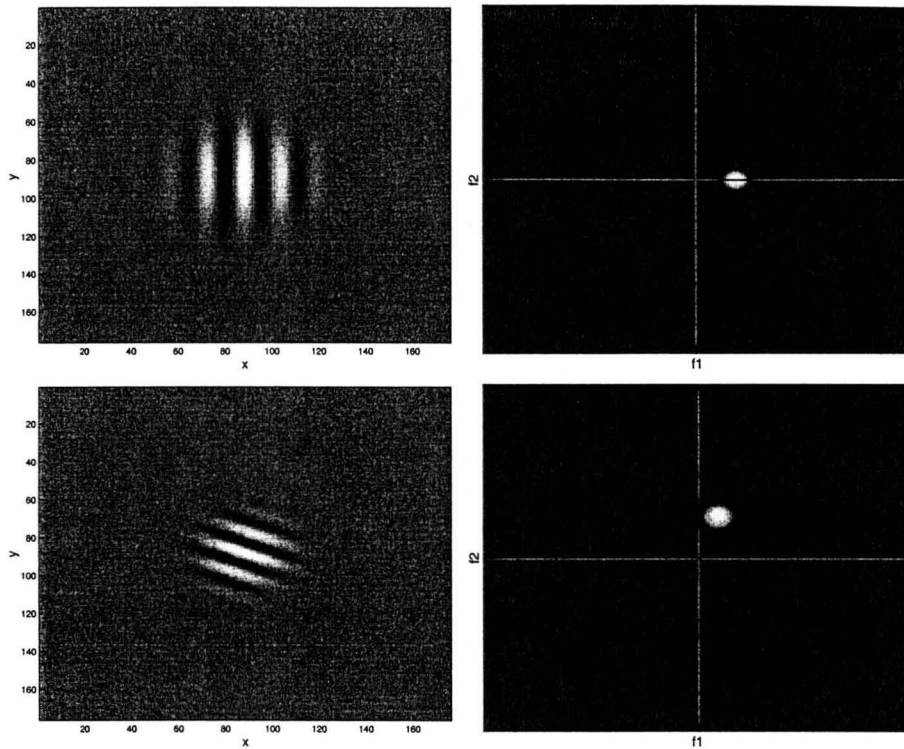
**Table 3.1:** The different wavelengths and orientations of the Gabor filters used by our system

means they are an excellent choice for analyzing the spatially localized frequency content of images. A properly tuned filter can be used as a correlation filter to look for energy in an image at a particular frequency and orientation, Figure 3.2.

At each facial landmark the image is sampled in frequency to produce a feature vector or label. This frequency sampling is conducted by convolving the images with a number of Gabor filters tuned to different frequencies and orientations determined by the wave vector  $\vec{k}$ , (3.1), Figure 3.3. The wavelets were also made DC-free to ensure invariance to changes in illumination. In this system 4 different wavelengths, spaced half an octave apart, and 8 different orientations were used, resulting in an even coverage of the frequency space Figure 3.4 by a filter bank containing 32 Gabor filters, Table 3.1.

The convolution is, in an attempt to speed up this preprocessing stage, carried out in the frequency domain by Fourier transforming the images, multiplying them with the stored transforms of the filterbank and then transforming the result back into time domain again.

The outputs from the convolution are used to form the feature vectors, or *jets*. A jet  $\vec{J}_i$  is defined as a vector containing the N complex-valued outputs from the Gabor filter bank



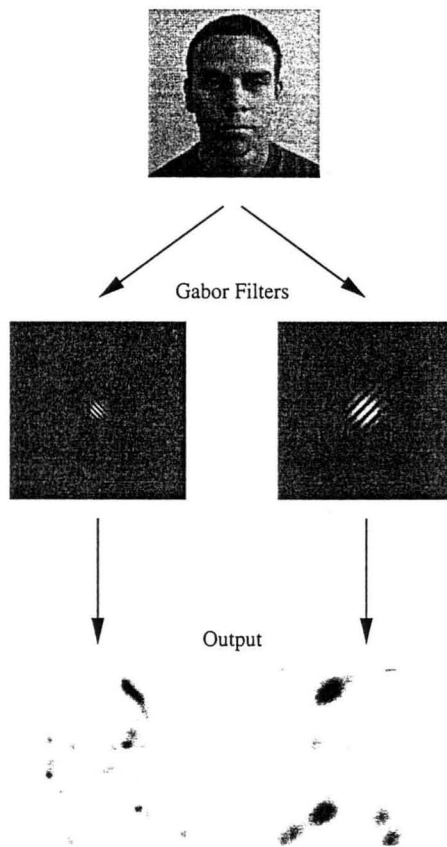
**Figure 3.2:** The absolute value of two example Gabor filters in the spatial domain, left, and in the frequency domain, right. Observe how the different orientation and scale of the filters determine the frequency domain positioning.

at pixel  $i$  in the image, Figure 3.5. We utilize these jets to perform facial feature location, stereo matching and recognition.

## 3.2 Similarity Functions

To be able to compare different jets  $\vec{J}_i$  and  $\vec{J}_i'$  with each other, we must define some type of similarity measurement. Three such similarity functions have been described by Wiskott *et al.* [5].

The  $k$ 'th element ( $k = 0, 1, \dots, N-1$ ) of the jets can be written as  $J_{ki} = a_{ki}e^{j\phi_{ki}}$ , where  $a_{ki}$  denotes amplitude and  $\phi_{ki}$  the phase of the  $k$ 'th element of the jet of the  $i$ 'th facial feature. The amplitude will vary slowly with position and the phase will rotate predominately with the center frequencies of the Gabor filters.



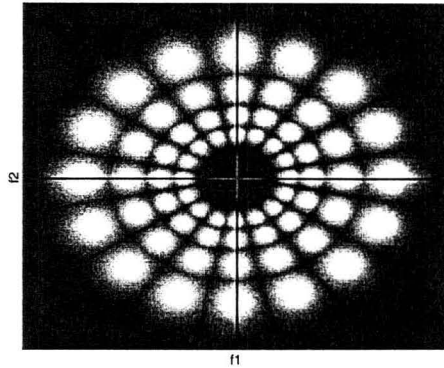
**Figure 3.3:** The filtering of a face with two different Gabor kernels. Note how the two filters light up different parts of the face, they are both sensitive to features with an orientation and scale similar to their own.

Jet coefficients with an amplitude lower than 5% of the maximal amplitude are set to zero, because the phases of such small outputs are poorly defined and numerically unstable.

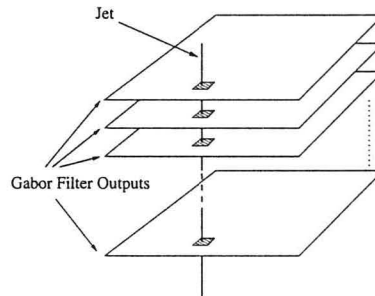
The comparative analysis of different jets will, to a very high degree, determine the performance of the whole system. Therefore a lot of emphasis is placed on the investigation and utilization of different methods of measuring jet similarities. In this work three different similarity functions will be discussed, the absolute, the phase and the displacement-estimation based similarity function.

The *absolute value based* similarity function  $S_a(\vec{J}, \vec{J}')$  is defined as the normalized dot product between the absolute values of  $\vec{J}_i$  and  $\vec{J}'_i$  and was first used by Buhmann et al. [18].





**Figure 3.4:** The position in the frequency domain of the 32 Gabor filters used by our system.



**Figure 3.5:** A jet containing the  $N$  complex outputs from the filtering procedure.

The phase is disregarded all together.

$$S_a(\vec{J}_i, \vec{J}_i') = \frac{\sum_{k=0}^{N-1} a_{ki} a'_{ki}}{\sqrt{\sum_{k=0}^{N-1} a_{ki}^2 \sum_{k=0}^{N-1} a'_{ki}{}^2}}. \quad (3.2)$$

This similarity function has, due to the slowly varying nature of the jet amplitudes, proven to be very robust to changes in facial expression, pose and scale. It is however, for the same reasons, not very accurate in locating matching jets.

The second approach to jet similarity measurement uses the phase  $\phi_{ki}$ , as well as the amplitude of the jets in the matching. The incorporation of phase into the similarity functions is a complicated matter, owing to the rapidly changing nature of the phase it will improve accuracy but also be less robust. The basic *phase based* similarity function is

defined as

$$S_\phi(\vec{J}_i, \vec{J}'_i) = \frac{\sum_{k=0}^{N-1} a_{ki} a'_{ki} \cos(\phi_{ki} - \phi'_{ki})}{\sqrt{\sum_{k=0}^{N-1} a_{ki}^2 \sum_{k=0}^{N-1} a'_{ki}{}^2}}. \quad (3.3)$$

However, since the phase  $\phi_{ki}$  rotates with the same rate as the center frequencies of the Gabor filter such a straight comparison of jets will result in a very erratic and rough similarity function. It is however the least computationally expensive method.

One way of improving the characteristics of a phase based similarity function was suggested by Theimer and Mallot [19] and involves compensating for this phase rotation by subtracting the phases of the jets with the product of an estimation of the jet displacement  $\vec{d}$  and the Gabor center frequencies  $\vec{k}_j$  resulting in the *displacement based* similarity function

$$S_d(\vec{J}_i, \vec{J}'_i) = \frac{\sum_{k=0}^{N-1} a_{ki} a'_{ki} \cos(\phi_{ki} - \phi'_{ki} - \vec{d} \vec{k}_j)}{\sqrt{\sum_{k=0}^{N-1} a_{ki}^2 \sum_{k=0}^{N-1} a'_{ki}{}^2}}. \quad (3.4)$$

Since the center frequencies for the different Gabor filters are known, an estimation of the spatial displacement  $\vec{d}$  is all that is needed to cancel out the phase rotation. This is accomplished by finding the displacement that maximizes  $S_d$  in its Taylor expansion.

$$S_d(\vec{J}_i, \vec{J}'_i) \approx \frac{\sum_{k=0}^{N-1} a_{ki} a'_{ki} \left[ 1 - 0.5 (\phi_{ki} - \phi'_{ki} - \vec{d} \vec{k}_j)^2 \right]}{\sqrt{\sum_{k=0}^{N-1} a_{ki}^2 \sum_{k=0}^{N-1} a'_{ki}{}^2}}. \quad (3.5)$$

by setting  $\frac{\partial}{\partial x} S_\phi = 0$  and  $\frac{\partial}{\partial y} S_\phi = 0$  and solving for  $\vec{d}$  [5] results in

$$\vec{d}(\vec{J}_i, \vec{J}'_i) = \begin{bmatrix} d_x \\ d_y \end{bmatrix} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \begin{bmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{bmatrix} \begin{bmatrix} \Phi_x \\ \Phi_y \end{bmatrix}. \quad (3.6)$$

$\Gamma$  and  $\Phi$  can be computed using

$$\Phi_x = \sum_j a_j a'_j k_{jx} (\phi_j - \phi'_j)$$

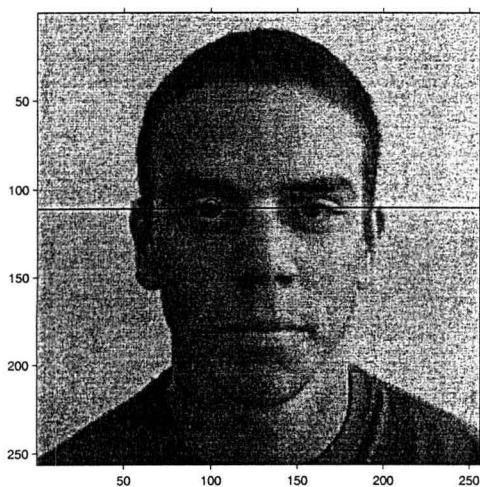
$$\Gamma_{xy} = \sum_j a_j a'_j k_{jx} k_{jy}$$

If  $\begin{bmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{bmatrix}$  is not singular then  $\vec{d}(\vec{J}_i, \vec{J}'_i)$  can be computed. With this estimation method, the displacements of two jets can be determined up to a distance of half the shortest wavelength of all the Gabor filters. This limitation arises owing to the low order of the Taylor expansion of cosine used, it is only valid between  $-\pi$  and  $\pi$ . The range of the displacement estimation can be extended however by first using the filters with the longest wavelength and the gradually including filters of higher and higher frequency in a coarse-to-fine manner. When advancing to a higher frequency level, the previous estimate is compared with the wavelength of the added frequency level. This will allow for possible corrections of the jet coefficients of the higher frequency filters by multiples of  $2\pi$ . With this added procedure, the maximum displacement can be as large as half the wavelength of the lowest frequency instead of the highest.

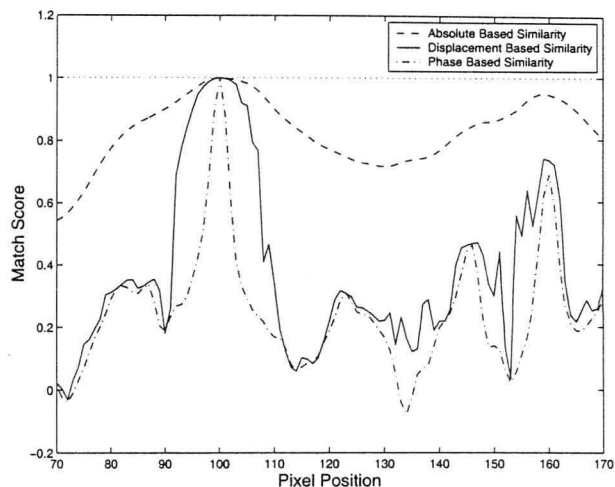
A comparison between these different similarity functions can be seen in Figure 3.7. A jet  $J_{eye}$  located at the left eye of a test subject is compared with all the jets  $J'_i$  with the same horizontal position as  $J$ , Figure 3.6. This example clearly shows the accuracy of the different methods with the absolute valued one being the least accurate and the phase based function displaying very good results.

Performance under ideal circumstances is not the only issue that has to be taken into consideration, actually more important is a methods ability to work under largely varying conditions. Figure 3.8 illustrates this. The left eye jet  $J_{eye}$  is now extracted from the right image, instead of the left image which was the case in Figure 3.7, in the stereo image pair. The absolute based approach that showed poor accuracy previously, displays a very high degree of robustness to image variations. The displacement method is also fairly invariant to changes in pose, the maximum similarity value only drops a few percent compared to





**Figure 3.6:** Illustration of the positions of the extracted jets. The eye jet  $J_{eye}$ , marked by a circle, is compared from left to right with all the jets positioned on the horizontal line,  $J'_i$ .



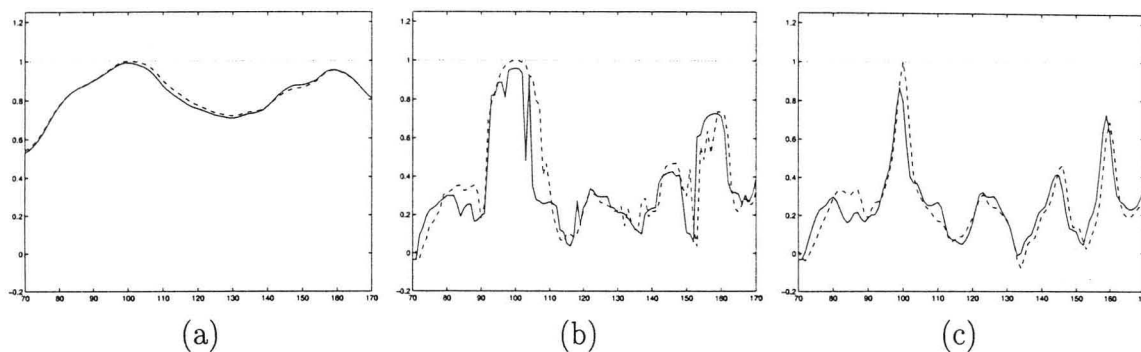
**Figure 3.7:** The outputs of the three different similarity functions when comparing jets extracted from Figure 3.6. Note the local maxima at  $x \approx 160$ , the position of the right eye.

the previous instance. Finally it can be observed that the phase based does not fare well with changes in the image, the similarity score is reduced by 15%. This constitutes a considerable reduction compared to the other two similarity measurements that only were reduced by 0.5% and 4% respectively.

The shape of the similarity functions must also be taken into account as it will have some influence on both the overall performance of the system as well as processing time. The smoother a curve is and the fewer local minima and maxima it shows the faster and the more accurate the location of the jets becomes. The absolute based similarity function possesses just these properties.

In conclusion, there are three methods available for comparing different jets all with advantages and disadvantages of their own. The absolute based function is very robust but less accurate, the phase based method is very accurate but extremely sensitive to jet changes, it is also the fastest method of them all. The displacement based similarity function, is a compromise between these two, thus it is moderately accurate and robust method of jet





**Figure 3.8:** The resulting similarities of a jet taken from the left eye of the same subject, but from another image, compared with the jets along a line at the same horizontal position using the three methods described. The results are compared with the similarity values from Figure 3.7, (dashed), for (a) the absolute based-, (b) the displacement based and (c) the phase based similarity function. In this figure the robustness of the different approaches can clearly be seen.

comparison, it does however require a lot of computational power.

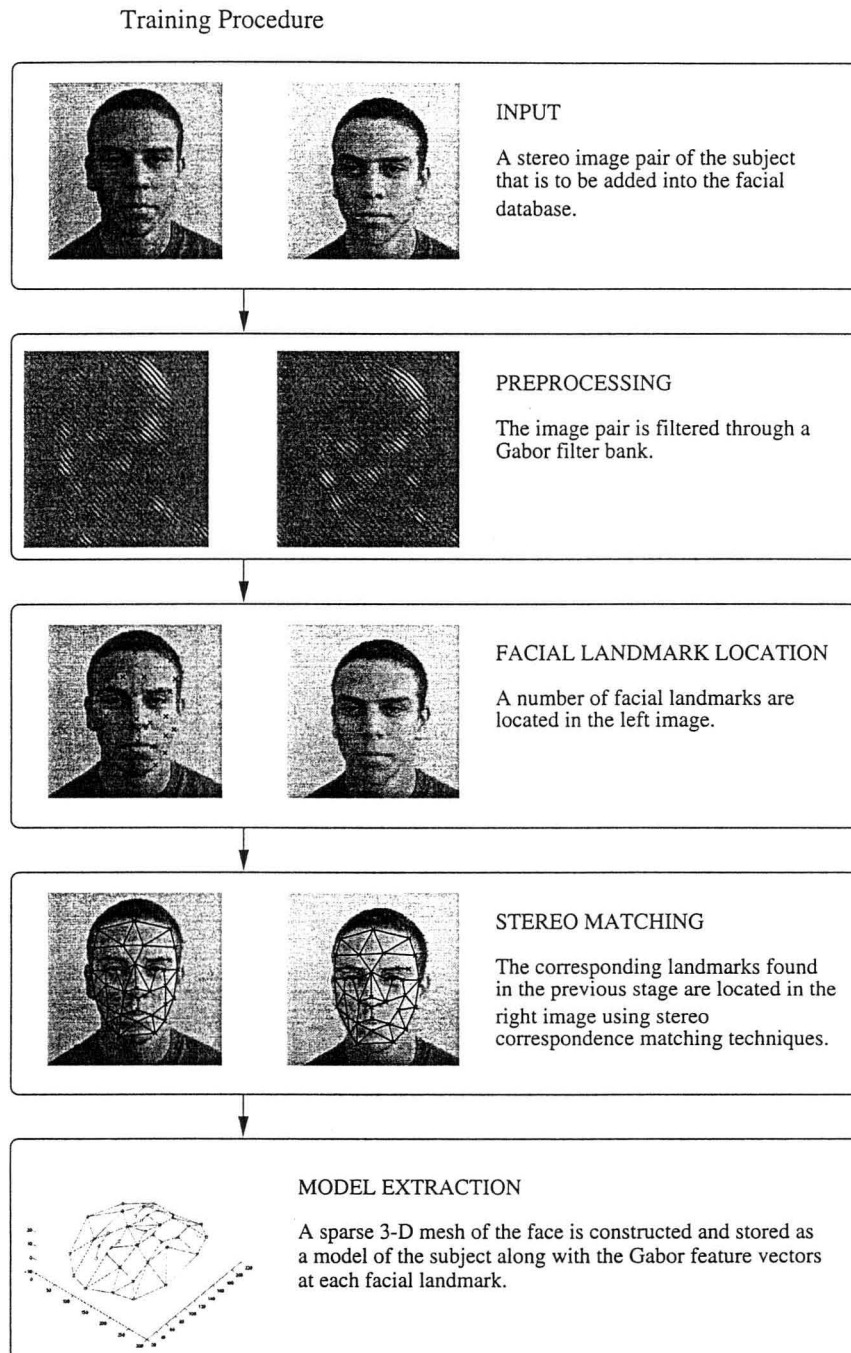
The similarity functions are therefore, owing to these different characteristics, employed depending on the circumstance.

## 3.3 Training

From each facial image presented to the training procedure a model or template is extracted, Figure 3.9. This model will be made up of the 3-D position of a number  $L = 40$  of facial landmarks along with the 32-dimensional jets obtained at each of these landmarks. In order to carry out this task, this procedure must be able to perform feature point location, stereo correspondence matching as well as 3-D reconstruction.

### 3.3.1 Facial Feature Points

The face is sampled spatially at a  $L = 40$  number of typical facial landmarks Figure 3.10, e.g the corners of the eyes and the pupil, the tip of the nose etc. The landmarks are located by searching the face for example jets of each facial feature point. These example jets are



**Figure 3.9:** A more detailed block diagram of the training procedure.



**Figure 3.10:** A face marked with the location of the facial landmarks used by the system.

acquired by manual marking of a  $G = 50$  number of training faces. These training faces were chosen from a garbage class containing subjects that only attended one single photo session.

This process can be a very computationally expensive one if the whole image has to be searched  $G$  times for each of the  $L$  facial landmarks. We therefore make use of the phase based similarity function  $S_\phi$  to find an initial positioning of the jets that can later be refined by other more complex means, its drawbacks of poor robustness is reduced on two accounts. First the large number of example jets used will increase the chance of a correct match. Second, not only the position of the best matched example jet is used, the  $S = 5$  best matches are taken into account. These  $S$  positions are weighted according to their matching score and averaged to used as an initial position of that facial feature point.

These results are then passed on to a refining stage that employs the displacement based similarity function  $S_d$  on the twenty best matches from the previous stage resulting in a facial feature point location procedure that has produced very satisfactory results.

### 3.3.2 Stereo Matching

Once the facial feature points have been located in the left image, stereo matching is performed. The task of disparity estimation has always been a complicated one and when

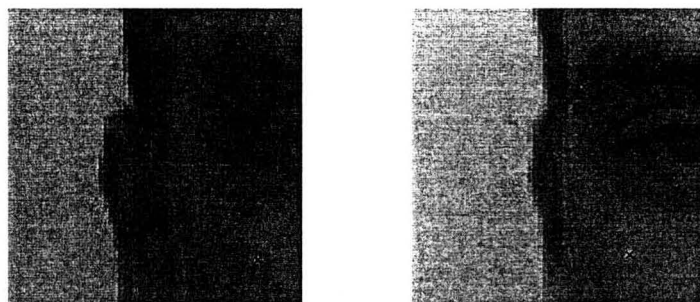


concerning human faces it is even more so.

Most existing systems [12, 13] deal with binocular stereo scenes that consist of a number of separable objects with differing distances to the image plane. For such instances the task is then to determine how much the objects have shifted and to locate singularities. A binocular stereo image pair of a human subject will show one single continuous object, a face, with very little or no occlusion. Owing to its customary convex shape, the edges of the faces will be distorted, one side compressed and the other one expanded horizontally. Faces are also difficult to process for another reason, they have large areas with low texture, eg. the forehead and the chins, that makes the correspondence matching even harder.

The matching is performed by searching in the right image for the best match to jets extracted from the facial feature points in the left image along their respective epipolar lines using the displacement based similarity function  $S_d$ . This function is chosen for its combined properties of robustness and accuracy, the phase based method proved to be too unreliable for this task.

The problem with feature points along the perimeter of the face is addressed by altering the filters used in the matching based on kernel size. It is obvious that larger filters will have difficulties in making out smaller image features and that small filters will not be able to utilize the more global information in the image. For a feature point located at for example the top of the left ear (Figure 3.11), assuming that occlusion does not occur, the change in viewpoint will result in that part of the image being contracted horizontally.

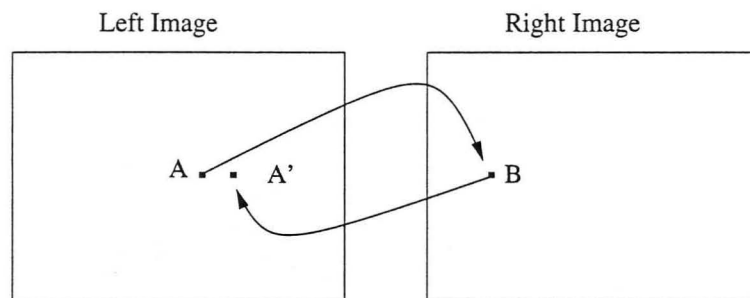


**Figure 3.11:** A closer look at a feature point located close to the top of the left ear.



this point is to be located in the right image, the stereo matching procedure will run into problems right away. The jet from the left image has registered a feature point that, according to the larger filter outputs, is placed a certain fraction of a wavelength from the left edge of the head and at the same time, now according to the smaller filters, is placed on a small strip of hair between two areas of skin. As the specific area where this feature point is located in the right image has been contracted horizontally its position now seems to be closer to the edge of the head but still placed on a thin strip of hair. The stereo matching algorithm now has to decide on the location of this feature point based on this contradicting information, the result is in most instances completely wrong. The solution to this particular problem is to remove the jet coefficients originating from larger filters and letting the matching be more based on the smaller more detail sensitive Gabor filters. However this can not be applied to all feature points, low textured landmarks need the globalised information that the larger filters provide to be correctly located. Performance is in this case reduced considerably if the large Gabor filter outputs are discarded. The feature points in the center part of the face will also be affected by such an act, even though these part of the image will not be distorted much, the accuracy of their placement will also be reduced by the removal of information.

So in order to determine if and when to remove the larger scale features from the jets and when not to, *back matching* or *left-to-right consistency* is implemented, Figure 3.12. This will allow for some degree of controlling how good a correspondence match really is.



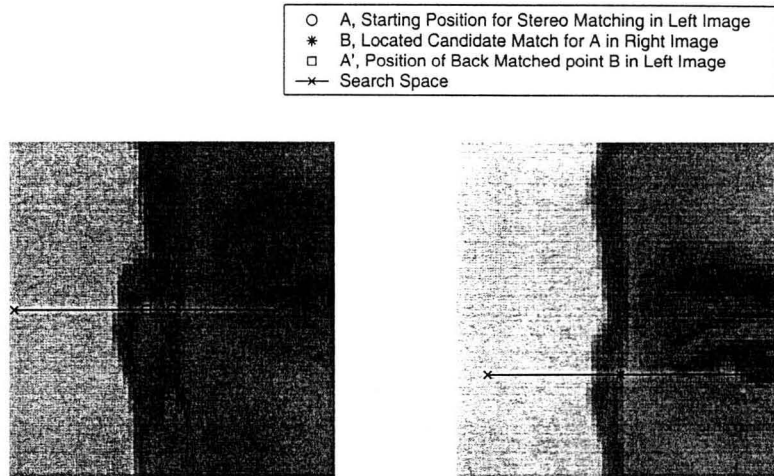
**Figure 3.12:** Point  $A$  in the left image is matched to point  $B$  in the left image.  $B$  is then matched back into the right image on to point  $A'$ . The match  $A \mapsto B$  is deemed left-to-right consistent if  $A - A'$  is smaller than a threshold  $\epsilon$ .

The stereo matching is performed by recursive jet reduction and back matching.

1. Start off by using all of the filter outputs and all of the facial feature points
2. Search along the epipolar lines in the right image for the best match  $B$  of the left image facial feature points  $A$  using the displacement based similarity function  $S_d$ .
3. Perform a back matching by reversing the matching order to right-to-left instead of left-to-right; Search along the epipolar lines in the left image for the best matched points  $A'$  of the results in the previous step  $B$  using the same similarity function.
4. Compare the resulting back matching points  $A'$  with their original position  $A$ . Feature points that are matched back correctly,  $A = A'$ , are marked as matched and stored.
5. Remove the jet coefficients originating from the largest filters, thus removing the larger scaled jet coefficients that contain more of the global information and allowing for a more detail sensitive matching. Since there are 8 orientations, all the jets will be reduced by 8 coefficients every time this reduction step is carried out.
6. Repeat steps 2-5 until no feature points remain unmatched or the jets can not be reduced further, this will occur after 4 iteration since there are 4 scales.
7. If a landmark still remains unmatched, choose the match that resulted in the smallest back matching error.

The gradual removal of the larger scales and the continuous consistency check ensures that the feature points with low distortion use all the available jet coefficients and that low textured landmarks can make use of the global information in the larger Gabor filters. It allows for better location of the distorted facial landmarks by the combinatorial testing of different filter size combinations during the matching.

By carrying out the stereo correspondence matching in this manner the system manages to locate all types of facial landmarks Figure 3.14 with sufficient accuracy.



**Figure 3.13:** The desired result of the stereo correspondence matching procedure.



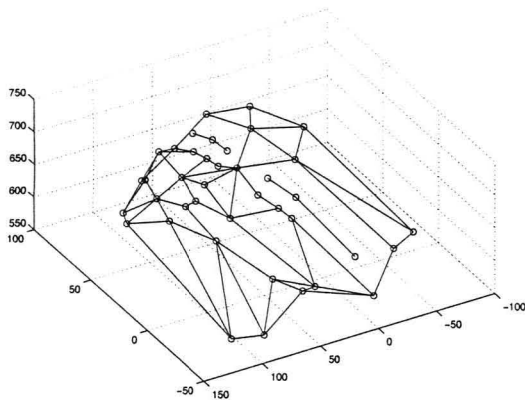
**Figure 3.14:** The desired result of the stereo correspondence matching procedure.

### 3.3.3 Model Extraction

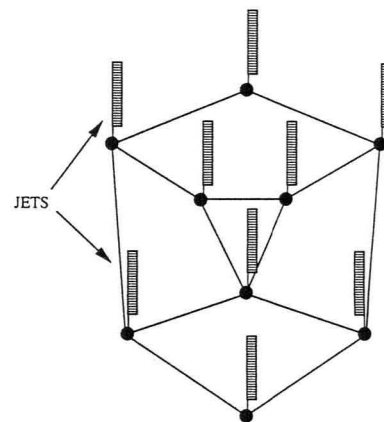
The final stage in the training procedure is the model extraction. It is here that the 3-D model of the face is constructed and stored, along with texture information in the form of jets, to make up the models later used for recognition.

3-D reconstruction is performed on the output from the stereo correspondence mapping procedure, the positions of all the 40 facial feature points in both images. The stereo





**Figure 3.15:** The reconstructed 3-D facial mesh.



**Figure 3.16:** The extracted model, the sparse 3-D facial mesh with the corresponding jets of each landmark node.

system is calibrated in the way described in Section 2.2. Using this camera calibration information the location of the facial landmarks in real world 3-D coordinates, Figure 3.15, can be easily be determined, Section 2.1.

The final model, Figure 3.16, will contain the following information

- The 3-D mesh of the face, Figure 3.15, i.e. the 40 facial landmark coordinates, referenced to the epipole.
- The jets at each of the facial feature points in the mesh, from both the left and the right image.
- The two images. They are not necessary for recognition purposes but are included so that manual verification of the recognition results can be carried out.

The models requires about 42 kB of storage each (excluding images). There are ways of reducing this requirement if necessary. If instead of storing the jets extracted from both images, only the average jet from the left and right image at each facial feature is stored, this will reduce the model size by 20 kB. Further compression can be achieved by clustering the jets using the regularities of the jet coefficients suggested by Krüger *et al.* in [20]. Such a clustering was also reported to reduce processing time.



### 3.3.4 Weighting of Features

It is quite obvious that different facial features are of different importance when humans perform face recognition, it is unsurprising that automatic face recognition also benefits from this. An analysis of the different facial landmarks influence on recognition performance was carried out using an algorithm developed by Krüger [21].

To be able to examine the influence of the weights, an evaluation function was firstly introduced

$$Q(T, w_1, \dots, w_n) \quad (3.7)$$

where  $T$  is the training set used to train the weights, the validation set  $V$  was utilized for this purpose,  $w_1, \dots, w_n$  signifies the individual weights for each of the  $n$  landmarks. In this work the recognition rate of the validation set was used as the evaluation function. When optimizing this evaluation function Krüger reported problems with generalization, a significant improvement in recognition rate on the training set was obtained but, using the same weights, the recognition rate on the test set decreased considerably.

In an attempt to reduce the dimensionality of the problem a different approach was proposed. It is based on the idea that the features with high discriminatory powers should be weighted during recognition. It can be assumed that the similarity between two jets from the same person are in general high and for jets of different persons lower. Under this hypothesis a facial feature can be classified as discriminatory if

$$\begin{aligned} S(\vec{J}_i^k, \vec{J}_k^l) \text{ is large for } k, \text{ and} \\ S(\vec{J}_i^k, \vec{J}_i^l) \text{ is small for } l \neq k \end{aligned}$$

This can be combined into

$$\Delta(k, l)_i = S(\vec{J}_i^k, \vec{J}_i^k) - S(\vec{J}_i^k, \vec{J}_i^l) \text{ is large for } l \neq k$$

$\Delta(k, l)_i$  is defined as the discriminatory power of feature  $i$  for persons  $k$  and  $l$ .

Using this, the total discriminatory power, the significance of landmark  $i$  was defined in [21] as

$$P_{disc,i} = \left( \sum_k \frac{1}{n} \sum_{l \neq k} \frac{1}{n} \arctan(\alpha_1 \cdot \Delta(k, l)_i) \right).$$

The arctan function is present to reduce the influence of extreme values of  $\Delta(k, l)_i$ .  $P_{disc,i}$  can be viewed as the average of the discriminatory power of feature  $i$  of all the subjects.

The weighting assigned to each facial landmark is finally defined as a function of that specific landmarks significance

$$w_i = \max(0, P_{disc,i})^{\alpha_2} \quad (3.8)$$

The evaluation function has now been reduced to a function of two variables,  $\alpha_1$  and  $\alpha_2$ .

$$Q(V, \alpha_1, \alpha_2) \quad (3.9)$$

This dimensionality reduction both solves possible generalization problems and reduces the optimization search space for the evaluation function.

An exhaustive search was performed to optimize the evaluation function on our data and to calculate the resulting landmark weights, Figure 3.17. This figure illustrates the different landmarks significance for face recognition, it can be seen how the facial features located in the center of the face is more important, with special emphasis on the region around the eyes. A result that intuitively seems to be correct.

### 3.4 Testing

In the testing procedure, image pairs of unknown faces are analysed for classification or verification. This procedure calculates a similarity value between an unknown face and a stored model. This comparison is performed, not by extracting a new model from the



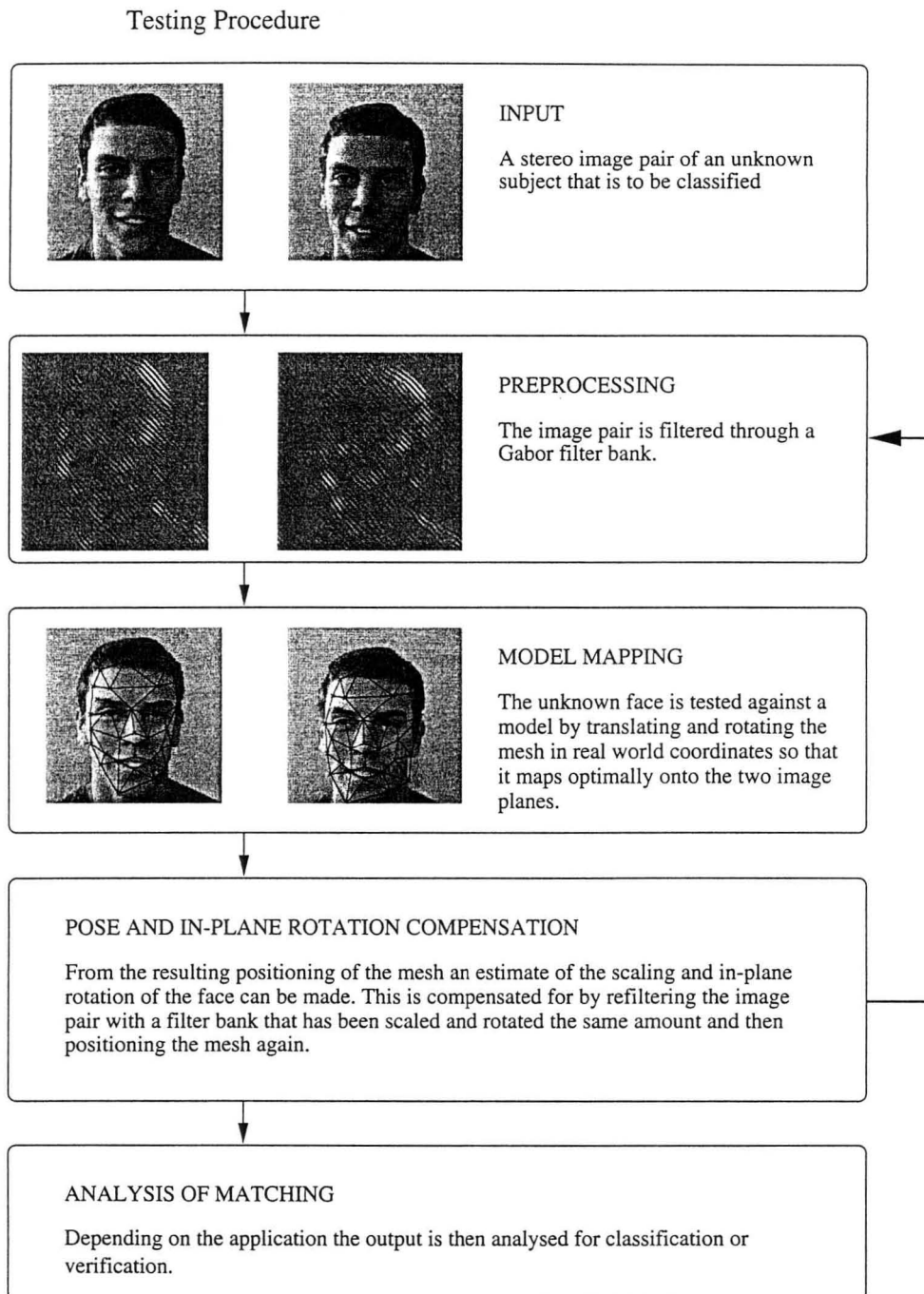
**Figure 3.17:** The resulting weighting of the facial landmarks. The higher significance of the eye regions indicates that an important part of the discriminatory information is located there.

image pair but by fitting the model onto the image pair using the camera calibration information, Figure 3.18. This approach was chosen mainly for a number of reasons, firstly it allows for carrying out testing on single images. Secondly, it makes the performance of the facial feature location less essential to the performance of the system. To perform recognition by extracting a new model from the test image pair the facial feature must be located very accurately. If the suggested approach is employed instead, a facial landmark that was badly located during training will not affect recognition performance to the same extent. The texture and stereo information extracted from the face is still correct but since recognition is performed by mapping the nodes of the mesh onto the unknown face and comparing the jets, incorrectly placed nodes will not affect performance considerably.

As in the training procedure, the image pair is initially filtered through a Gabor filter bank to allow for the extraction of the jets in the two images.

### 3.4.1 Optimal Mapping of the Model

The next step is the matching of the model with the unknown face and this is carried out by transforming the position of each landmark node in 3-D coordinates to the 2-D image coordinates for both image planes using the camera calibration information for the test



**Figure 3.18:** A more detailed block diagram of the testing procedure.



setup.

$$P_f = f(C, P_w) \quad (3.10)$$

where  $P_f$  is the position of the facial landmarks in 2-D frame coordinates,  $P_w$  is the position in 3-D real world coordinates,  $C$  the camera calibration information and  $f$  the 3-D real world to 2-D frame coordinates transformation function described in Section 2.2.2.

The model jets  $\vec{J}$  are then compared with the jets extracted at the points where the mesh nodes are projected on the images  $\vec{J}'$ . The extracted jets  $\vec{J}'$  will be a function of the position of the facial landmarks in 2-D frame coordinates  $P_f$  and the image  $I$  they are mapped onto.

$$\vec{J}' = g(I, P_f) \quad (3.11)$$

The similarity between a model with jets  $\vec{J}$  and an unknown face with extracted jets  $\vec{J}'$  is defined as the weighted sum of the individual jet similarities

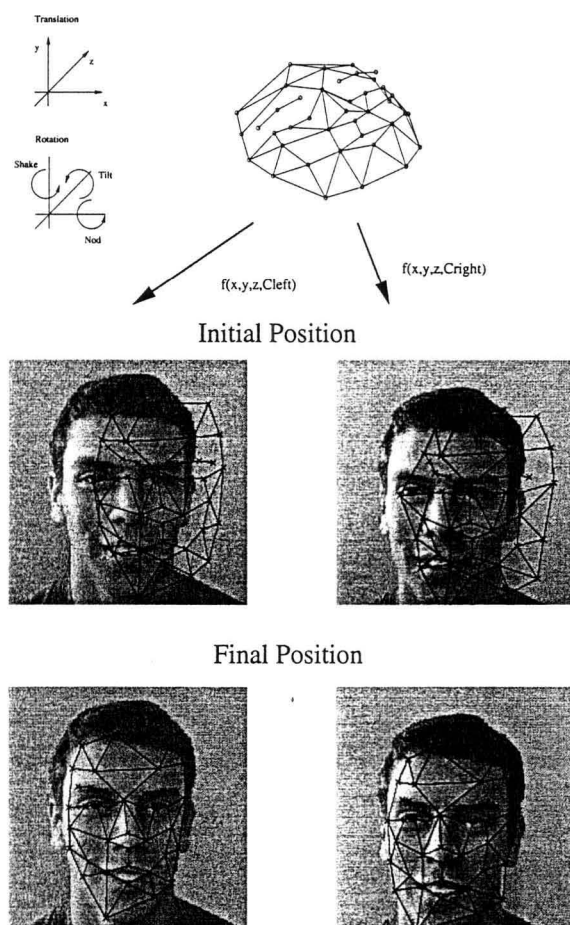
$$S(M, I) = \sum_{i=0}^{L-1} w_i S_{a,d}^i(\vec{J}_i, \vec{J}'_i); \quad (3.12)$$

It is more than likely that the subjects pose is different from the training session, so to ensure pose independence, the mesh must be translated and rotated to optimize the mapping, Figure 3.19.

This rotation and translation of the mesh will change the position of the facial features in 3-D real world coordinates

$$P_w = R(\theta, \phi, \psi)P_0 + [x, y, z] \quad (3.13)$$

$P_0$  is the initial position of the mesh, extracted during the training procedure. The rotation



**Figure 3.19:** The mapping of a model onto an image pair with and without the rigid transformation that constitutes the pose compensation.

$R$  matrix is defined as

$$R = \begin{bmatrix} \cos \psi \cos \theta & \sin \psi \cos \theta & -\sin \theta \\ -\sin \psi \cos \phi + \cos \psi \sin \theta \sin \phi & \cos \psi \cos \phi + \sin \psi \sin \theta \sin \phi & \cos \theta \sin \phi \\ \sin \psi \sin \phi + \cos \psi \sin \theta \cos \phi & -\cos \psi \sin \phi + \sin \psi \sin \theta \cos \phi & \cos \theta \cos \phi \end{bmatrix} \quad (3.14)$$

where  $\theta$ ,  $\phi$  and  $\psi$  are the Euler angles yaw, pitch and roll.

By combining equations (3.10-3.13) the extracted jets can be expressed as a function of

the 6 rotation and translation parameters.

$$\vec{J}_{left,right}(x, y, z, \theta, \phi, \psi) = g(f(C_{left,right}, I_{left,right}, R(\theta, \phi, \psi)P_0 + [x, y, z])) \quad (3.15)$$

The *facial* similarity is then finally defined as the sum of the model-image similarity, (3.11), obtained from the left and right images.

$$S_{\text{face}}(M, I) = S(M, I_{\text{left}}) + S(M, I_{\text{right}}) \quad (3.16)$$

Inserting (3.15) into (3.16) gives the facial similarity as a function of rotation and translation, the optimal mapping of the mesh is found by maximizing this function.

$$\begin{aligned} S_{\text{face}}(M, I, x, y, z, \theta, \phi, \psi) = & \sum_{i=0}^{L-1} w_i S_{a,d}^i(\vec{J}_i, \vec{J}'_{\text{left}}(x, y, z, \theta, \phi, \psi)) \\ & + \sum_{i=0}^{L-1} w_i S_{a,d}^i(\vec{J}_i, \vec{J}'_{\text{right}}(x, y, z, \theta, \phi, \psi)) \end{aligned} \quad (3.17)$$

The model mapping task has now been reduced to an optimization problem but not a trivial one. The six degrees of freedom, three components of translation and the three angles of rotation, makes the search space extremely large. An exhaustive search of the whole pose-space is thus intractable. Genetic algorithms were also discarded for the same reason. Gradient based methods proved too sensitive to local maxima to be feasible.

A direct search method called the Nelder-Mead simplex search method [22] was the one that was finally chosen. Direct search methods are often used in practical applications where highly accurate solutions might be undesirable because of the high computational costs required to attain it. This approach has been very successful in applications where an improvement rather than an optimization is the goal. The Nelder-Mead search method only uses function values and does not try to approximate function gradients. It has been proven by Wright [23] that this search method can converge to an acceptably accurate solution substantially faster than multi-directional searches or the gradient based, steepest descent method.

The optimization is divided up in a number of steps to reduce the possibility of errors

caused by local maxima. It was assumed that the largest variations would be found in the translation components, especially the in-plane translation. The depth translation and the rotation can be assumed to be smaller if the subjects are cooperative. By this, we mean that the subjects will try to keep their heads upright and facing the cameras at approximately the same distance.

The two in-plane translation components  $x$  and  $y$  are first optimized, the scale component and then later the rotation angles are added so that finally optimization is performed in all six dimensions.

### 3.4.2 Scale and In-plane Rotation Compensation

Translation and rotation will not only alter the location of the facial landmarks but also the jet coefficients. There are four different kinds of changes that will effect the jets, scaling, in-plane rotation and two types of in-depth rotation, caused by nodding and shaking of the head. All of these can be compensated for. However, we address only the first two types of changes. These compensations are performed by reshaping the Gabor filter bank and then filtering the image pairs again.

Since the real world 3-D coordinates are referenced to the epipole and the two image planes one of the rotation components  $\phi$  will represent the in-plane rotation. The variation to the jet caused by this rotation is easily compensated for by rotating the whole Gabor filter bank by the same angle in plane, Figure 3.20.

Scaling is estimated by determining how much the perspective projection of the head has changed. According to the pinhole camera model Chapter 2 the depth translation component  $z$  is inversely related to the magnification of projected points

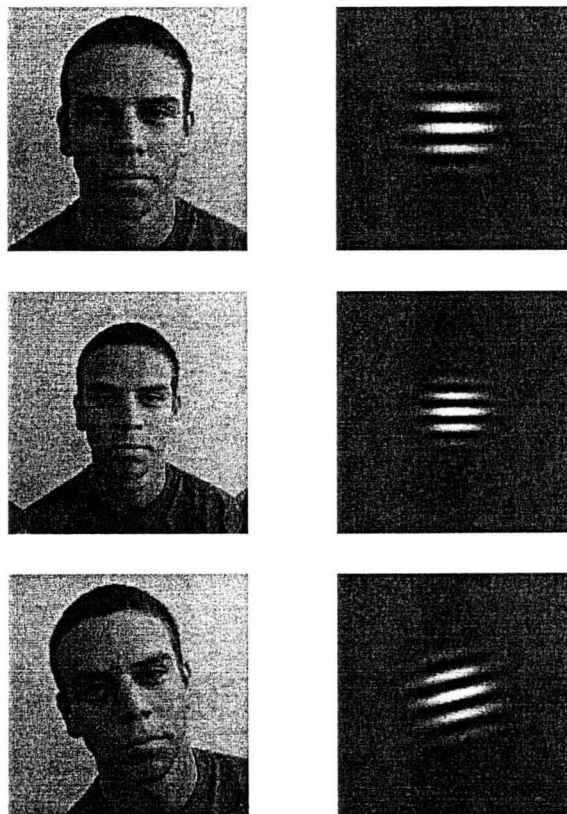
$$G = \frac{f}{z_w} \quad (3.18)$$



The scaling factor between the faces from image 1 & 2 can then be written as

$$s = \frac{f_1 z_{w2}}{f_2 z_{w1}} = [\text{Assume equal focal length, } f_1 = f_2] = \frac{z_{w2}}{z_{w1}} \quad (3.19)$$

where  $z_w$  is the, by the positioning algorithm estimated distance to the face from the cameras. The Gabor filter bank is then scaled up or down to compensate for the change in perspective scale of the face, Figure 3.20



**Figure 3.20:** Examples of scaled and in-plane rotated heads with the corresponding filter compensations.

It is possible to, instead of carrying out additional time consuming convolutions, adjust the jets for scaling and rotation by linearly combining the jet coefficients from filters of differing sizes and orientations. However, for the sake of higher accuracy this method was discarded in preference to compensation by refiltering.

### 3.4.3 Recognition & Verification Analysis

Depending on the current application of the system the resulting score is processed in differently.

For classification, or recognition, the unknown face is simply compared to all the models in the database and the model that yields the highest similarity value is chosen as the correct match.

In the case of identity verification the aim is to determine whether or not a subjects claim to an identity is legitimate. To do this it must be determined if the similarity value of the match between the unknown face and the model of the tested identity  $S_0$  is significantly higher than other models. This is done by using a method suggested by Lades *et al.* [24]. The unknown face is not only compared to the supposedly correct model but also against a number of  $P$  models chosen randomly from the model database. The mean  $m$  and standard deviation  $\sigma$  of the scores for the  $P$  models  $\{S_j | j = 1, 2, \dots, P\}$  are calculated. The criteria of acceptance is defined as

$$\kappa = \frac{S_0 - S_1}{\sigma} \quad (3.20)$$

with  $S_1$  is the best match among the randomly chosen models. The subject is accepted if  $\kappa$  is larger than a certain acceptance threshold  $t_{acc}$  and rejected if it is lower. The acceptance threshold will determine the trade-off made between ruling out the false identity claims and accepting the correct ones. The level of the threshold depends on what type of installation the system is supposed to safeguard. Lower security areas will typically have a low false rejection rate and a somewhat higher rate of false acceptations. More sensitive posts will on the other hand require a low false acceptance rate with the concession of having more false rejections.

## Chapter 4

# Experimental Evaluation

This chapter contains the results and analysis of the different experiments that were conducted to evaluate the performance of our system. Tests were carried out to investigate how the system handles pose changes as well as its performance in recognition and verification tasks.

### 4.1 Investigation of the Performance of the Pose Compensating Algorithms.

This section details results obtained by testing the system on the pose database, POSIM. This database was constructed by photographing a person performing a number of predetermined poses in front of the camera instead of manual modification of a template image pair, Figure 4.1 and 4.2. The latter will allow for a more accurate control of the scaling and rotation but the former method will supply more realistic images. Since this system is planned to be used in a realistic environment we decided to make our experiments as true to life as possible and therefore we chose the former of the two methods. All the images, training as well as testing images, were taken in one session in an attempt to eliminate unwanted variables such as lighting variations, differing facial expression and changes in appearance due to lapsed time.





**Figure 4.1:** A posed scaling versus a manually constructed one, both images are scaled down by 30% compared to the training image. The posed scaling is very difficult to control and is therefore carried out by first capturing a number of images with different, unknown scales and then manually measuring scaling afterwards. When constructing scaled images by manipulation of an original image issues such as how to deal with image edges will arise. In the left image above an attempt to correct such an edge effect by stretching the image borders can be observed.



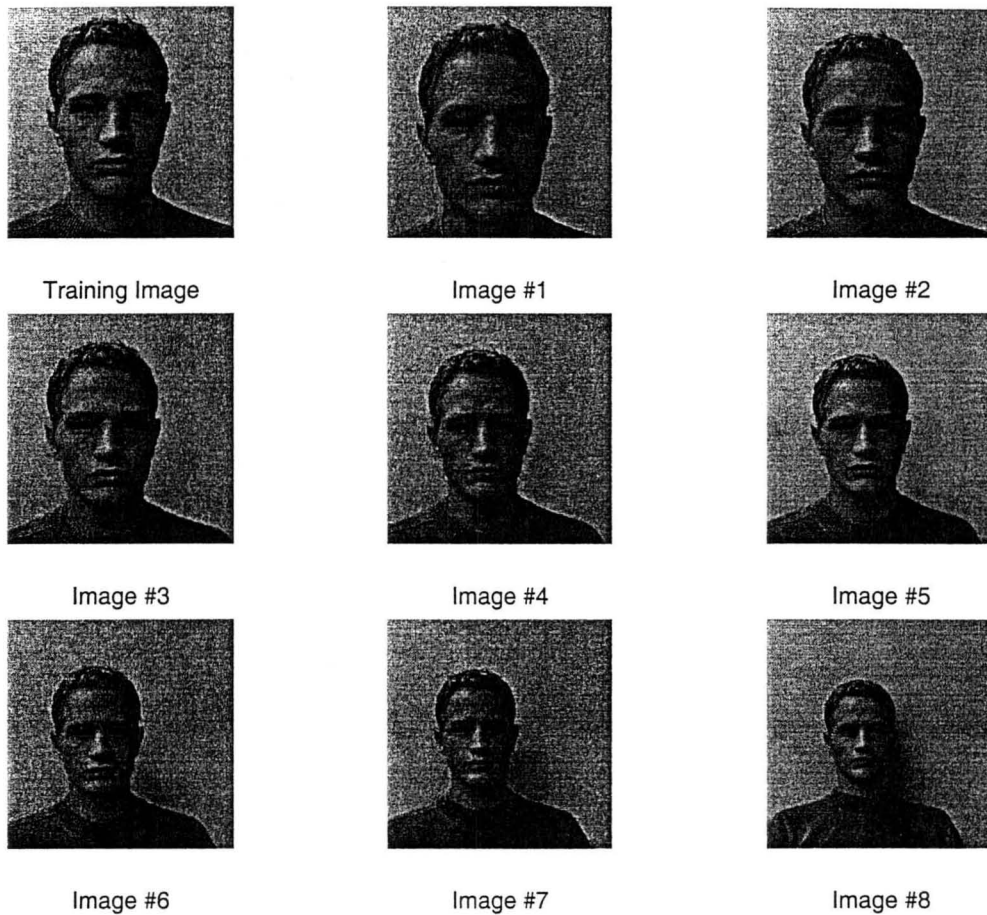
**Figure 4.2:** Two versions of in-plane rotated faces. The same problems as in Figure 4.1 will occur in this instance as well.

The pose test set can be divided into two parts, scaled and rotated, or tilted, faces. One image pair of the subject in a normal pose was used for training.

The scale set was acquired by having the subject to move, starting close to the cameras, further and further away from the setup in as small steps as possible. This resulted in a set of 8 image pairs where the scaling ranges from +15% to -45%, Figure 4.3. The scaling was determined manually from the image set by marking a number of points in the image and calculating how much their relative distance had changed compared to the training image pair.

In a similar fashion the rotated set was acquired. The cooperation of the subject was again utilized to capture a number of images of a face tilted at different angles. Owing to the

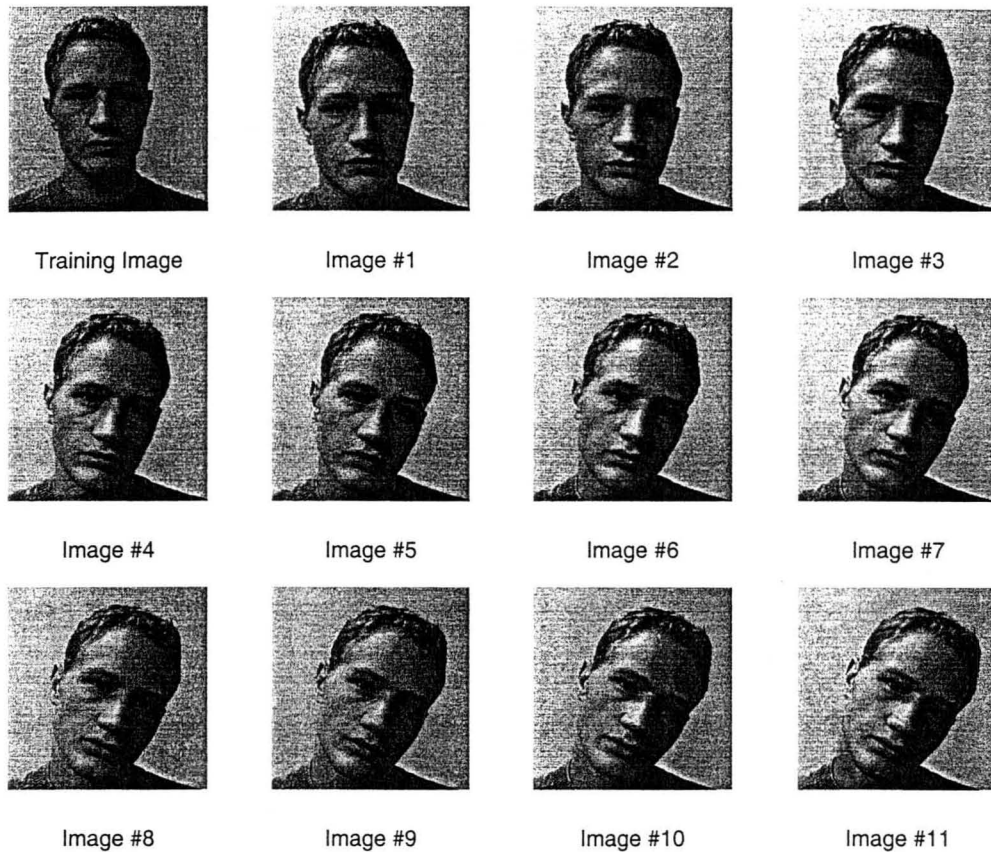




**Figure 4.3:** The images in the pose database used to examine scale compensation, scaling ranges from +15% to -45% compared to the training image.

symmetry of the problem only tilting to one side was carried out. The rotation angles were once again estimated manually by examining how certain points in each images change compared to the training images. The rotated image set consequently contain 11 image pairs rotated from 0 to 30° in fairly equally spaced steps, Figure 4.4.

The experiment was carried out by processing the image pairs as described in Section 3.4 with the small alteration that the outputs from the different stages of this procedure were analysed as well as the final result. These outputs were used to conduct a number of tests to examine the performance of the different pose compensation schemes suggested in the previous chapter. The system's ability to estimate the physical values of these two



**Figure 4.4:** The images in the rotated test set.

pose variations were also thoroughly examined as this will greatly influence recognition performance. Even the most accurate scaling and rotation compensation algorithm will be rendered useless if a good estimate of the degree of scaling and rotation can not be established.

The first experiment was conducted to investigate the influence different methods of compensating for pose affect the similarity between the model and test images. The following compensation steps were examined

**Face finding :** The significance of locating a face, in even a partially controlled environment such as this one, is determined by examining at how much the facial similarity value improves from the initial mapping to an optimization of the in-plane translation variables  $x$  and  $y$ .



**Pose estimation :** Here all of the six translation- and rotation parameters were optimized to determine how much is gained by extending the face finding algorithm to include pose estimation as well.

**Jet compensation :** In the final step the importance of compensating the jet values themselves is examined by again evaluating the improvements made in the facial similarity value from the previous steps.

The performance of the pose estimation procedure was determined by examining how well its estimates of the scaling and rotation compares with the manually measured values of these parameter. How the scale and rotation is calculated from the mapping parameters is discussed later in this section.

The model matching results on the scaled image set can be seen in Figure 4.5. The system managed to fit the mesh onto the test face successfully in all instances. From this mapping the scaling of the test face can be determined by utilizing the perspective geometry described in Section 2.1 and looking at the relative distance to the cameras.

The results from estimating scale this way, (Figure 4.6), proved to be very accurate, the average error of this algorithm was less than 1.5%. For each scaled image the facial similarity value of the different pose compensation steps described above was also calculated, (Figure 4.7). This figure clearly illustrates how the similarity between the test image pairs and the model changes for different scale at different stages of the pose estimation and compensation algorithm. Starting with the bottom dashed line in Figure 4.7 that shows the facial similarity value without any kind of optimization of the mapping and comparing that with the dotted line, where an estimation of the in-plane translation has been performed, the significance of an accurate location of the face is obvious. Even in cases where the faces have the same scale and are fairly aligned to begin with, Figure 4.8.

When the mesh mapping is extended to include depth translation and rotation as well the facial similarity value increases further even for the cases where scaling and rotation is not noticeable. This indicates a high sensitivity for these types of pose variations as well.

Even though the mesh is fitted successfully onto all the faces in the scaled image set the

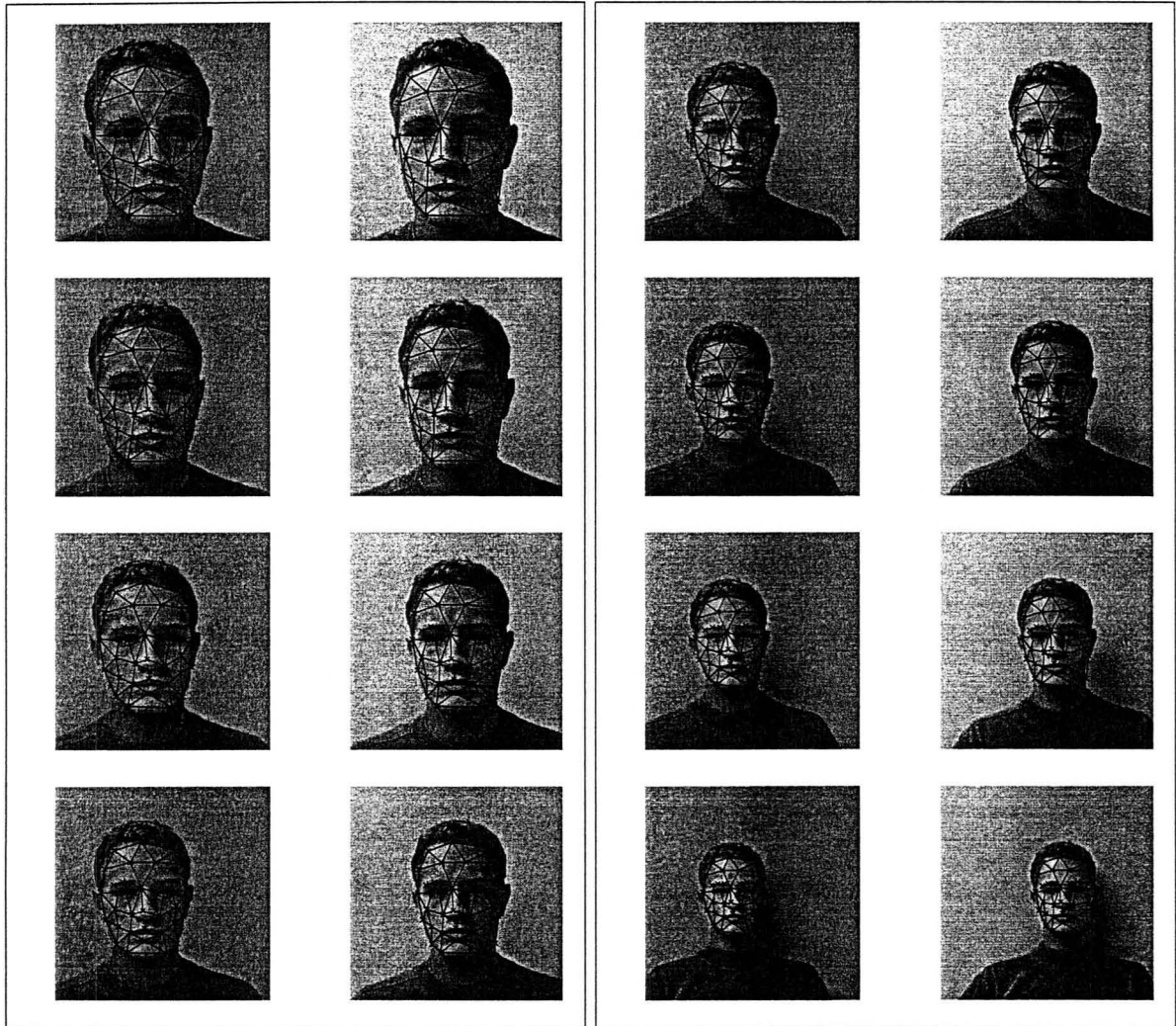
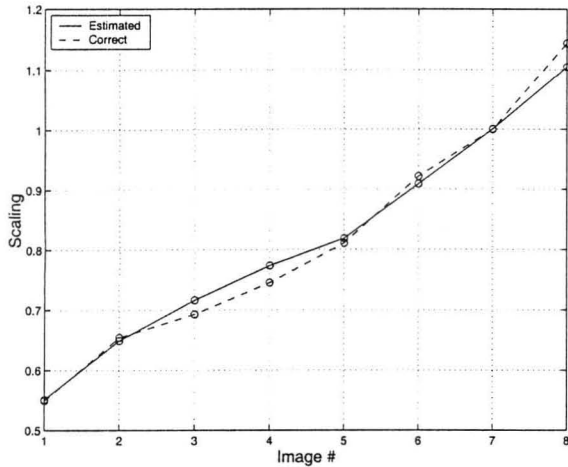
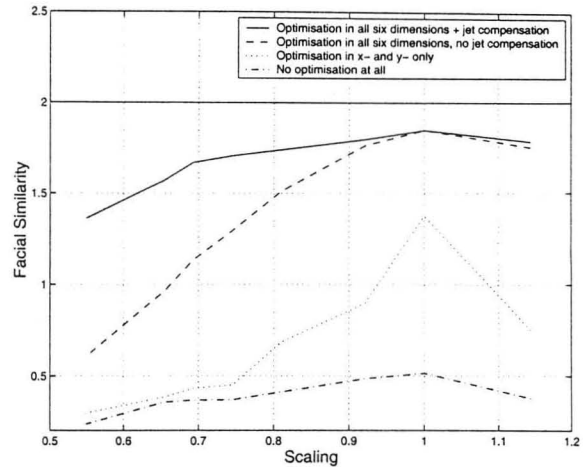


Figure 4.5: The resulting model mapping onto the scaled image set.





**Figure 4.6:** The manually measured scaling along with the system's estimated scaling referenced to the training image of the 8 image pairs in this test set.

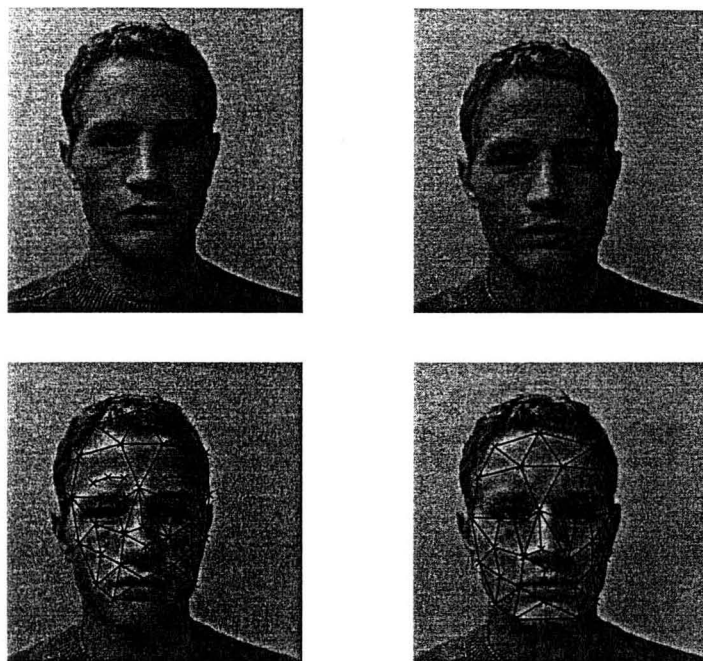


**Figure 4.7:** The facial similarity values for different compensation methods as a function of scale.

similarity value rapidly decreases when the scaling of the test image starts to deviate from the training image. This is due to the changes in the jets caused by the fact that the Gabor filters are of constant size and leads us to the next and last compensation performed, the jet compensation. Even though the facial similarity value is based on the displacement based similarity function, a reasonably robust method of comparison as discussed in Section 3.2, as the faces are scaled further and further the extent to which the jets are transformed exceeds the capability of this function. Using the estimated scaling, (Figure 4.6), the filter bank is rescaled the same amount and the test images are re-filtered and the optimal mapping of the mesh is calculated resulting in a more correct similarity value, solid line in Figure 4.7. The reason for why the similarity values still decrease can be explained by the loss of details that occurs as result of the camera resolution. It will always be more difficult to recognize a person at a distance.

The rotated test set was tested and analysed in the same fashion as the scaled test set, the results of the mapping can be seen in Figure 4.9.

Once again the pose estimation proved to be extremely accurate, the estimated rotation angles are presented in Figure 4.10. The average error of the rotation angle algorithm was

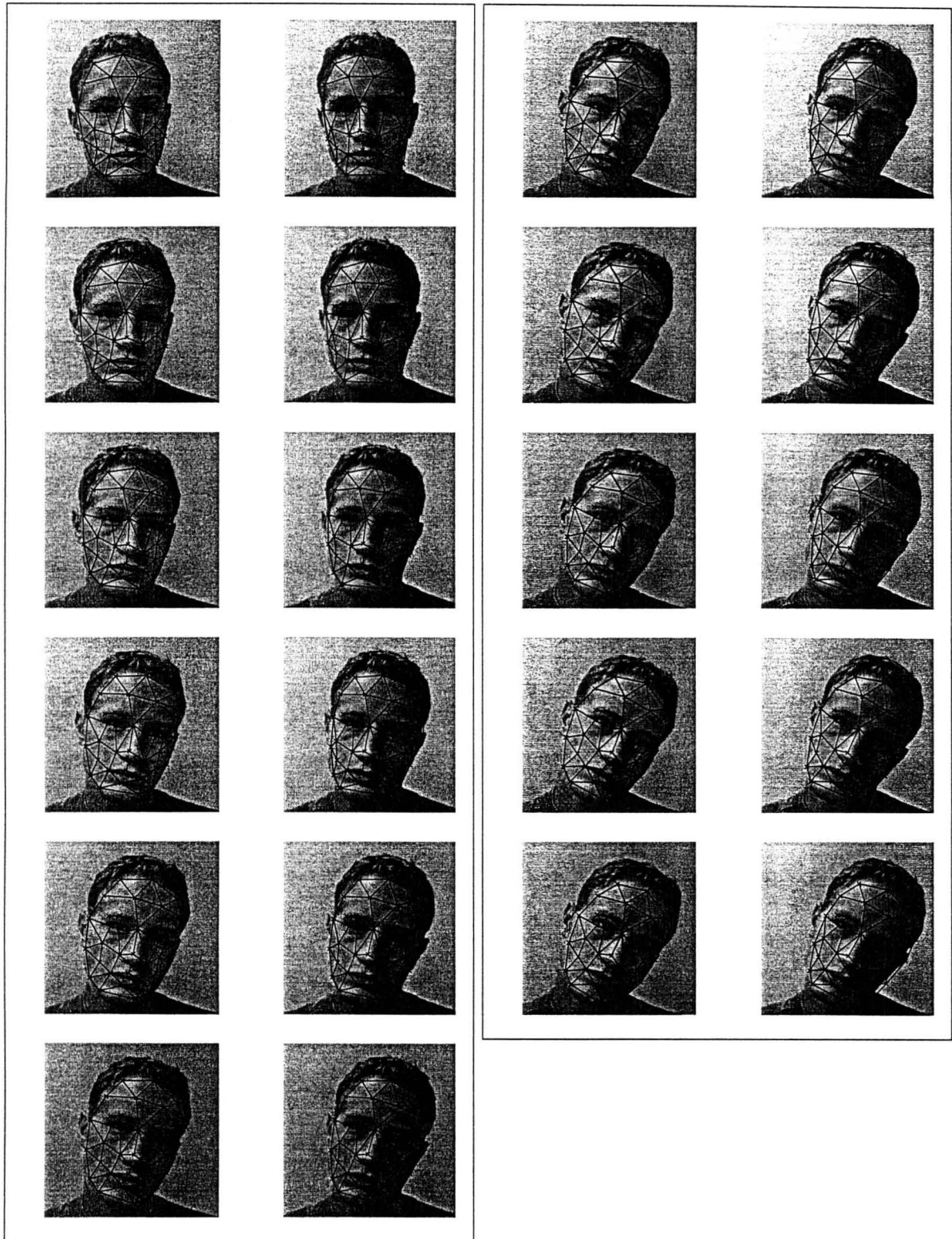


**Figure 4.8:** This figure illustrates how two initially well aligned faces, the training image, top left, and test image no. 7, with the same scale as the training face, in the top right need further in-plane translation compensation. Comparing the un-optimized mapping of the extracted mesh, bottom left, and in-plane translation optimized mapping, bottom right, the facial similarity value almost tripled when the  $x$  and  $y$  parameters are optimized. This again points out the importance of accurately locating the face to be recognized, the optimal mapping has a projection only 20 pixels away from the initial one.

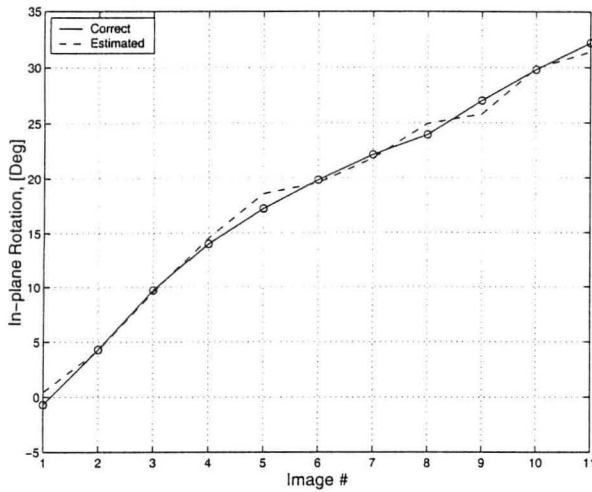
as low as 0.64 degrees.

An analysis of the facial similarity values, (Figure 4.11), was also carried out and from which similar conclusions can be drawn. The importance of locating the face can be seen in this test as well as the gain of allowing for optimizing of all the six translation and rotation parameters. The jet compensation makes the final similarity values less sensitive to pose to the same extent as it did for the scaled test set. The reason for why the similarity values decrease in this test as well even though no loss in detail is present can be found in the choice of image set acquisition method. It can be seen by examining the images in the test set that the subjects face is inadvertently distorted by the somewhat unnatural head position. The presence of the neck and shoulders also contributes to additional distortion since they do not move with the head.

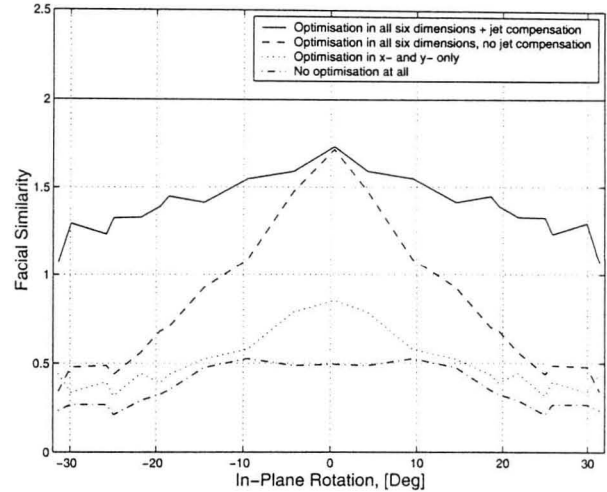




**Figure 4.9:** The resulting model mapping onto the rotated image set.



**Figure 4.10:** The manually measured and the, by the system, estimated in-plane rotation of the 11 faces in the test set.



**Figure 4.11:** The facial similarity values for different compensation methods as a function of rotation angle. The symmetry of the setup was utilized to extend the figure to included rotation in both directions.

Even though jet compensation was shown to improve the performance of the system considerably it can also be seen in figures 4.7 and 4.11 that it makes the biggest difference when the scaling and/or rotation is of significant magnitude. Jet compensation is therefore only employed when any of these two parameters exceed a certain threshold in an attempt to reduce processing time, we used  $\pm 3\%$  for scaling and  $\pm 2^\circ$  for rotation.

To conclude this experiment, the investigation strongly confirms the assumptions made in Chapter 3 on pose compensation and the efficiency of the implemented algorithms. It also displayed the impressive accuracy of the system's pose estimation algorithms.

## 4.2 Recognition & Verification

The recognition and verification experiments conducted are explained in this section, followed by a discussion of the results.

The system's ability to perform recognition and verification was assessed by testing it on the



FASIM database. The models were trained using the procedure described in Section 3.3, resulting in a gallery of 68 models. Recognition and verification tests were carried out on the test set in a number of different ways. Firstly, the algorithm was tested on 85 test image pairs of subjects without glasses from the FASIM database. The same test was then repeated but this time the mesh was only mapped onto a single image of each subject. This test was devised in an attempt to quantify the gain made from using stereo information. The negative effect of spectacles on performance was also investigated by testing the system on a number of images of subjects with glasses.

The recognition results were evaluated by examining the matching scores between the test images and all of the models and choosing the best match as the correct one. It is very inconclusive however to only take into consideration whether a correct identification was made or not when analyzing recognition performance. The ranking made by the system in instances where correct matches were not made will also reveal information about the system's performance. Two face recognition systems that produces identical first rank recognition rate can still have very different performance. If, in all the occurrences of mismatches, one system ranks the correct subject as its second guess it should be considered more accurate than a system that ranks the correct subjects third or lower. Even though a system more correctly ranks the test subjects it will not affect the recognition rate but it will influence the verification results.

Most publications in the area of face recognition today only present their results in the form of first rank recognition rate. In biometric applications an algorithms verification capabilities are much more important as they are most commonly used for security purposes such as access control. Our system verifies identity using the assumption that if a persons identity claim is genuine he or she will score significantly higher than other identities, Section 3.4.3. This significance is tested by not only matching the unknown image pair against the model for the assumed identity but also against  $P$  randomly chosen models. If the correct model scores significantly higher than the other  $P$  models the persons claim is accepted. Our methods verification performance was evaluated using the FASIM database. The results are presented in the form of *False Acceptance Rate* (FAR) and *False Rejection*

*Rate* (FRR) curves, or acceptance/rejection curves, that shows the probabilities of incorrectly accepting or rejecting identity claims for different acceptance threshold levels. The point where the false acceptance rate and false rejection rate are equal is called the *Equal Error Rate*, (EER), and is a commonly used performance measurement in biometrics.

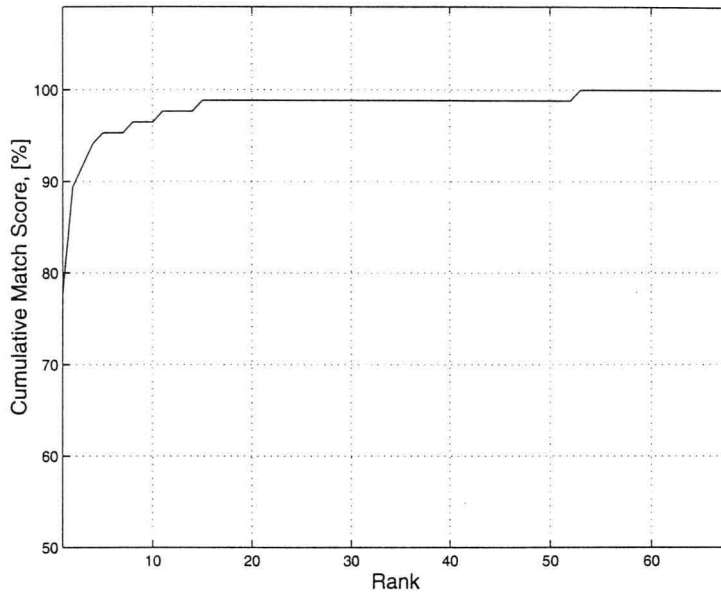
### 4.2.1 The Testing Procedure

All the experimental results in the following sections were acquired using the testing procedure described in Section 3.4. Only a few alterations were made for practical reasons. The implemented algorithm has been optimized only to a very small extent, resulting in an extremely computationally expensive procedure even in its most basic version. To process one image pair on a Pentium II 450 MHz with 192 MB of memory takes approximately 40 seconds, 19 seconds for the filtering and 21 seconds to optimize the mapping. In an attempt to keep the time requirements for the testing procedure within reasonable limits the algorithm was reduced by limiting the employment of the scale & in-plane rotation compensation algorithms described in Section 3.4.2. Only the top ten scoring models are fed through this refinement stage. It was also noticed that problems with local maxima in the mapping optimization only occur in approximately 5% of the cases and that the matching score then are usually very low. To again reduce processing time, the procedure described in Section 3.4.1 for the purpose of escaping local maxima was only employed when the matching score is lower than a certain threshold. This strategy will reduce the performance of the system as local maxima located near the global maxima will not be corrected for.

### 4.2.2 Stereo Image Mapping Results

This system was designed to perform recognition and verification on calibrated stereo image pairs. The following experiment evaluates how well this task is carried out. The result from the recognition test can be seen in Figure 4.12.

Before verification can be performed, the size of the verification set (the  $P$  randomly chosen



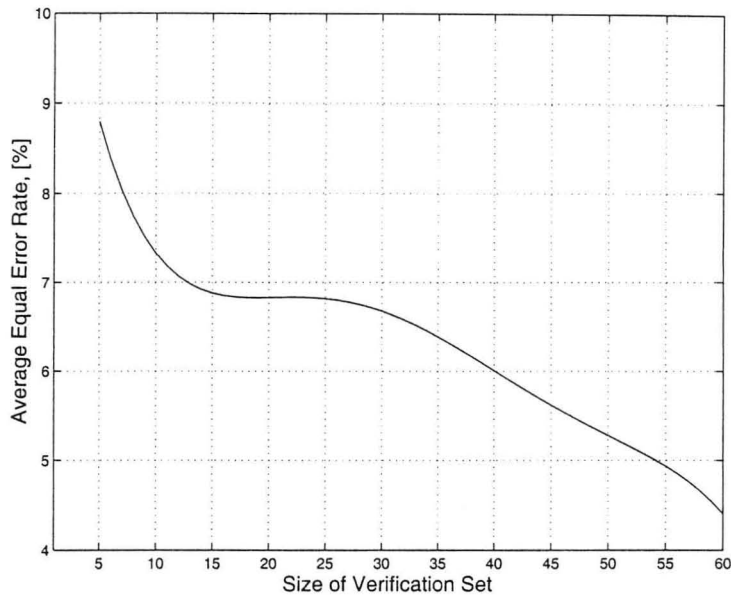
**Figure 4.12:** The recognition results on the FASIM database. Cumulative match score as a function of rank, 78% of the subjects were correctly identified and in 92% of the cases the correct match was found within the top three ranks.

models that are tested against) must be chosen. It will be a trade-off between performance and processing time. The verification set is meant to be representative of the matching score distribution so that a decision on whether the claimed identity matches significantly higher than the other models can be made. The smaller the verification set the higher is the probability that such significance is falsely achieved by chance. The influence of this size on verification performance was investigated by examining how the equal error rate change as a function of verification set size. The averaged results from 100 repeated runs can be found in, Figure 4.13.

A trade-off between performance and processing time has to be made when deciding on the size, the more models used will reduce the equal error but will also require more processing time. The final decision will have to be largely based on the type of application.

In the following experiments  $P = 10$  models were used in the verification set, a fairly low number, chosen to keep processing time down but also to ensure that the comparative results do not become saturated. When the performance of a certain algorithm is very close





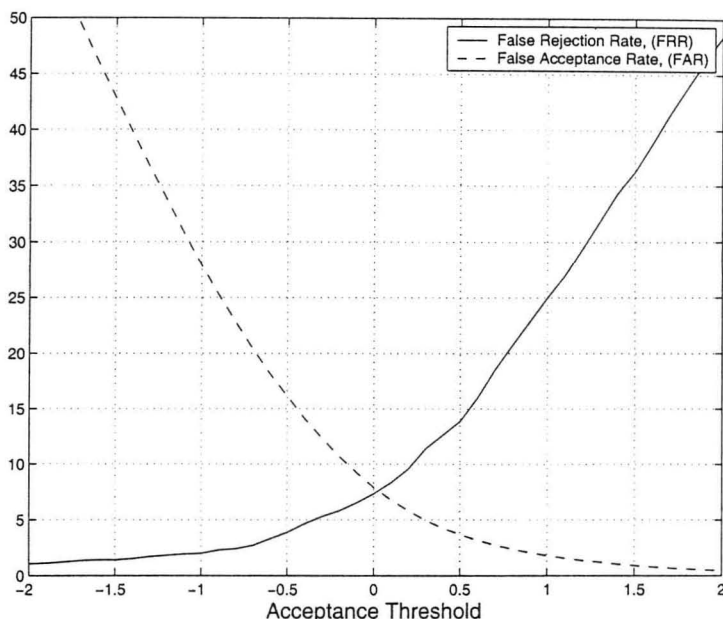
**Figure 4.13:** The averaged equal error rate of a 100 test runs as a function of verification set size. It can here clearly be seen how performance improves as the verification set grows.

to ideal, improvements made to that algorithm will be difficult to evaluate as there will be little or no increase in performance. By making the task more difficult improvements will be put into much better perspective.

The average acceptance/rejection results obtained from FASIM database is shown in Figure 4.14.

### 4.2.3 Single Image Mapping Results

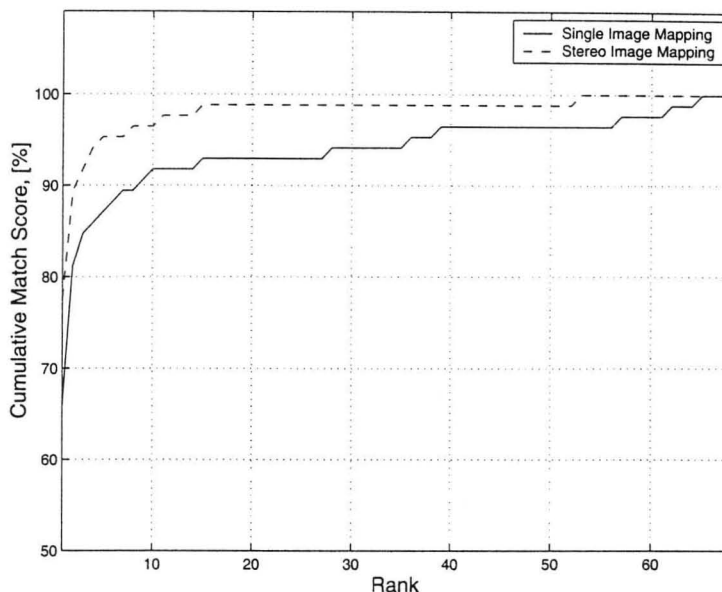
In an attempt to investigate how much is gained by using stereo information in face recognition/verification the recognition algorithm was again tested against the FASIM database, but now with a slight alteration. The mapping of the 3-D mesh was performed on only the left of the images in the image pair. This change also has the advantage that in a test setup only one camera is required, hence reducing costs, i.e. in a practical setup one stereo camera training stations would be assigned to perform all the training image acquisition so that at each access controlled point only one single camera would be necessary. It is



**Figure 4.14:** The average false rejection and false acceptance rate from 30 test runs for different acceptance thresholds,  $t_{acc}$ . The equal error rate is 7.5% for  $t_{acc} \approx 0$ , this should be compared with the lowest obtainable EER, 4.5%, see Figure 4.13.

noteworthy that even though only single images are used during testing, some stereo information is still utilized in the form of the 3-D mesh extracted in the training procedure. The results from the recognition and verification experiments, using single image mapping, can be seen in Figure 4.15 and Figure 4.16.

As expected the performance of the single image mapping proved to be somewhat lower than the stereo mapping method. A detailed analysis revealed, however, that approximately 50% of the errors made compared to the stereo mapping, were caused by unsuccessful mapping and was not a result of lacking discriminatory properties of the underlying algorithm. The task of finding the optimal position of the mesh is more difficult when the solution becomes less restricted. To simultaneously map the mesh onto two images has unquestionably more constrains than only mapping onto single images. The scheme for avoiding local maxima, described in Section 3.4.1 and successfully utilized by the stereo mapping algorithm was not efficient enough under these conditions. By implementing more advanced optimization algorithms this type of mapping could very well become feasible.



**Figure 4.15:** Single image mapping recognition results on the FASIM database compared with the stereo image mapping results from Figure 4.12. A first rank recognition rate of 66% and a third rank recognition rate of 85% constitutes a performance reduction of about 10% from the stereo image mapping method.

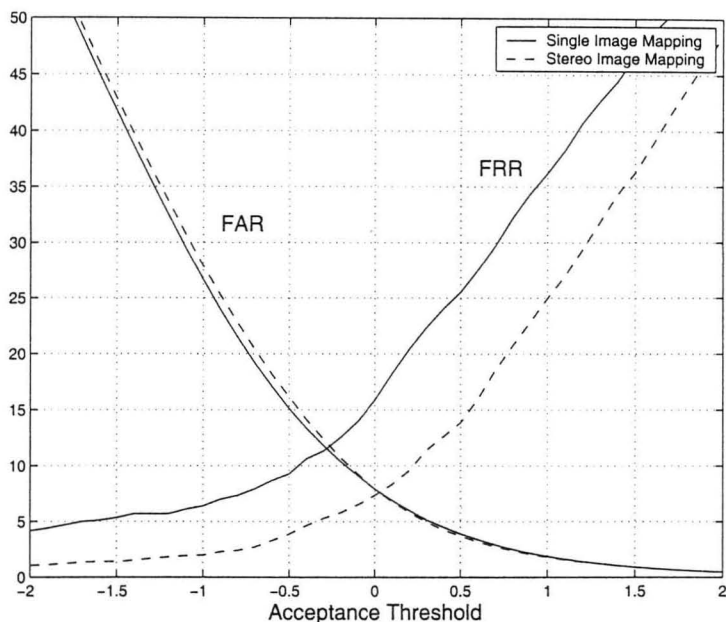
#### 4.2.4 Glasses

The effects of subjects wearing glasses was also investigated, however, owing to the circumstances under which these images were acquired, special consideration to the results have to be taken. These image pairs were taken during the same session as an image pair of the same subject without glasses, see the database description in Section 1.4.1. Since the time factor can greatly affect the performance of the system, Section 1.4, this must be taken into account when analyzing the results. Prior to testing this database, containing subjects wearing glasses, was thoroughly examined and classified according to what session each image pair was acquired during.

There were 58 image pairs of subjects wearing glasses, out of which 16 were used in the validation set and 42 in the test set. 13 of those 42 image pairs were of subjects that were part of the FASIM database.

These 13 image pairs were tested against the models in the FASIM database both for



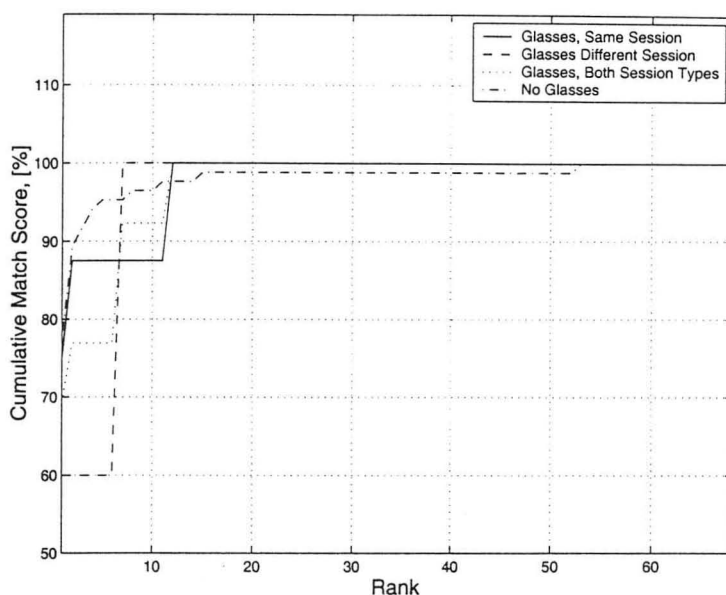


**Figure 4.16:** The acceptance/rejection curve, averaged from 30 test runs, for single image mapping compared to stereo image mapping. The equal error rate increased from 7.5% to 12%.

recognition and verification, Figure 4.17 and Figure 4.18. The reason for initially only using these 13 images for testing was so the results could be correctly and fairly compared to previous results.

The system managed to handle glasses very well, both from a recognition and a verifying point of view, as expected the recognition rate dropped somewhat and the equal error rate increased slightly, but still within reasonable limits. The results also managed to call attention to the influence of the time factor, the lapsed time between the acquisition of the training and the testing images, on overall performance.

The final experiment was devised to investigate this relationship between recognition performance and lapsed time between sessions further and with more statistical significance. By incorporating image pairs of subjects that only attended a single session but were wearing glasses we were able to construct both a new training and test set. The training set consisted of 21 subjects and the test set of 42 image pairs, 18 of which were from the same session and 24 from a different session. The result from this experiment is shown in



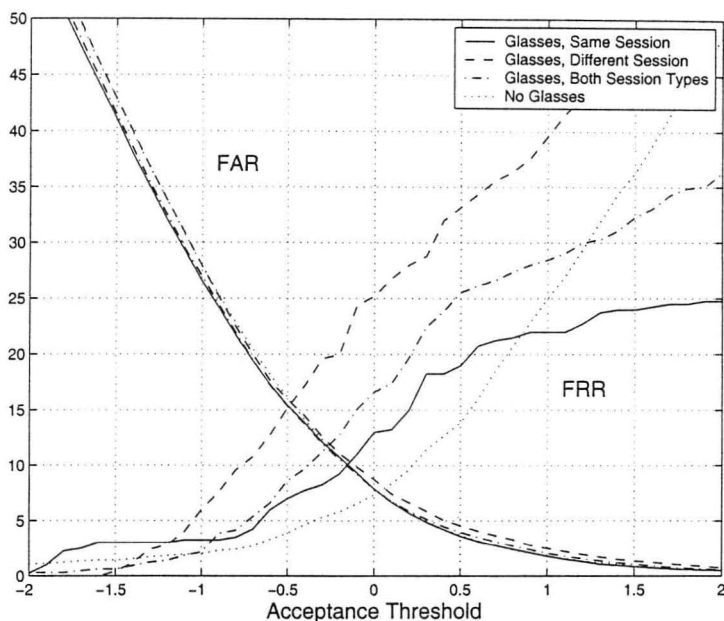
**Figure 4.17:** The recognition rate when testing subjects wearing glasses against the models from the FASIM database. Same session first rank recognition rate is 75% and different session recognition rate 60%.

Figure 4.19.

This effect that the time factor has on recognition rate is not only caused by a change in appearance of the subject but can also be explained by the fact that lighting conditions are the same throughout one session, the subjects pose, scale and facial expression do not alter very much between the two acquisitions. These types of changes actually seem to be the larger cause of dissimilarity between the sessions than changes in the actual subject.

### 4.3 Analysis & Discussion

The experiments conducted on this system has shown some very promising results. The pose estimation algorithm managed to estimate scale and in-plane rotation extremely well, within a few percent and degrees respectively. Since pose and scale are variables that most face recognition systems today have grave difficulties with, such an accurate estimation procedure will provide a very good foundation for which to base this system.



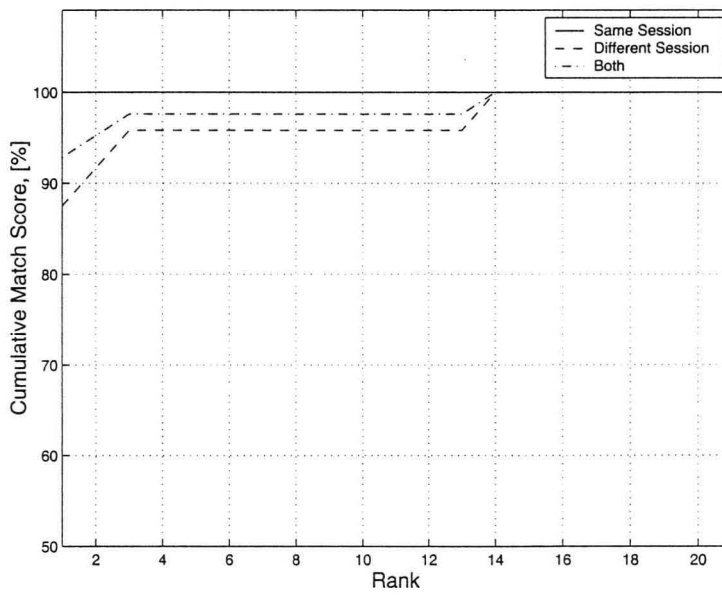
**Figure 4.18:** The average acceptance/rejection curve for subjects with and without glasses. Note the differing equal error rates for same session ( $\approx 10\%$ ) and different session images ( $\approx 15\%$ ).

The recognition and verification procedures also proved very reliable, with a first recognition rate of 78% and an equal error rate of down to 4.5%. As previously mentioned these results can not be compared directly to any other existing algorithms tested on different databases. A correlation based algorithm was implemented and tested on the FASIM database, the faces were located by our system and the images scaled and cropped. The results were not very convincing with a first rate recognition of approximately 15%, we chose however not to include these results in a more extensive form as correlation is extremely sensitive to changes in lighting conditions and no attempts were made to compensate for this.

The feasibility of a single image testing procedure was also examined, the results from that experiment indicated strongly that such an approach would be worth further pursuit

The system also proved to be fairly robust, recognition dropped by an average of 9% to 69% and the equal error rate increased by 7.5% when the system was tested on a set of image pairs of subjects wearing glasses.





**Figure 4.19:** Comparison of recognition rates for test images acquired during the same session as the training image and images acquired in different photo session. Note that these results can not be compared directly with the previous ones as the size of the training set is different, see Section 1.4.

# Chapter 5

## Conclusions

A novel way of recognizing human faces was presented in this thesis. This chapter contains a summary of the experiments that were carried out along with a conclusive discussion of the results. Suggestions on the direction of future research are also included.

### 5.1 Summary of Results

In order to evaluate the proposed algorithm several different experiments were conducted. The first test was designed to study the performance of the different pose compensation schemes and to allow for an examination of how well the system could handle largely varying pose and scale by testing it on the POSIM dataset. The system successfully located and fitted the mesh on the face optimally on all of the 8 scaled and 11 rotated image pairs. In this test it was shown that very accurate estimates of the in-plane rotation and scaling was made by the system, the mean errors were less than 1.5% units for scaling and  $0.65^\circ$  for rotation. The pose compensation algorithms proved equally efficient, the matching scores were increased, in some cases by as much as 1100%. An accurate pose estimation and compensation procedure, such as this, is an essential part of any face recognition algorithm trying to recognize people under largely varying conditions.

The system also produced convincing recognition results, with a first rank recognition rate

of 78% on the 85 test image pairs in the FASIM database. Single images and image pairs of subjects wearing glasses were also classified with satisfying accuracy. These results are also emphasizes the high level of robustness that our system possesses.

An equal error rate of as low as 4.5% was achieved on the FASIM database during verification testing. This algorithm also proved to be fairly robust, single image mapping and the introduction of glasses only increased the equal error rate with 5.5% and 7.5% respectively. These results can not, however, be evaluated to its fullest extent until a thorough comparison with established face recognition algorithms under similar conditions can be made.

## 5.2 Future Work

Even though the work carried out in this thesis managed to successfully reach many of the goals initially set out, large number of possibilities for future research and further development of this system still remains open.

A detailed investigation into the choice of facial landmarks should be carried out. The intuitive choice of feature points made at this stage should be replaced by a more formalized procedure. Extensive studies on the importance of different facial features for face recognition has been carried out by the psychology community, [25–27] and could be utilized. A more analytical approach is to choose the landmarks individually for each face based on facial texture or stereo matching potentials. One such procedure was suggested by Manjunath where a generic feature detector [28], responsive to short lines, line endings and corners, was utilized to place feature points in a face for the purpose of recognition, [29].

Methods to compensate for in-depth rotation should be implemented. An approach to the former has been proposed by Maurer *et al.* [30]. This method attempts to compensate for in-depth rotation by modeling it with a rigid rotation of a face with locally flat object shape near the feature points. Since the use of stereo information in this system will supply the true object shape an improvement of this procedure appears viable.



The effects of how localised analysis of the jets for the detection of serious changes in facial expression or the presence of glasses should also be investigated. At the moment no note of these changes are taken into account when weighting the features. By allowing for the weights to be altered in the testing procedure if any irregularities are encountered, performance should improve. If the presence of glasses is detected, less emphasis should be placed on the eye region during recognition. The mesh should also be made elastic at certain nodes to model the non-rigid areas of the face such as the mouth.

Further research is also needed into the configuration of the filterbank, the filters are currently spread out to evenly cover the frequency space. A Karhunen-Loeve analysis of the discriminatory powers of the filters and corresponding jet coefficients discriminatory power should reveal how essential the different scales and orientations are for recognition. Such an analysis should also allow for a considerable reduction of the number of filters, thus lowering the number of convolutions needed in the pre-processing stage and making the system faster. Kalocsai *et al.* [31] has shown that such a filter reduction is possible without the sacrifice of any discriminative power of the representation.

An improved optimisation algorithm should be implemented to make way for the use of single image mapping procedures. Even though this mapping was proven to be inferior to the stereo mapping, it should not be abandoned as it might reduce hardware implementation costs considerably.

Finally, the issue of processing time should be addressed, at the moment the algorithm is implemented mainly in MATLAB and is too slow to be used practically. Optimizing the code into a compilable language should reduce processing times considerably.

### 5.3 Summary

This thesis has presented a novel approach to face recognition using binocular stereo information. A fully functional 3-D face recognition system, with excellent performance, was designed and implemented. It includes a framework for accurate face location, as well as pose and scale estimation under varying lighting conditions. The performance of the sys-

---

tem, as well as the pose and scale estimations, was determined through extensive testing of the algorithm on a facial stereo image database constructed as a part of this work. A recognition rate of 78% was achieved on the FASIM database containing 85 image pairs of 68 different subjects.

# Bibliography

- [1] J.P. Phillips et.al., "The FERET September 1996 Database and Evaluation Procedure," in *Proceedings of the First International Conference on Audio and Video-based Biometric Person Identification, Crans-Montana, Switzerland, March 1997*.
- [2] T. Kanade., "Computer recognition of human faces," in *In Interdisciplinary Systems Research*, Birkhäuser Verlag, 1977.
- [3] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1042–1052, 1993.
- [4] X. Jai and M. Nixon, "Extending the Feature Vector for Automatic Face Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 1167–1176, 1995.
- [5] L.Wiskott et.al., "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 755–799, 1997.
- [6] K. Okada, et.al., *The Bochum/USC Face Recognition System. And How it Fared in the FERET phase III Test*. Springer Verlag, in press.
- [7] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.
- [8] B. Moghaddam, W. Wasiuddin and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition," in *3rd IEEE International Conference on Automatic Face & Gesture Recognition*, April 1998. Nara, Japan.
- [9] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *2nd IEEE Workshop on Applications of Computer Vision*, Dec. 1994.
- [10] T. Fromherz, *Shape from Multiple Cues for 3D-enhanced Face Recognition*. PhD thesis, University of Zürich, Zürich, Switzerland, September 1997.
- [11] B. Achermann et.al., "Face Recognition Using Range Images," in *Proceedings International Conference on Virtual Systems and Multimedia '97 (VSMM'97)*, pp. 129–136, September 1997. Geneva, Switzerland.



- [12] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiments," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 920–932, September 1994.
- [13] D.J. Fleet and A.D. Jepson, "Phase-Based Disparity Measurement," *CVGIP: Image Understanding*, vol. 53, pp. 198–201, 1991.
- [14] J. Magarey and A. Dick, "Multiresolution Stereo Image Matching using Complex Wavelets," in *Proceedings 14th International Conference on Pattern Recognition (ICPR'98)*, vol. I, pp. 4–7, August 1998. Brisbane, Australia.
- [15] C. Venter and D. Weber, "Three Dimensional Reconstruction of Faces," in *IEEE Africon '99*, vol. 1, pp. 113–116, September 1999.
- [16] R.Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision and Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE Journal of Robotics and Automation*, vol. RA-3, no. 4, pp. 323–344, 1987.
- [17] J.G. Daugman, "Complete Discrete 2-D Gabor Transform by Neural Networks for Image Analysis and Compression," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1169–1179, 1988.
- [18] J. Buhmann, *Object Recognition in the Dynamic Link Architecture: Parallel Implementation on a Transputer Network*, ch. Neural Networks for Signal Processing, pp. 121–159. Prentice Hall, 1992.
- [19] W.M. Theimner and H.A. Mallot, "Phase-Based Binocular Control and Depth Reconstruction Using Active Vision," *CVGIP: Image Understanding*, vol. 60(3), pp. 343–358, 1994.
- [20] N. Krüger, M. Pötsch and C. von der Malsburg, "Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs.," *Image and Vision Computing*, 1997.
- [21] N. Krüger, "An Algorithm for the Learning of Weights in Discrimination Functions using a priori Constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1997.
- [22] J.A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.
- [23] M.H. Wright, "The Nelder-Mead Method: Numerical experimentation and Algorithmic Improvements," tech. rep., AT&T Bell Laboratories, NJ, USA, 1996.
- [24] M. Lades et.al., "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Transaction on Computers*, vol. 42, no. 3, 1993.
- [25] V. Bruce, *Recognising Faces*. Laurence Erlbaum Associates, 1988.



- [26] R. Diamond and S. Carey, "Why Faces are not Special : an Effect of Expertise," *Journal of Experimental Psychology: General*, vol. 115, pp. 107–117, 1986.
- [27] A.W. Young and V. Bruce, *Face Recognition*, ch. Perceptual Categories and the Computation of "grandmother", pp. 5–49. Laurence Erlbaum Associates, 1991.
- [28] B. S. Manjunath, C. Shekhar and R. Chellappa, "A new approach to image feature detection with applications," *Pattern Recognition*, vol. 29, pp. 627–640, April 1996.
- [29] B. S. Manjunath and R. Chellappa, "A Feature Based Approach to Face Recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition '92*, pp. 373–378, June 1992.
- [30] T. Maurer and C. von der Malsburg, "Learning Feature Transformations to Recognize Faces Rotated in Depth," in *International Conference on Artificial Neural Networks*, October 1995. Paris.
- [31] P. Kalocsai, H. Neven, J. Steffens and I. Biedermann, "Statistical Analysis of Gabor-filter Representation.," in *3rd IEEE International Conference on Automatic Face & Gesture Recognition*, 1997. Nara, Japan.