University of Nevada, Reno

Robust 3D Head Pose Classification Using Wavelets

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
with a major in Computer Engineering.

by

Mukesh C. Motwani

Dr. Frederick C. Harris, Jr., Thesis advisor

May 2003

US Patent Application No. 10/266,481 titled

METHOD FOR DISCRIMINATING POSES USING WAVELETS

was filed on October 5, 2001 and is in pending status.

We recommend that the thesis

prepared under our supervision by

MUKESH C. MOTWANI

entitled

Robust 3D Head Pose Classification Using Wavelets

be accepted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

Dr. Frederick C. Harris, Jr., Ph.D., Advisor

Dr. Yaakov Varol, Ph.D., Committee Member

Dr. William Kuechler, Ph.D., At-Large Member

Marsha H. Read, Ph.D., Associate Dean, Graduate School

May 2003

Abstract

This thesis describes a robust method for estimating face pose from a video sequence featuring views of a human head under variable lighting and facial expression conditions. Wavelet transform is used to decompose the image into multiresolution face images containing both spatial and spatial-frequency information. Principal component analysis (PCA) is used to project a mid-frequency, low-resolution sub-band face pose onto a pose eigenspace where the first three eigencoefficients are most sensitive to pose and follow a trajectory as the pose changes. Any unknown pose can then be estimated finding the Euclidean distance of the first three eigencoefficients of the query image from the estimated trajectory. Wavelet transform reduces the computational load on the PCA and makes the algorithm robust against illumination changes and facial expression. An efficiency of 84% was observed for test images under different environmental conditions not included during training.

Acknowledgments

I would like to thank Professor Harris, my advisor, for his generous help when it is most needed, for his guidance, support, and encouragement throughout the thesis process. I am very grateful to Professor Varol and Professor Kuechler for serving on my committee and for their valuable time. Finally, I would like to thank my sister, Rakhi and my parents for their love and support throughout my Master's thesis work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The problem of recognizing human body parts and detecting their pose in 3D space is well-known. Determining human head pose is just one of many aspects of this problem. Head pose provides good cues about the general focus of attention of a person. The estimation of head pose is necessary for head gesture recognition. For face and facial expression recognition, it is necessary to estimate head motion in order to track a head continuously [27].

A real-time pose estimator can be used to drive graphical models for applications such as virtual teleconferencing. It can also be used to index a more detailed view-specific representation for identity recognition and expression analysis. In addition, pose prediction is useful for overcoming display lags in real-time interactive and visual communication applications. An automatic video-conferencing system has used head pose classification to decide which camera is best suited to capture the scene [36]. Head pose estimation could also be used instead of typed manual commands to guide a mobile robot. Another application could be to classify the focus of attention of a car driver as straight-ahead, toward the rear-view mirror, toward the dashboard, *etc.* In some approaches, an approximate pose is estimated, but other approaches attempt to be precise so that they can be used for gaze detection such as in [14].

A human head rotating in depth (out of the image plane) induces nonlinear transformations in the projected image of the face. Facial features become occluded, and the outline of the face alters its shape, causing interference with the background. The appearance of the face is also affected by other parameters such as identity change, distance from the camera, facial expression, noise, illumination changes, and occlusion. Another major problem in automatic pose classification is that of knowing where the face is located in the current image prior to pose classification. Because most face detection algorithms assume that the face is frontal, they must work independent of the face pose. Alignment of the faces is equally important. For these reasons and more, pose classification is a difficult task. This thesis proposes a fast, low-cost method for determining head pose position.

The rest of this thesis is structured as follows: Chapter 2 gives a literature review of the techniques which have been used for pose discrimination. Chapter 3 describes the motivation behind our approach. Chapter 4 provides a detailed description of the experimental setup and the algorithm used to discriminate pose. Chapter 5 discusses results of the all the experiments performed to test the robustness of the system. Chapter 6 concludes the thesis by highlighting the limitations of the algorithm and presenting possible future work.

Chapter 2

Literature Review

There are a variety of approaches for face representation of poses, and they can be classified into three general categories: appearance-based, feature-based, and model-based.

2.1 Feature-based Approaches

In a feature-based approach, estimation based on the relationship between human facial features [6, 14] relies heavily on the accuracy of the facial feature detection schemes. Detection of facial features is not accurate and often fails because of the changes of the shape of facial features during changes in illumination conditions and facial expressions.

Baluja [2] used probabilistic modeling methods to solve the problem of discriminating between five facial orientations (left profile, left semi-profile, frontal, right semi-profile, and right profile) with very little labeled data. Azarbayejani *et al.* [1] used an extended Kalman filter to recover head pose from between ten and twenty tracked feature points. However, the estimation accuracy was found to degrade as the rotation angle increased.

Kruger *et al.* [20] used computationally intensive elastic graph matching to locate

and estimate the pose of faces. The graphs consisted of connected nodes of Gabor filter jets. Different graph models were needed for different poses, leading to a poorly integrated and computationally expensive approach.

Borovikov [5] described a method that relies on the idea that the orientation of a human head in 3D space can be recovered from an image by locating a set of crucial facial features within the head silhouette boundaries. He experimented on a series of 320×240 true color images, each displaying a single individual in quasi-frontal view. It was shown that the method could reliably estimate a person's head pose, provided that the head silhouette was pre-segmented correctly and that the crucial facial features were not obscured by anything. The method spends a lot of time locating the head silhouette and then locating the facial features within it.

2.2 Appearance-based Approaches

Appearance-based approaches attempt to capture and define the face as a whole. This approach has received lot of attention because of the limitations of feature-based approach. We can further classify this approach according to the type of classifier selected.

2.2.1 Templates

In one technique, templates representing facial feature are used in determining head position and orientation. The image is compared with various templates to determine which template most closely matches the image. An example of a template matching system is illustrated in [26] by S. Niyogi and W. Freeman. The significant differences in images resulting from different lighting conditions and different appearances makes matching with templates difficult. The most attractive advan-

tage of template matching is its simplicity; however, it suffers from large memory requirements and an inefficient matching algorithm, which makes it very slow.

2.2.2 Neural Networks

Rae *et al.*[28] describe a neural-network-based system to estimate the pan and tilt of a person's head. In another technique as described by V. Kruger *et al.* [19], Gabor wavelet networks were developed combining the rotation-invariant properties of radial basis function networks with the feature extraction of Gabor filters. However, neural networks cannot work well for previously unseen changes. Each neural net, once trained, can do only what it has learned to do. Neural net learning algorithms involve a large number of training examples and long training periods compared to their symbolic counterparts.

2.2.3 Support Vector Machines

Support vector machines (SVM) have also been used for pose discrimination as illustrated by J. Huang *et al.* in [18]. SVMs are currently considered slower at run time than other techniques with similar generalization performance.

2.2.4 Principal Component Analysis

In another appearance-based approach, principal component analysis (PCA) represents a face as a linear combination of weighted eigenvectors, known as eigenfaces. Faces rotating across views form continuous manifolds in a pose eigenspace (PES) [15]. There are essentially two ways of approaching the problem of pose estimation in an eigenspace framework: view-based and parametric.

Pentland *et al.* and Darrell *et al.* [10] have suggested a view-based PCA approach to face recognition under varying pose and face pose estimation. In this approach, a

separate set of eigenfaces is computed for each possible pose. The pose is identified by computing the eigenspace projection of the input image onto each eigenspace and selecting the one with the lowest residual error. View-based representations of human faces using sets of 2D views, rather than explicit 3D models, are becoming increasingly attractive for computer vision. Their popularity is partly due to the fact that computation can be simplified by avoiding the need to build 3D models or to perform explicit 3D reconstruction. In addition, view-based representations do not directly encode prior knowledge of 3D shape. An important consequence of this representation is that it can be learned directly from a (possibly labeled) set of images. However, the resulting surface is a highly complex and non-separable manifold because the appearance of the face is the combined effect of its shape, sensor parameters, pose in the scene, and illumination conditions.

With the parametric eigenspace approach, given N individuals under M different poses, one can perform recognition and pose estimation in a universal eigenspace computed from the combination of $N \times M$ images. In this way a single parametric eigenspace will encode both identity and pose. Such an approach has recently been used by Murase and Nayar [25] for general 3D object recognition and pose estimation. J. Sherrah *et al.* [30] used orientation-selective Gabor filters to enhance differences in pose before projecting the faces on the pose eigenspace.

2.3 Model-based Approaches

Model-based approaches map the face to a 3D model to estimate pose of the head. Some model-based methods [16, 17, 21] estimate the 3D pose from the 2D pose by detecting features and matching model features with them. Horprasert *et al.* [17] presented an approach for computation of head orientation by employing face

symmetry and statistical analysis of face structure from anthropometry. Detection of facial features in this manner is not robust and is difficult. 3D model-based head pose estimation based on color of the entire face has also been used [3].

In [8], a method is outlined to estimate the 3D pose of human heads in a single image. It uses color information and fuzzy theory to extract the skin region and the hair region and to detect faces in images. Geometrical properties such as the area, center, and axis of least inertia of the skin region and the hair region are calculated and then used to estimate the 3D pose of the head. However, color models are sensitive to background and illumination.

2.4 Other Approaches

In another technique, optical flow is used to determine positions of features. An example of the optical flow technique is shown in [23]. In the optical flow technique, the sequence of images is used to follow features and determine a change in position from one image to the next. This technique requires fast processing so that the head movements between images are small. It also requires significant processing power in order to meet these processing speed requirements.

A Harvard University project [36] used the hairline contour within the hair region for pose estimation. This contour is generated by clustering brightness color and connectivity into hair blobs. The hairline is horizontally segmented into six regions and a six-point feature vector is formed. The orientation of the head is estimated using a Bayesian classifier for optimal mapping from the feature vector to the orientation angle.

In [4], a representation technique for 3D objects is presented unifying both the viewer-centered and model-centered object representation approaches. This approach,

called volumetric frequency representation (VFR) that encapsulates both the spatial structure and a continuum of the 2D discrete Fourier transform views of the model in the same data structure. Pose estimation is carried out using a VFR model constructed from the range data of a person and gray level images to index into the model. The pose of a given face image is determined by correlating the intensity image signature with the VFR in the 4D pose space. The pose estimation errors are quite low at about 4.05 degrees in azimuth, 5.63 degrees in elevation, and 2.68 degrees in rotation. However, this approach is not very practical and cannot be used for real time application because it is slow and uses laser range scanners.

In spite of all the advances in head pose recognition, a need exists for a head pose determining system that is robust even in an uncontrolled environment and, at the same time, has real-time capability. It should be low in cost and not use laser range scanners or require high-speed processors or special hardware.

Chapter 3

Motivation

Human motion capture is the process of recording human or other movement in physical space and transforming that information in a computer-usable form. With few exceptions (*e.g.*, finger manipulation), human body motion has many constraints. Some of the constraints come from the laws of physics; for example, the head cannot turn 360 degrees without moving the rest of the body. Gaits are another example of highly constrained motion. Some physical constraints induce a rhythmic and repetitive pattern of motion. By recognizing the presence of these constraints it is possible to recognize the motion and pose of the head. We investigate the methods used for human motion capture to influence the design of our algorithm.

Recently, comprehensive surveys of vision-based human motion capture have been published [13, 24]. The method suggested by Campbell and Bobick [7] is impressive. They developed techniques for representation of body movements based on space curves in subspaces of a phase space. Phase space is a Euclidean space with axes for each of the independent variables of a system and their time derivatives. Each point in phase space represents a state of the system, and, as the system evolves over time, the point moves along a phase path. If the parameters of the system do not change with time, then only one phase path passes through each point in the

space. Thus two coincident phase paths represent the same motion. Each different categorical movement embodies a different set of constraints on the motion of the body parts. These constraints are most easily observed in a phase space that relates the independent variables of the body motion.

Motivated by the success of using phase space to learn classical ballet steps, we choose to identify the properties for an efficient face subspace representation and to build face pose distribution curves. Marr and Nishara [22] consider representations of 3D shapes for object recognition and present five criteria: accessibility, scope, uniqueness, stability, and sensitivity. Accessibility refers to the ease in computing the representation from image data; scope is the class of movements for which the representation is designed; uniqueness corresponds to the property that a particular movement should always result in the same unique description; stability allows additions of images even after the representation is built; and sensitivity refers to discrimination between instances with respect to minor changes in the head movement (perturbation theory). Any representation will need to allow a compromise between stability and sensitivity. There are other questions to be considered. Because two different people will not transverse identical paths, questions remain about the uniqueness of the representation: Can two motions judged to be the same by a human observer also be placed in the same category by a system using this representation? In other words, is there a way to represent both the commonality and the differences between movements? Does the degree of similarity in the representation reflect similarity in the motions? Can subtle differences be expressed in the representation? Is the representation robust in the presence of missing data?

Principal component analysis (PCA), linear discriminant analysis (LDA), or independent component analysis (ICA) could be used as face representations. PCA is a

widely used representation model for faces and has also been used to model poses. LDA is useful when we would expect clusters in the representation. Because we expect the face pose distribution curve to be smooth, LDA may not be the optimal representation for our purposes. ICA is generally useful for blind source separation and is a generalization of PCA and is not a popular face representation. We thus use PCA as our face representation model.

We were inspired by the work proposed by Nayar *et al.* [25] who used PCA to project face poses on a pose eigenspace for pose estimation. This thesis proposes the use of wavelet transform as a pre-processor to enhance technique and improve performance. In the past, Englehart [11] has used a wavelet packet-based feature set in conjunction with PCA to improve myoelectric signal classification. We explore the application of wavelet transform combined with PCA to improve the performance of pose estimation.

Chapter 4

Review of PCA and Wavelets

4.1 Principal Component Analysis

4.1.1 An Overview of PCA

Principal component analysis (PCA) [34] involves a mathematical procedure that transforms a number of (possibly) correlated variables into a smaller number of uncorrelated variables called principal components. The principal components are the eigenvectors of the covariance matrix. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The second principal component is constrained to lie in the subspace perpendicular to the first principal component. Within that subspace, it points in the direction of the maximum variance. Then, the third principal component points in the maximum variance direction in the subspace perpendicular to the first two, and so on. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix, and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix. There can be as many principal components as

there are variables.

Let the training set of face images be $\Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_M$. The average face of the set is defined by $\Psi = \frac{1}{M} \sum_{n=1}^{M} \Gamma_n$. Each face differs from the average by the vector $\Phi_i = \Gamma_i - \Psi$.

The covariance matrix is $C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$ where A=[Φ_1 Φ_2 . . . Φ_M]. The covariance matrix for a set of faces is highly non-diagonal. The goal is to construct a face space, where each component is not correlated with any other component. This means that the covariance matrix of the new components should be diagonal. This diagonal form of the covariance matrix implies that the variance of a variable with itself will be maximized and the covariance of a variable with any other variable will be zero. The variables will no longer be correlated. Thus this construction finds those directions which maximize the variance.

The amount of information that the i_{th} principal component carries is given by its eigenvalue, λ_i. The basic approach is to compute the eigenvectors of the covariance matrix and to approximate the original data by a linear combination of the eigenvectors. Let λ_n n=1,. . . ,D, denote the eigenvectors, and let k be the $D \times D$ matrix whose columns are the vectors $\lambda_1, \ldots, \lambda_D$. In many applications, the eigenvectors in k are sorted in descending order according to the eigenvalues.

4.1.2 Limitations of PCA

Standard PCA-based methods suffer from two limitations: proper alignment of faces and large computational load. PCA assumes that all the faces are properly aligned. The faces are typically normalized by aligning the eyes of the training images. However, this is a tedious job and requires a manual operator to align the faces. Exact alignment is even more difficult for profile views of the head.

The second problem in the PCA-based method is the high computational load in finding the eigenvectors. The computational cost for finding eigenvectors from the covariance matrix C is cubic in complexity, i.e., $\mathcal{O}(D^3)$. The typical image resolution is 128×128, *i.e.*, $D = 320 \times 240$.

If the number of training images, M, is smaller than the value of D, the computational complexity will be reduced to $\mathcal{O}(M^3)$. If M increases, the computational load will be increased in cubic order. The computation complexity can be expressed as $\mathcal{O}(r^3)$, where $r = min(M,D)$.

Discrete wavelet transform (DWT) used as a preprocessor can help reduce the above limitations and increase the performance. Applying DWT on all the training images and selecting a low resolution sub-band allows further dimensionality reduction. The computational load on PCA is significantly reduced if DWT is used as a preprocessor. PCA achieves reduction in dimensionality by using statistical features of the training data in comparison with DWT, which uses low resolution to achieve dimensionality reduction.

DWT also provides immunity against exact alignment of faces. Englehart [11] showed that using DWT prior to PCA gives better noise immunity and shift invariance. This thesis investigates the use of DWT in conjunction with PCA to provide a robust framework for PCA.

4.2 Wavelet Transform

4.2.1 Why Wavelets?

Discrete wavelet transform has been a popular tool for image analysis in the past ten years. The advantages of DWT, such as speed and frequency localizations, have been discussed in many research articles. By decomposing an image using DWT, the

resolutions of the sub-band images are reduced. In turn, the computational complexity will be reduced dramatically by working on a lower resolution image. Wavelet decomposition provides local information in both space domain and frequency domain.

4.2.2 What Are Wavelets?

Wavelets are functions that satisfy certain mathematical requirements and are used in representing data or other functions. This idea is not new. Approximation using superimposition of functions has existed since the early 1800s, when Joseph Fourier discovered that he could superimpose sines and cosines to represent other functions. The wavelet analysis procedure involves adopting a wavelet prototype function, called an analyzing wavelet or mother wavelet. Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, and frequency analysis is performed with a dilated, low-frequency version of the prototype wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients. Instead of decomposing a signal using sines and cosines of different frequencies, we use dilates and shifts of the wavelet and the scaling function (implemented in the form of digital filter banks).

The wavelet function extracts information about differences between adjacent positions within the data. Figure 4.1 shows the shape of a 2D wavelet. A companion function which retains information about the averages between adjacent points is the scaling function. In the multi-resolution analysis framework, the orthogonal wavelet transform is based on the scaling function. The scaling function is a continuous and, in general, real-valued function on the set of real numbers. The scaling function is

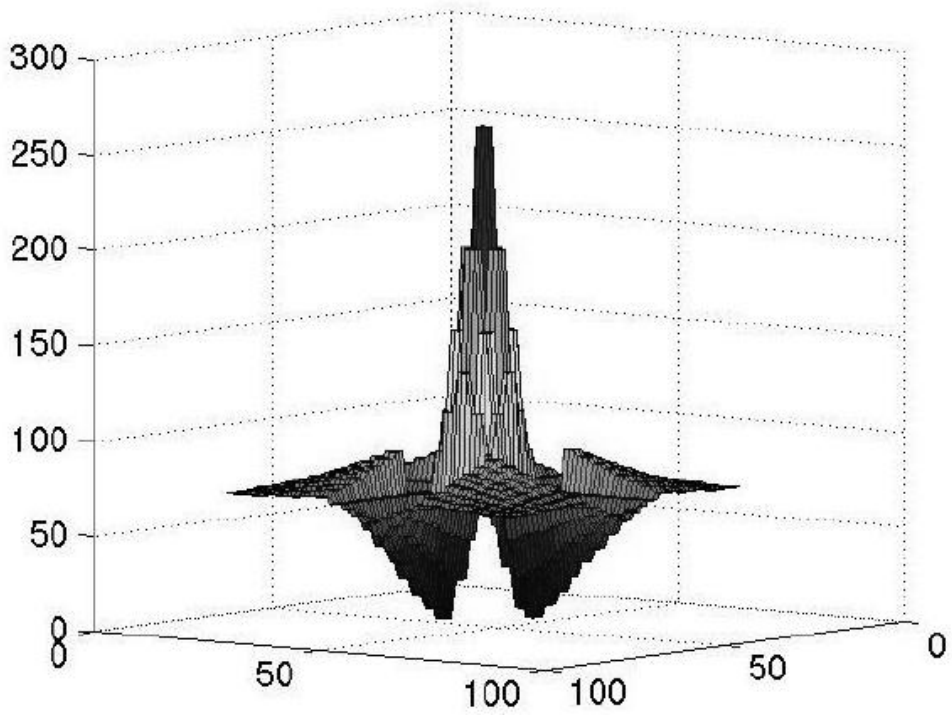

Figure 4.1: 2D plot of wavelet function.

different from the wavelet. It does not satisfy the wavelet admissibility condition. The scaling function plays the role of the average function. The correlation between the scaling function and a continuous function produces the averaged approximation of the function.

The orthogonal wavelet bases are generated from the scaling functions bases. The wavelet basis is orthogonal to the scaling basis within the same scale. These two series of coefficients are each half as long as the original data set. The smoothed data series corresponds to the coefficients generated from inner products with the scale function. The detail of the data series corresponds to the coefficients generated from the inner products with the wavelet function. Therefore, an original data set of 512 values will be decomposed into two series of 256 values. When the results are saved to a file, the smoothed data is saved first, followed by the detail of the data. This transform

will be called a first-level transform, and it applies the wavelet function at only the smallest scale. If the above analysis is done along with an analysis using the function expanded by a factor of two, the result will be called a second-level transform.

4.2.3 Lifting Scheme

The lifting scheme [31, 32] is a new flexible tool for constructing wavelets and wavelet transforms that does not rely on the Fourier transform. Lifting can be used to construct second generation wavelets, *i.e.*, wavelets which are not necessarily translates and dilates of one function. The latter we refer to as first-generation wavelets. In the case of first-generation wavelets, the lifting scheme will never produce wavelets that could not be found by the Cohen-Daubechies-Feauveau machinery. Nevertheless, it has the following advantages:

1. It allows a faster implementation of the wavelet transform. Traditionally, the fast wavelet transform is calculated with a two-band sub-band transform scheme. In each step, the signal is split into a high-pass and a low-pass band and is then subsampled. Recursion occurs on the low-pass band. The lifting scheme makes optimal use of similarities between the high-pass and the low-pass filters to speed up the calculation. The number of operations can be reduced by a factor of two.

2. The lifting scheme allows a fully in-place calculation of the wavelet transform. In other words, no auxiliary memory is needed, and the original signal image can be replaced with its wavelet transform.

3. In the classical case, it is not immediately clear that the inverse wavelet transform actually is the inverse of the forward transform. Only with the Fourier

transform can one be certain of the perfect reconstruction property. With the lifting scheme, the inverse wavelet transform can be found immediately by undoing the operations of the forward transform. In practice, this comes down to simply changing each plus into a minus sign.

4.2.4 2D Wavelet Transform

The implementation of DWT is carried out by applying one dimensional transform to the rows of the original image data and to the columns of the row transformed data. Wavelet decomposition provides local information in both space domain and frequency domain. For simplicity, we borrow the notation from the filtering literature [35]. The letter L stands for low frequency, and the letter H stands for high frequency. The left upper band is called LL band because it contains low frequency information in both the row and column directions. The LL band is a coarser overall information about the whole image because it is essentially a downsampled version of the original image. The LH sub-band is the result of applying the filter bank column wise, and it extracts the facial features. The HL sub-band, which is the result of applying the filter bank row wise, extracts the outline of the face boundary. Although the HH band shows the high frequency component of the image in non-horizontal and non-vertical directions, it was not considered to contain significant information about the face. The first-level decomposition of an image is shown in Figure 4.2. Further decomposition is conducted on the LL sub-band. Despite the equal sub-band sizes, different sub-bands carry different amounts of information. This observation was made at all resolutions of the image [12].

Figure 4.2: Wavelet transform of an image.

Chapter 5

Algorithm

The system as outlined in Figure 5.1 consists mainly of two stages: the training stage and the classification stage.

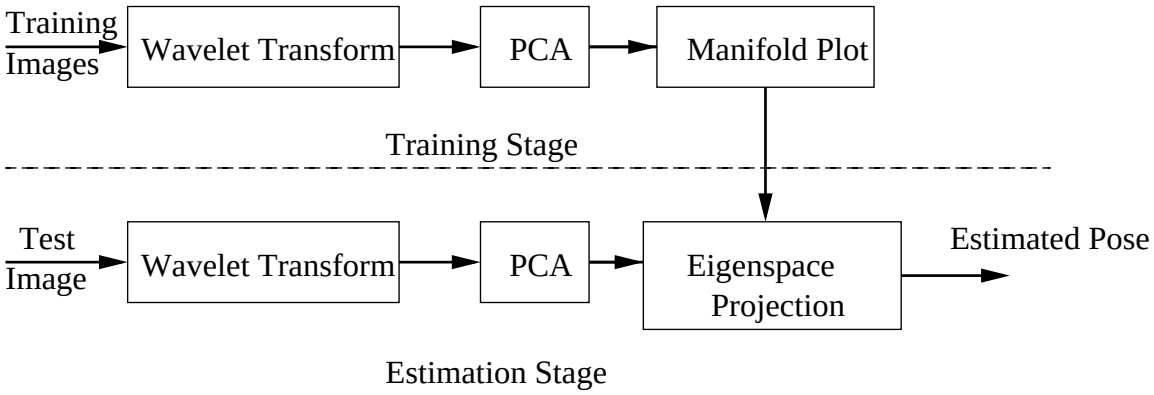

Figure 5.1: System block diagram.

Wavelet transform and PCA are the common blocks in the training and test stage. Both training and test images are preprocessed using wavelet transform. Wavelet transform uses the same wavelet function and level for training and test images. PCA then projects these low resolution subbands onto the Pose eigenspace. The following subsections explain the individual blocks in the block diagram in more detail.

The Training Stage

5.1 Acquisition of Training Data

The training stage of the algorithm can be either in online or offline mode. In the offline mode, an MPEG video clip of a person in various poses is captured using a Logitech QuickCam Express webcamera. After the video clip is finished recording, the MPEG video is then decompressed into frames. A typical decompressed video clip of a person moving his head from left to right would be composed of 600 to 800 frames. However, the training data typically consists of 63 gray scale images of size 240×352. If a smaller number of learning poses is used, classification tends to be unreliable when the test images correspond to poses that lie inbetween the learning poses. If the pose increments used in the learning stage are small, we obtain a larger number of learning samples and a larger number of points on the parametric manifold. If the person slows down the motion of the head in a particular pose, more closely spaced points are obtained in that region on the manifold. We manually select images from the decompressed sequence and evenly distribute the images among nine classes: extreme left, left, mid-left, center-left, center, center-right, mid-right, right, and extreme right. Similar classes are defined for up-down sequences of the head, although up-down head motion has less degree of rotation than left-right motion. Manual selection of images introduces human error in assigning classes to the training images. Vector quantization would be more appropriate in assigning the training images to the classes. Although using all the decompressed images instead of sampling the images increases the computational load on training, the percentage increase in performance is interesting to observe. To achieve real-time performance, we selected 63 images for training to obtain 63 three-dimensional data points on the manifold.

In the online mode, training images are stored in an image buffer, and no decompression is required. Microsoft Vision SDK was used to acquire images in real time from the USB camera for the online mode.

The image sequence is captured with a relatively plain static background because background changes in the query image affect the performance of the algorithm. Instead of segmenting the face using face detection techniques, we use the whole image to make the algorithm independent of the performance of the face detection routines.

5.2 Discrete Wavelet Transform

5.2.1 Introduction

Discrete wavelet transform is applied to all the images in the training set. In most wavelet transform applications, the original signal must be synthesized from the wavelet coefficients. This condition is referred to as perfect reconstruction. In some cases, however, such as pattern recognition, this requirement can be relaxed. In general, the goal of most modern wavelet research is to create a mother wavelet function that will give an informative, efficient, and useful description of the signal of interest. It is not easy to design a uniform procedure for developing the best mother wavelet or wavelet transform for a given class of signals. Thus there does not exist a unique wavelet that can be used for constructing representations of all poses in the database so that these representations are different enough to classify the poses with a desired dissimilarity margin between them. Choosing a wavelet for face pose estimation thus depends heavily on trial and error.

A lifting scheme [32] is used to implement a CDF (2,2) bi-orthogonal wavelet. This wavelet filter is an finite impulse response (FIR) filter with compact support. The

lifting scheme is preferred over the conventional convolution approach because of its advantages [31]. Gabor wavelets seem to be the most probable candidate for feature extraction, but they suffer from certain limitations. They cannot be implemented using the lifting scheme, and they form a non-orthogonal set, making the computation of wavelet coefficients difficult and expensive. Special hardware such as DataCube MaxRevolution is required to make the algorithm work in real time. Also, automatic selection of the resolution level with Gabor wavelets is not easy. Images with an aspect ratio other than 1:1 are considered, carefully taking into account the boundary conditions using reflection at the edges.

5.2.2 Selection of Sub-band and Resolution Level

The computational complexity of a traditional PCA-based method is in the cubic order of image resolution or number of training images, depending on which value is smaller. To minimize the computational complexity, we prefer to work with low resolution sub-bands which have sufficient information for pose estimation. In turn, working on a lower resolution image will reduce dramatically the computational complexity. We compute the energy in the sub-bands to measure the information content in the sub-bands. It is well known that the magnitudes of wavelet coefficients of a signal at each resolution level are proportional to the distributed energy of the signal at the associated frequency band. Because our representation is constructed based on these coefficients, its magnitudes at each level are also related to the distributed energy. If sufficient information is not preserved in the down-sampled images, the PCA block will not function properly because there is insufficient information. The energy in the sub-bands represents the amount of information, which includes primarily information about pose. Ideally, we have to preserve information which represents

only pose. If we preserve the whole information (image without wavelet transform), we also include information about illumination and other parameters. We wish to exclude this information while keeping sufficient information about the pose in the image. The information in a sub-band cannot be neglected if it is comparable to the information in other sub-bands. Figure 5.2 shows the distribution of energy in the LL sub-band at different levels of the wavelet transform. Because most of the energy is stored in the LL sub-band, we choose to project the LL sub-band onto the Pose eigenspace.

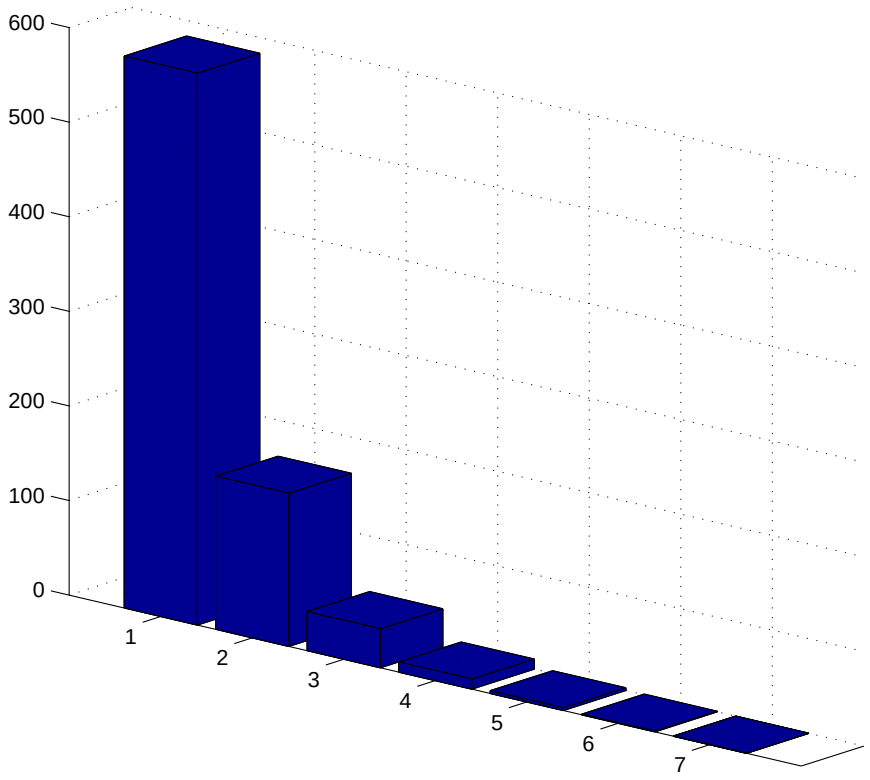

Figure 5.2: Distribution of energy of the image in the LL sub-band.

Figure 5.3 shows the distribution of energy among the LH, HL, and HH sub-bands at different levels of the wavelet transform. From Figure 5.3 we observe that the information in sub-band LH is comparable to the information in the LL sub-band

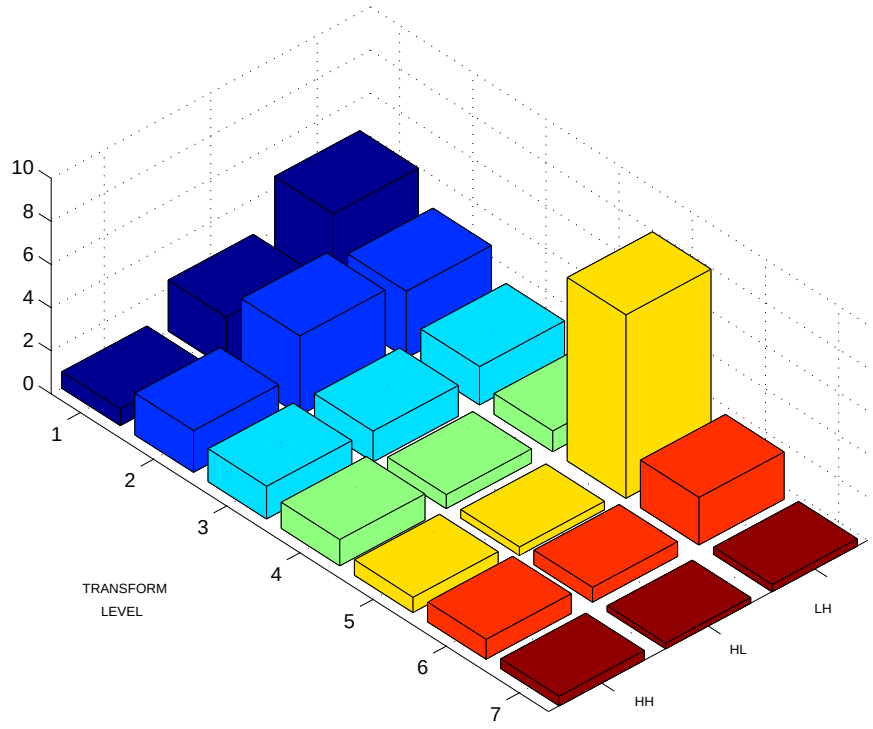

Figure 5.3: Energy concentration in wavelet sub-bands.

at a wavelet transform level greater than five. Thus we cannot select the LL sub-band at a level greater than five because a part of the information is also stored in other sub-bands. The LL sub-band is not shown in the Figure 5.3 since most of the energy is stored in the LL sub-band, and the information then seems negligible in the LH, HL, and HH sub-bands. In the case of face images, the LH sub-band contains information about the facial features, and the HL sub-band contains information about the facial contour, which cannot be neglected for pose classification. Thus the optimum choice of sub-band is the sub-band with a sufficient amount of energy that is the most predominant sub-band at that level of the DWT. The LL sub-band at level four seems to be the optimum choice. However, it was observed that, although the performance of the algorithm was the same for levels three and four, the algorithm performed poorly in certain difficult cases. Thus we conclude that the

energy preserved at level four was not sufficient for pose classification. Going back to level three gave the optimum performance. It was observed that the combined energy in the LH, HL, HH sub-bands was 30% of the energy in the LL sub-band at level three of the wavelet transform. Previously low-resolution information has been used for face pose estimation [26], recognition [9], and gender classification [33].

5.3 Principal Component Analysis

In the second stage of training, PCA [34] is applied on the low-frequency LL sub-band of resolution 30×44 instead of on the original image resolution of 320×240. The output of this step will be a set of eigenvectors and eigenvalues. These eigenvectors constitute the dimensions of the eigenspace.

5.4 Manifold Plot

By arranging the eigenvalues in descending order, we select three eigenvectors with the largest eigenvalues. The first three eigenvectors encode information about the pose. The rest of the eigenvectors encode information about facial expression and other details. The performance declines when more than three eigenvectors are used. Thus all images in the training set are represented by a linear combination of three representational bases by projecting then into the pose eigenspace. All training images are projected on the eigenspace to obtain a set of points in a 3D eigenspace. These points lie on a manifold that is parameterized by pose. Figure 5.4 shows the pose distribution curves of the training and the query images. Each point on the curve represents an image projected on the pose eigenspace.

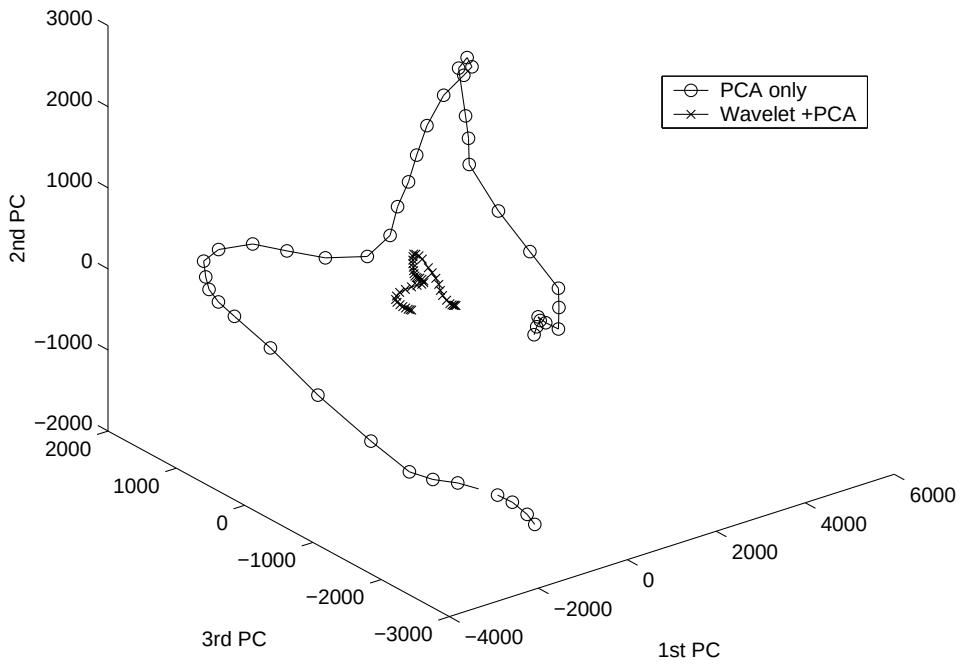

Figure 5.4: Face pose distribution curve for a 45-frame sequence.

The Classification Stage

In the estimation or the classification stage, we find the wavelet transform of the image to classify the pose of the query image, and we project the LL sub-band of resolution 30×4 onto the pose eigenspace using the first three eigenfaces. We use the L2 distance to find the distance between the first three eigencoefficients of the query image from the manifold points in the 3D space. The L2 or Manhattan distance is used to keep the computations to a minimum without compromising the performance. This is the simplest distance to calculate and is more robust to outliers. When a test image is discriminated to the closest image in the database, a simple mapping is done to assign a class to the test image. Because the face pose distribution curve is a smooth curve, clustering is not required for assigning the class.

Chapter 6

Performance

Because face pose estimation is so new, there are no standard face sequences yet available for comparison. A number of experiments were done to test the robustness of the algorithm and to increase the classification accuracy. Selecting low-resolution subbands makes the intensity image insensitive to changes in facial expression, lighting, occlusion, and distance from the camera.

Two sets of experiments were done on all the images. In the first set, the faces were projected on the pose eigenspace without any preprocessing. Figure 6.1 shows the pose distribution curve when PCA was used without any preprocessing. The query image was captured under different illumination.

In the second set of experiments, wavelet transform was applied to all the images and a low-resolution subband was projected onto the pose eigenspace. Ideal results would be obtained if the query manifold coincides with the training distribution curve. The closer and more similar in shape the query curve is to the training manifold, the more accurate the discrimination. Figure 6.2 shows the manifold for training images which were preprocessed using wavelet transform and projected onto the pose eigenspace.

Figure 6.3 shows the pose distribution curve for a person wearing glasses in the

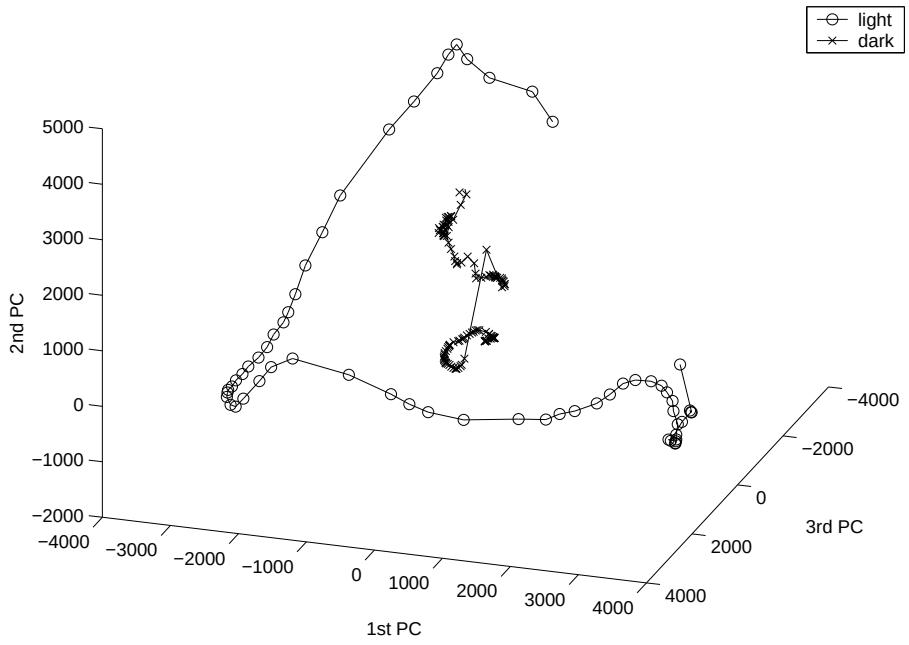

Figure 6.1: Manifold curve: PCA only.

query images. The training set of images didn't have the subject wearing glasses. 100% accurate results were obtained in the case of query images with such minor changes.

In Figure 6.4 the query manifold and the pose distribution curve vary widely, suggesting that the system is not robust against identity changes. In most of the cases, the query image returned the pose class as looking straight ahead.

Experiments were done to test the performance of the system at different levels of the wavelet transform. Figure 6.5 shows the pose distribution curve when the training and the query images were preprocessed using DWT of level 4 instead of level 3.

In an uncontrolled environment, the distance between the subject and the camera frequently changes. An experiment was done in which the subject kept moving away from the camera, and query images were sampled at varying distance from the camera.

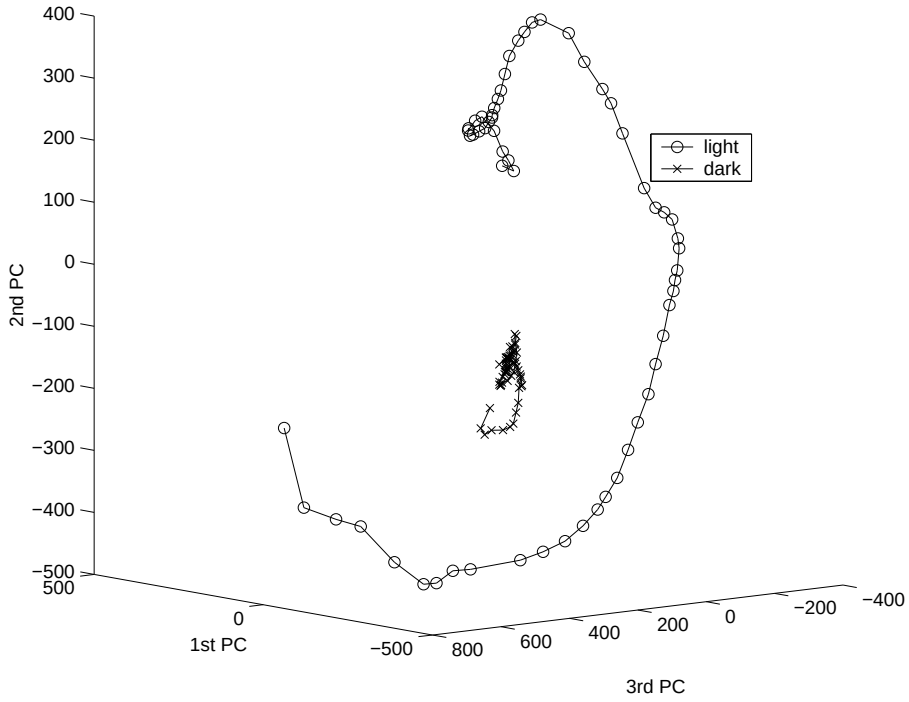

Figure 6.2: Manifold curve: wavelet and PCA.

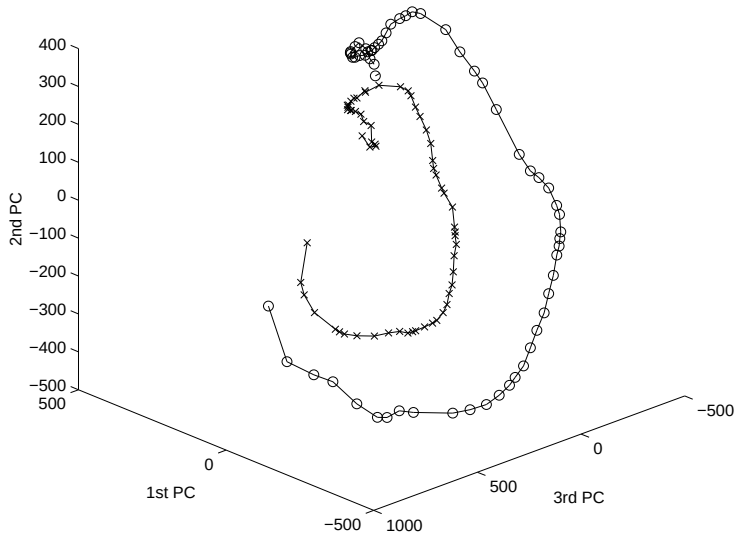

Figure 6.3: Manifold curve: a person wearing glasses with and without wavelets.

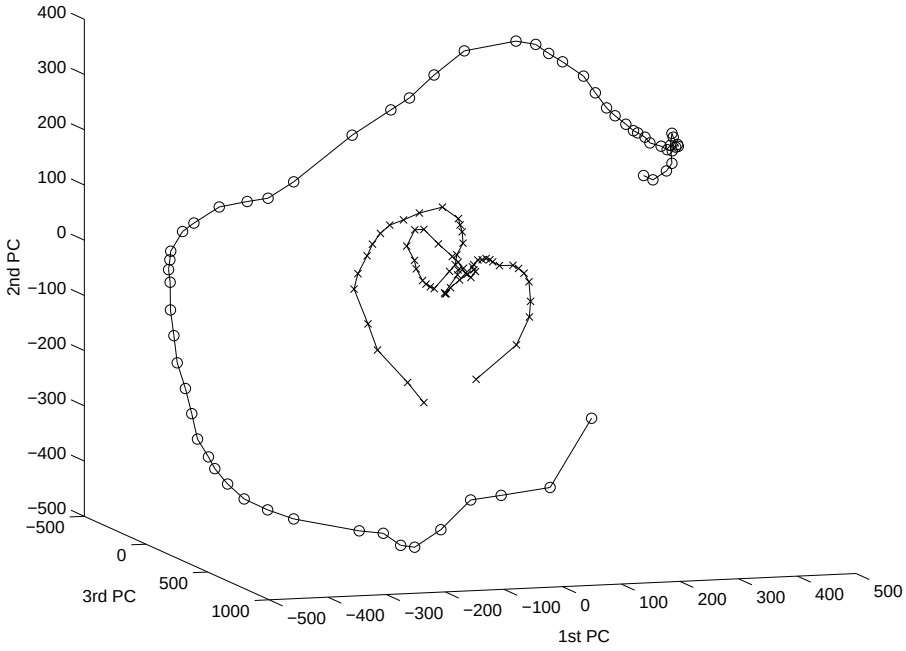

Figure 6.4: Manifold curve: persons with different identity.

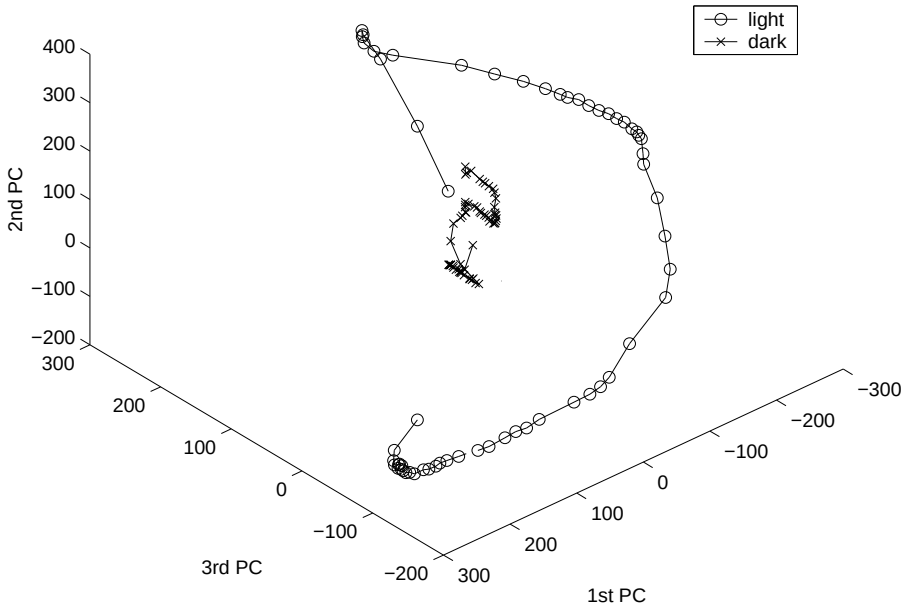

Figure 6.5: Manifold curve: different illumination using DWT level 4.

The manifold of these query images is shown in Figure 6.6.

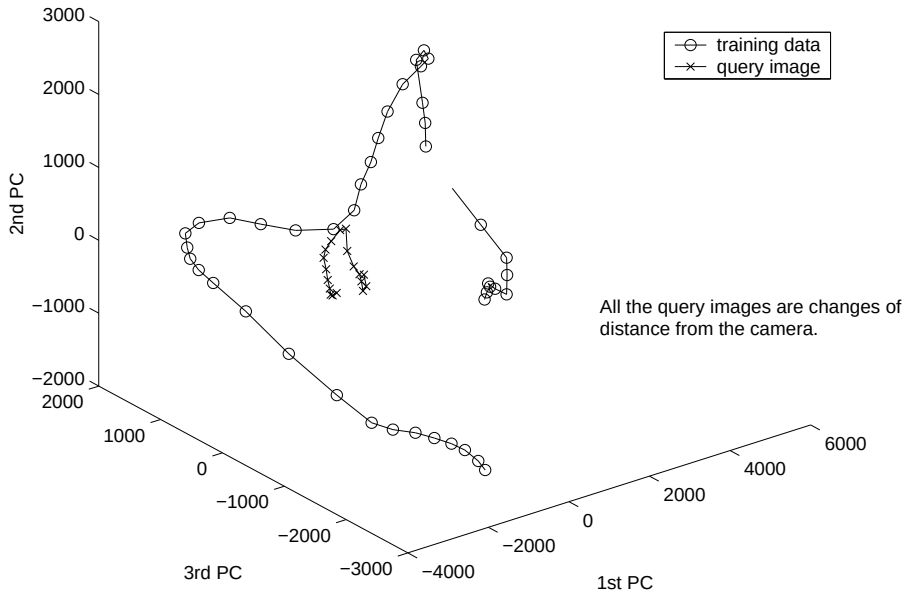

Figure 6.6: Manifold curve: changing distance from the camera.

In the above experiment, the procedure was repeated and performance was evaluated with normalized PCA coefficients as shown in Figure 6.7.

In Figure 6.8 we discriminate face pose with different facial expressions and illuminations and show the closest match. The results clearly show the robustness of the system.

A confusion matrix as shown in Table 6.1 was used to test the performance of classification. The confusion matrix is a 9×9 matrix because there are 9 classes. The algorithm performed fairly well with an accuracy of 84% for images under different environment conditions than those included while training. An accuracy of 100% was observed for poses taken from the same video sequence and not included in the training data set. PCA performed poorly, giving an accuracy of 68%. We also

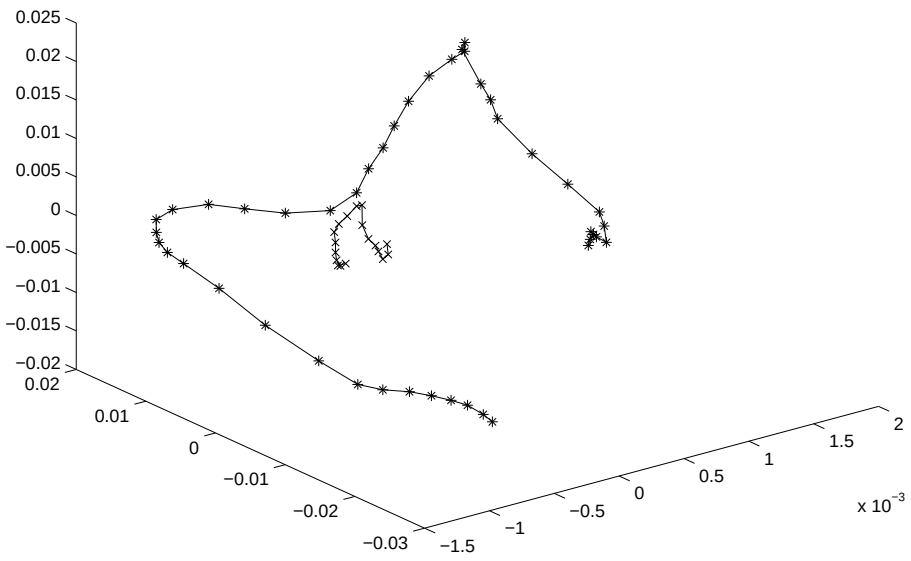

Figure 6.7: Manifold curve: changing distance using normalized PCA coefficients.

tested the system on people who were not in the training set. The identity of the closest matched head constantly changed, but the pose of the best match generally matched well with the pose of the input image. When the distance between the camera changed, the algorithm performed poorly if an LH, HL, or HH sub-band was used. The LL sub-band gave the optimum performance. The results are not perfect, but they are reasonable. The results were tested with four training sequences and used a combination of 159 test images.

6.1 Timing Analysis

The system we proposed and implemented requires 17.26 seconds for training of 63 images compared to the traditional PCA method requirement of 337.6 seconds. This system is clearly more computationally efficient than the traditional PCA method. All experiments were performed on an Ultra 360 MHz Sun Workstation.

Figure 6.8: Test images on the left. Closest match is shown on the right.

	c1	c2	c3	c4	c5	c6	c7	c8	c9
c1	10	2	0	0	0	0	0	0	0
c2	1	10	1	0	0	0	0	0	0
c3	0	0	20	2	4	0	0	0	0
c4	0	0	0	15	2	0	0	0	0
c5	0	0	0	0	25	0	0	0	0
c6	0	0	0	1	4	15	0	0	0
c7	0	0	0	0	0	4	12	1	0
c8	0	0	0	0	0	0	0	15	2
c9	0	0	0	0	0	0	0	1	12

Table 6.1: Confusion Matrix.

6.2 Discussion on Experiments and Results

The algorithm failed (as expected) when the LH, HL, HH sub-bands were projected on the pose eigenspace. Because PCA exploits the second-order correlation among the data samples, there was no strong correlation among the wavelet coefficients in these sub-bands. There has not been any published study yet that shows exactly how wavelet coefficients are related to the coefficients in their neighborhoods, to other coefficients within the same sub-band, or to coefficients located at the same scale in other sub-bands at the same level of the transform. Embedded Zero Tree Wavelet [29] exploits the relationships among wavelet coefficients in the same sub-band at different levels of the transform but doesn't take into account the relation among coefficients with respect to other wavelet coefficients at the same level of transform. Exploiting the relation among wavelet coefficients could take the step of data compression further and provide a technique to better exploit information in other sub-bands.

Wavelet transform is used to obtain a low-resolution sub-band, which reduces the computational load on PCA. However, it appears that using local averaging techniques to obtain a low-resolution sub-band might suffice because we are not using information in other sub-bands obtained using wavelet transform. Local averaging would work, too, if the size of the image is always fixed. However, we have provided a framework based on the energy of wavelet coefficients to compute the most efficient resolution required for pose discrimination, and there are other advantages if DWT is used instead of local averaging. Englehart [11] has shown how wavelet combined with PCA helps in filtering noise from ECG signals. A similar advantage is achieved in the case of pose discrimination. Wavelet transform relaxes the restriction of exact alignment of faces, which is otherwise a required condition for using PCA.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis has outlined a wavelet sub-band approach using PCA for face pose classification. Wavelet transform is adopted to decompose an image into multi-resolution sub-bands. A low-frequency sub-band is selected for PCA representation. Results show that the proposed method gives better accuracy and class separability in classifying the head poses than applying PCA on the whole original image containing all frequency components. Another significant advantage is the increase in computational speed. Our system is robust even in an uncontrolled environment and, at the same time, has real-time capability. No specialized hardware is required, providing a low cost solution to the face pose discrimination problem.

7.2 Limitations

Although this thesis considers robustness against changes in environment, no attempt was made to evaluate the performance of the system in the case of partial occlusion of the face. The algorithm is not extremely robust against identity changes. If two persons of different races are considered, the algorithm fails because of the difference in face outline. To overcome this limitation, we propose using an ellipse to

fit the faces at low resolutions in the LL sub-bands. The inclination of the ellipse will give an indication of the pose and will not be dependant on face outliers.

If the change in distance from the camera is substantial, the point on the manifold always tends to be in the middle of the manifold, indicating central pose. This effect could be studied more once we use face detection instead of a static background. The multi-resolution capability of the algorithm will fix this problem.

7.3 Future Work

Wavelets combined with PCA provide a robust framework for PCA. Although discrete wavelet transform (DWT) is a powerful tool for signal and image processing, it has some limitations. Future work involves overcoming these limitations. Other issues for future consideration include:

- DWT is shift sensitive because input-signal shifts generate unpredictable changes in DWT coefficients. Undecimated wavelet transform could further improve the performance of the algorithm because of the shift invariance property. Using undecimated wavelet transform as a pre-processor for PCA can give immunity against noise and an allow for minor misalignment of faces.

- DWT suffers from poor directionality because DWT coefficients reveal only three spatial orientations. In its present form, we are using discrete wavelet transform to filter the effects of other distortions. Wavelets are not used like Gabor filters for feature extraction because the wavelet does not specifically enhance the pose. The choice of wavelet didn't seem to affect the performance much. Thus it is suggested that multi-wavelets be tried for better feature extraction along with the filtering of other data. Banana wavelets, and complex

wavelets could also be investigated for pose discrimination.

- Instead of using just the LL sub-band, a combination of LH, HL, and HH sub-bands at different scales and resolutions could be projected on the pose eigenspace to improve performance. Instead of computing the eigenvectors of the covariance matrix of wavelet coefficients in these sub-bands, a different statistical measure needs to be computed.

- This thesis relies totally on the visual inspection of the manifold curves and achieved results to justify the use of wavelets as a pre-processing block. A rigorous mathematical analysis could further validate our conclusions.

- It would be interesting to combine the underlying principles of wavelets and PCA to form a new framework. Instead of computing the eigenvectors of the covariance matrix of the wavelet coefficients in the LH, HL, and HH sub-bands, the new framework would compute the eigenvectors of a matrix that best defines the relationships of the wavelet coefficients among themselves.

- Faster search methods using k-trees should be used for searching for the shortest distance of the query point from the manifold.

- Other metrics such as entropy should also be used to select the resolution level.

- Genetic algorithms could be used to search for the most expressive features in the wavelet sub-bands for pose estimation.

- The same algorithm could be extended to object poses instead of faces poses. However, the manifold will no longer be 3D because more eigenvectors would be required to encode the object pose.

- It also would be interesting to combine simultaneous detection of face along with pose.

- The ultimate test of the capability of the system would be a graphic model mimicking the poses of a human face in real time.

- Hidden Markov models (HMM) have been used to classify body poses. They can also be used to model the face poses as states of the HMM.

- Vector quantization could be used for classifying the poses rather than manually assigning the classes. Thus, instead of sampling the images, all the images in the image sequence can be used in the training data set.

Bibliography

[1] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(6):562–575, 1995.

[2] Shumeet Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In *Neural Information Procesing systems (NIPS '98)*, 1998.

[3] P.A. Beardsley. Pose estimation of the human head by modelling with an ellipsoid. *IEEE Conf. on Automatic Face and Gesture Recognition*, 1998.

[4] J. Ben-Arie and D. Nandy. A volumetric/iconic frequency domain representation for objects with application for pose invariant face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(5):449–457, 1998.

[5] E. Borovikov. Human head pose estimation by facial features location. University of Maryland Institute for Computer Studies, College Park, MD.

[6] Roberto Brunelli. Estimation of pose and illuminant direction for face processing. Technical Report AIM-1499, 1994.

[7] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn*, page 8, 1995.

[8] Qian Chen, Tetsuo Shimada, Haiyuan Wu, and Tadayosi Shioyama. 3d head pose estimation using color information. *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, 1, 1998.

[9] Y. Dai and Y. Nakano. Recognition of facial images with low resolution using a hopfield memory model. *Pattern Recognition*, 31:159–167, 1998.

[10] T. Darrell, B. Moghaddam, and A. Pentland. Active face tracking and pose estimation in an interactive room. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'96)*, pages 67–72, San Francisco, CA, June 1996.

[11] Englehart. *Signal Representation for Classification of the Transient Myoelectric Signal.* PhD thesis, University of New Brunswick, Fredericton, Canada, 1998.

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

[13] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[14] A. H. Gee and R. Cipolla. Determining the gaze of face in images. Technical Report CUED/F-INFENG/TR 174, Trumpington Street, Cambridge CB2 1PZ, England, 1994.

[15] S. Gong, S. McKenna, and J. Collins. An Investigation into Face Pose Distributions. In *Second International Conference on Automated Face and Gesture Recognition*, Killington, Vermont, October 1996.

[16] Kazuyuki Hattori, Shinichi Matsumori, and Yukio Sato. Estimating pose of human face based on symmetry plane using range and intensity images. *14th International Conference on Pattern Recognition*, 2, 1998.

[17] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3-d head orientation from a monocular image sequence. *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, pages 242–247, 1996.

[18] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines. *Proc. of 14th Int'l Conf. on Pattern Recognition (ICPR'98)*, pages 154–156, 1998.

[19] V. Kruger and G. Sommer. Gabor wavelet networks for object representation. In *Proc. of the Int. Dagstuhl 2000 Workshop*, 2000.

[20] Volker Kruger, Sven Bruns, and Gerald Sommer. Efficient head pose estimation with gabor wavelet networks. *Proc. British Machine Vision Conference, Bristol, UK*, 2000.

[21] M. Marius, F. Preteux, and V. Buzuloiu. 3d global head pose estimation : A robust approach. In *International Workshop on Synthetic-Natural Hybrid Coding and Three-Dimensional Imaging*, 1999.

[22] D. Marr and H. K. Nishihara. Representation and reconition of the spatial organization of three dimensional shapes. *Proceedings of Royal Society of London B*, 200:269–294, 1978.

[23] T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. *Proceedings, Second International Conference on Automatic Face and Gesture Recognition*, pages 176–181, 1996.

[24] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 2001.

[25] S. Nayar, H. Murase, and S. Nene. Parametric appearance representation, 1996.

[26] S. Niyogi and W.T. Freeman. Example-based head tracking. Technical Report TR96-34, MERL Cambridge Research, 1996.

[27] T. Otsuka and J. Ohya. Real-time estimation of head motion using weak perspective epipolar geometry. *Proc. 4th IEEE Workshop on Application of Computer Vision*, pages 220–225, 1998.

[28] R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on Neural Networks*, 9(2):257–265, 1998.

[29] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. on Signal Processing*, 41(12):3445–3462, 1993.

[30] J. Sherrah, S. Gong, and E-J. Ong. Understanding pose discrimination in similarity space. *Proc. British Machine Vision Conference, Nottingham.*, 1999.

[31] Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, 1998.

[32] Wim Sweldens and Peter Schrder. Building your own wavelets at home. *Wavelets in Computer Graphics, ACM SIGGRAPH Course Notes*, pages 15–87, 1996.

[33] S. Tamura, H. Kawai, and H. Mitsumoto. Male/female identification from 8×6 very low resolution face images by neural network. *Pattern Recognition*, 29(2):331–335, 1996.

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

[35] Martin Vetterli. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[36] C. Wang and M. Brandstein. Robust head pose estimation by machine learning. *IEEE International Conference on Image Processing (ICIP)*, 2000.