Face Pose Estimation in Monocular Images

by

Muhammad Shafi

A Doctoral Thesis

Submitted in partial fulfilment of the requirements
for the award of Doctor of Philosophy of Loughborough University

January, 2010

ABSTRACT

People use orientation of their faces to convey rich, inter-personal information. For example, a person will direct his face to indicate who the intended target of the conversation is. Similarly in a conversation, face orientation is a non-verbal cue to listener when to switch role and start speaking, and a nod indicates that a person has understands, or agrees with, what is being said. Further more, face pose estimation plays an important role in human-computer interaction, virtual reality applications, human behaviour analysis, pose-independent face recognition, driver's vigilance assessment, gaze estimation, etc. Robust face recognition has been a focus of research in computer vision community for more than two decades. Although substantial research has been done and numerous methods have been proposed for face recognition, there remain challenges in this field. One of these is face recognition under varying poses and that is why face pose estimation is still an important research area.

In computer vision, face pose estimation is the process of inferring the face orientation from digital imagery. It requires a serious of image processing steps to transform a pixel-based representation of a human face into a high-level concept of direction. An ideal face pose estimator should be invariant to a variety of image-changing factors such as camera distortion, lighting condition, skin colour, projective geometry, facial hairs, facial expressions, presence of accessories like glasses and hats, etc.

Face pose estimation has been a focus of research for about two decades and numerous research contributions have been presented in this field. Face pose estimation techniques in literature have still some shortcomings and limitations in terms of accuracy, applicability to monocular images, being autonomous, identity and lighting variations, image resolution variations, range of face motion, computational expense, presence of facial hairs, presence of accessories like glasses and hats, etc. These shortcomings of existing face pose estimation techniques motivated the research work presented in this thesis. The main focus of this research is to design and develop novel face pose estimation algorithms that improve automatic face pose estimation in terms of processing time, computational expense, and invariance to different conditions.

To my parents,

for their inspiration,

for their love,

for their sacrifices.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Paul Chung for his guidance, encouragement, time and source of funding throughout my PhD. I am grateful to my brothers and sisters for their support, encouragement and sacrifices. Special thanks to my friend Muhammad Irfan Khattak for his continuous support and sincere help. I am indebted to Shufu Xie (Ph.D Candidate, Face Group, JDL, Chinese Academy of Sciences) for providing the CAS-PEAL pose database. Finally, I would like to thank all my friends especially, Amir Ehsan, Fayaz Ahmed, Mohammad Saleem, Muhammad Ishaq, Rizwan Faiz, Salman Hussain, Benoit Terrier and Zarish Khan for their support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 INTRODUCTION

In computer vision, face pose estimation is the process of inferring the face orientation from digital imagery. It requires a series of image processing steps to transform a pixel-based representation of a human face into a high-level concept of direction. An ideal face pose estimator should be invariant to a variety of image-changing factors such as camera distortion, lighting condition, skin colour, projective geometry, facial hairs, facial expressions, presence of accessories like glasses and hats, etc. In computer vision, it is common to assume that human head can be treated as a disembodied rigid object. Based on this assumption, a human face is limited to 3 degrees of freedom in pose, which can be represented by pitch, roll, and yaw angles as shown in Figure 1-1.

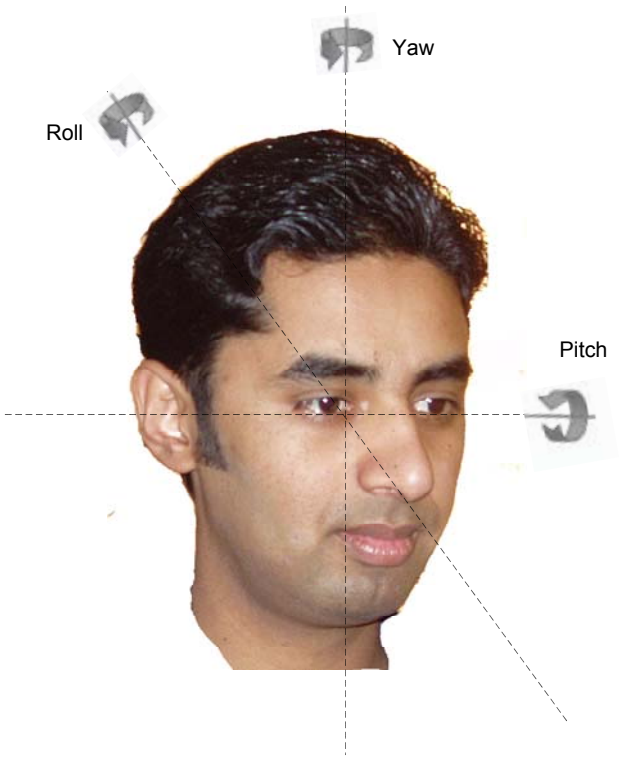


Figure 1-1 Three Degrees of Freedom of a Human Face i.e. Yaw, Pitch and Roll

This thesis focuses on improvements to existing face pose (specifically Yaw) estimation in terms of computational expense, processing time, and invariance to different conditions e.g. bad illumination, different skin colours, facial hairs, presence of glasses, etc. Since facial features localization is generally used as an intermediate step in face pose estimation, it is also considered as a subtopic in this thesis. The terms "face pose estimation", "head pose estimation" and "pose estimation" have been used interchangeably in the thesis.

1.2 RESEARCH MOTIVATION

People use orientation of their faces to convey rich, inter-personal information. For example, a person will direct his face to indicate who the intended target of the conversation is. Similarly in a conversation, face orientation is a non-verbal cue to listener when to switch role and start speaking, and a nod indicates that a person understands, or agrees with, what is being said. Further more, face pose estimation plays an important role in human-computer interaction, virtual reality applications, human behaviour analysis, pose-independent face recognition, driver's vigilance assessment, gaze estimation, etc. Robust face recognition has been a focus of research in computer vision community for more than two decades. Although substantial research has been done and numerous methods have been proposed for face recognition, there remain challenges in this field. One of these is face recognition under varying poses and that is why face pose estimation is still an important research area.

Face pose estimation has been a focus of research for about two decades and numerous research contributions have been presented in this field. According to Murphy-Chutorian and Trivedi (2009), face pose estimation schemes for static images can be classified into seven categories: Appearance Template Methods, Detector Array Methods, Non-linear Regression Methods, Manifold Embedding Methods, Flexible Models, Geometric Methods, and Hybrid Methods. In appearance template methods, the input image is compared with already created templates. The template that gives the maximum similarity is assigned as the pose to the input image. Appearance template methods have the advantages of being applicable to both low and high resolution images, also the set of templates can be expanded to adapt to changing conditions. The disadvantage of these methods is that they are capable of estimating discrete poses only. In detector array methods, multiple face detectors are trained, with each to a specific different discrete pose. For arrays of binary classifiers, successfully detecting the face will specify the pose of the

face while for detectors with continuous output; pose can be estimated by the detector with the greatest support. An advantage of these methods is that steps for head localization are not required; however, it is burdensome to train detectors for each discrete pose. Nonlinear regression methods use nonlinear functions to map the image space to one or more pose directions. The high dimensionality of an image is a challenge for regression tools and this kind of methods. In manifold embedding methods, a low dimensional manifold is modelled, and embedding technique is used to project a new sample to the manifold. The challenge lies in creating a dimensionality reduction algorithm that successfully recovers face pose while ignoring other variations of an image. The problem with this kind of methods is that there is no guarantee that primary components will relate to pose variations rather than appearance variations. In flexible models methods, a non-rigid model is fit to the image such that it conforms to the facial structure. These methods require training data with annotated facial features, but it enables them to do comparisons at feature level rather than the global image. Geometric methods use head shape and the precise locations of facial features to estimate pose. Geometric methods have the advantage of being fast and simple, but the facial features need to be localized precisely which make these methods vulnerable to bad illumination and low resolution. Hybrid methods combine two or more of the aforementioned techniques to estimate face pose. These systems are designed to overcome the limitations of each individual method but they can be computationally expensive.

All the aforementioned face pose estimation techniques have some shortcomings and limitations in terms of accuracy, applicability to monocular images, being autonomous, identity and lighting variations, image resolution variations, range of face motion, computational expense, presence of facial hairs, presence of accessories like glasses and hats, etc. These shortcomings of existing face pose estimation techniques motivated the research work presented in this thesis. The main focus of this research is to design and develop novel face pose estimation algorithms that improve automatic face pose estimation in terms of processing time, computational expense, and invariance to different conditions as already mentioned.

1.3 AIM AND OBJECTIVES

The aim of the research is to design and develop novel efficient and robust algorithms for face pose estimation in near-field images that improve the automatic face pose estimation in terms of processing time, computational expense, and invariance to different conditions.

The objectives of the research are:

- To study the existing face pose estimation methods, classify them into different groups, and identify their advantages and limitations.
- To study the existing standard pose databases and their capturing schemes in detail.
- To suggest novel improvements to existing pose estimation methods in terms of speed, computational expense, and invariance to different conditions, e.g. bad light, different skin colour etc.
- To suggest novel improvements to facial features (i.e. eyes and mouth) localization as facial feature localization is used as an intermediate step in the proposed pose estimation methods.
- To test the proposed pose estimation methods using publically available pose databases to measure the accuracy and compare with existing pose estimation methods.
- To identify the limitations of the proposed methods and outline future research directions.

1.4 CONTRIBUTIONS

The following original contributions have been made by the research presented in chapters 4, 5 and 6. A list of publications is included in Appendix B.

1.4.1 USE OF EDGE-DENSITY FOR EYES LOCALIZATION IN FACIAL IMAGES

The novel edge-density based eyes localization method (discussed in chapter 4) is based on the observation that edge-density is high in eye regions as compared to other regions in a face. Most colour variations occur in eye regions in facial images are due to colour difference between eyelids and skin, skin and sclera, sclera and iris, and iris and pupil. So if edge detection is applied to a facial image, high edge density is present in the eye

regions. After edge detection, morphological operators i.e. erosion and dilation are used to get connected regions. Shape and geometry based rules are then applied to these connected regions to extract and verify eyes. The proposed scheme was tested on PICS (http://pics.psych.stir.ac.uk/) images database with very good results.

1.4.2 USE OF A COMBINATION OF COLOUR, EDGE AND ILLUMINATION CUES FOR EYES LOCALIZATION IN FACIAL IMAGES

The proposed hybrid method (discussed in chapter 4) combines intensity, edge and illumination cues to form a more accurate eyes localization system. The system was tested using the PICS (http://pics.psych.stir.ac.uk/) images database. The result demonstrates that the proposed method overcomes the weaknesses of each of the individual methods.

1.4.3 USES OF EYES-MOUTH TRIANGLE SHAPE FOR FACE POSE ESTIMATION

This proposed method (discussed in chapter 5) is based on the observation that eyes-mouth triangle has a distinct shape for distinct yaw angle. The method uses the shape of eyes-mouth triangle to estimate face pose. Eyes and mouth centres are localized first using the methods discussed in chapter 4. To minimize inter-subject variations, the triangle obtained from the eyes and mouth centres is rotated and normalized, i.e. resized to a standard size. The shape of this normalized triangle is then used to estimate face pose. The proposed method has a number of advantages over existing methods. It is based on eyes and mouth which are the most salient features of a face. It is based on only three feature points, instead of five, which makes it faster computationally and hence suitable for real-time applications. Because it uses centres instead of corners of eyes and mouth, the range of pose estimation is increased. The proposed method was tested using Pointing ’04 (Gourier et al. 2004), CAS-PEAL (Gao et al. 2008) and an in-house (discussed in detail in appendix A) databases with very good results.

1.4.4 USE OF DISTANCE TRANSFORM AND NORMALIZED CROSS-CORRELATION FOR POSE ESTIMATION IN NEAR-FIELD IMAGES

The proposed method (discussed in chapter 6) uses distance transform and normalized cross-correlation for face pose estimation. Distance transform face pose template is first computed for each discrete yaw angle. To estimate the pose for an input image, the image is first converted to grey level if it is in colour. Distance transform is calculated and then it is resized and normalized so that the facial feature locations coincide with the already

created templates. The pose is then estimated using a normalized cross-correlation between this resized distance transform image and the set of templates. The distance transform property of being invariant to intensity makes the proposed method suitable for different skin colour because the skin colour variation appears as intensity variation in grey level images. Also, since the distance transform is relatively invariant to illumination, the proposed method is suitable for different illumination conditions. The proposed method is also relatively invariant to facial hairs and whether the subject is wearing glasses. The proposed method was tested using the CAS-PEAL pose database (Gao et al. (2008)). To show that the method works for different skin colours, the method was also tested using an in-house database (discussed in appendix A) which contains darker colour faces as compared to other publically available pose databases.

1.5 THESIS ORGANIZATION

This thesis is organized into seven chapters:

Chapter 1 provides an overview of the thesis while highlighting the research motivation, aim and objectives of the research, contributions of the thesis and thesis organization.

Chapter 2 presents a detailed review of face pose estimation from the literature. The face pose estimation methods are classified into different groups. Advantages and disadvantages of each group are discussed in detail.

Chapter 3 presents some of the fundamental image processing concepts which are utilized in the proposed approaches described in chapters 4, 5 and 6.

Chapter 4 presents the proposed facial feature localization methods with detailed experimental results and analysis.

Chapter 5 presents a proposed face pose estimation method based on the eyes-mouth triangle with detailed results and analysis.

Chapter 6 presents another proposed face pose estimation method based on distance transform and normalized cross-correlation with detailed results and analysis.

Chapter 7 concludes the research presented in the thesis with future research directions.

2 LITERATURE REVIEW

2.1 INTRODUCTION

Gee and Cipolla (1994) proposed one of the earliest pose estimation schemes and since then numerous contributions have been presented in this field. Murphy-Chutorian and Trivedi (2009) have recently conducted a very detailed and comprehensive survey of the head pose estimation schemes from literature. The pose estimation schemes can be divided into two broad categories: the ones which estimate pose in still images and the ones which work for videos while tracking face pose in each frame of a video. Some methods combine these two by estimating the pose in the first frame of a video using the first type of method and then tracking the pose in consecutive frames using the second type method. Since the focus of this research is pose estimation in still images, therefore, only the pose estimation schemes of the first type i.e. the one for static images, are discussed.

According to Murphy-Chutorian and Trivedi (2009), face pose estimation schemes for static images can be classified into the following seven categories.

- **Appearance template methods** compare a new image of a face to a set of exemplars (each labelled with a discrete pose) in order to find the most similar view.
- **Detector array methods** train a series of face detectors each attuned to a specific pose and assign a discrete pose to the detector with the greatest support.
- **Nonlinear regression methods** use nonlinear regression tools to develop a functional mapping from the image or feature data to a face pose measurement.
- **Manifold embedding methods** seek low-dimensional manifolds that model the continuous variation in head pose. New images can be embedded into these manifolds and then used for embedded template matching or regression.
- **Flexible models** fit a non-rigid model to the facial structure of each individual in the image plane. Face pose is estimated from feature-level comparisons or from the instantiation of the model parameters.

- **Geometric methods** use the location of features such as the eyes, mouth, and nose tip to determine pose from their relative configuration.
- **Hybrid methods** combine one or more of these aforementioned methods to overcome the limitations inherent in any single approach.

Each category is described in detail in the following sections. Since the geometric methods use facial feature extraction as the first step for pose estimation, a brief review of facial features localization schemes is also presented in section 2.7.

2.2 APPEARANCE TEMPLATE METHODS

In appearance template methods, an input image is compared with a set of templates (each labelled with a discrete pose). The template that gives the maximum similarity is assigned as pose to the input image. Generally, appearance template methods use image-based comparison metrics to match the pose of an input image with templates. Use of normalized cross-correlation at multiple image resolutions (Beymer 1993) and mean squared error (MSE) over a sliding window (Niyogi, Freeman 1996) are examples of appearance template methods. Figure 2-1 shows an illustration of appearance template methods.

Appearance template methods have some advantages over other methods. The system can adapt to changing conditions by adding more templates to the set of templates. Also, appearance template methods do not require any negative training examples or facial feature points. Creating a corpus of training data requires only cropping face images and providing face pose annotations. These methods are also suitable for both low and high resolution images.

The appearance template methods also have disadvantages. Since the templates can represent discrete poses only, these methods are applicable to discrete pose and can not be used for continuous pose estimation. Also, these methods assume that face boundary has already been localized. Face localization error can degrade the accuracy of these systems. Moreover, adding more templates means more comparisons which increases the computational expense of these methods. One solution to these problems is proposed by Ng and Gong (1999; 2002), which suggest training a set of Support Vector Machines (SVMs) to localize the face and subsequently use the support vectors as appearance templates for pose estimation.

Apart from the above stated limitations, the most significant problem with these methods is that they work on the principle that pair-wise similarity in image space can be equated to similarity in pose. If two images of the same person with different poses and two images of different persons with same pose are considered, the first one will probably give a more metrics because of similarity in identity. This means that the effect of identity can cause problem to appearance template methods. Although the effect is likely to be lessened for wide varying poses, it can still lead to erroneous pose estimation when the poses are not varied widely. Many approaches have been suggested to overcome this pair-wise similarity problem which are based on various distance metrics and image transformations that reduces face pose estimation error by reducing the identity effect. For example, the image is convolved with a Laplacian-of-Gaussian filter (Gonzalez, Woods 2002) to emphasize more common pose related contours while removing the identity specific texture variations. Similarly, the images can be convolved with complex Gabor-wavelet to emphasize only the directed features such as vertical lines of nose and horizontal lines of mouth (Sherrah, Gong and Ong (1999; 2001)). The magnitude of this convolution is relatively invariant to shift which reduces the error generated due to variance in facial features locations between different subjects. Lablack et al. (2008) present a template based face pose estimator using support vector decomposition and Gabor wavelets features. Features vectors are first extracted using both Support vector decomposition and Gabor wavelets. The pose is then estimated using support vector machine (SVM) and K nearest neighbours (KNN) algorithms. The comparison of the two learning algorithms (SVM and KNN) is also presented.

Table 2-1 presents a summary of appearance template methods, highlighting the approach, advantages, disadvantages and representative works.

Table 2-1 Appearance Template Methods Summary

Approach	Compare a new image with a set of templates in order to find the most similar view
Advantages	• Templates can be expanded to a large set, allowing systems to adapt to changing conditions • Do not require negative training examples or facial feature points • Suited for both low and high resolution images
Disadvantages	• Capable of estimating discrete poses only • Work on the assumption that face boundary has already been localized. Face boundary localization error degrades the accuracy of these systems • Adding more templates makes the method computationally expensive • Work on the principle that pair-wise similarity in image space can be equated to similarity in pose, so effect of identity can harm the accuracy
Representative Works	• Mean Squared Error (Niyogi, Freeman 1996) • Normalized Cross-correlation (Beymer 1993) • Gabor Wavelets (Sherrah, Gong & Ong 2001)

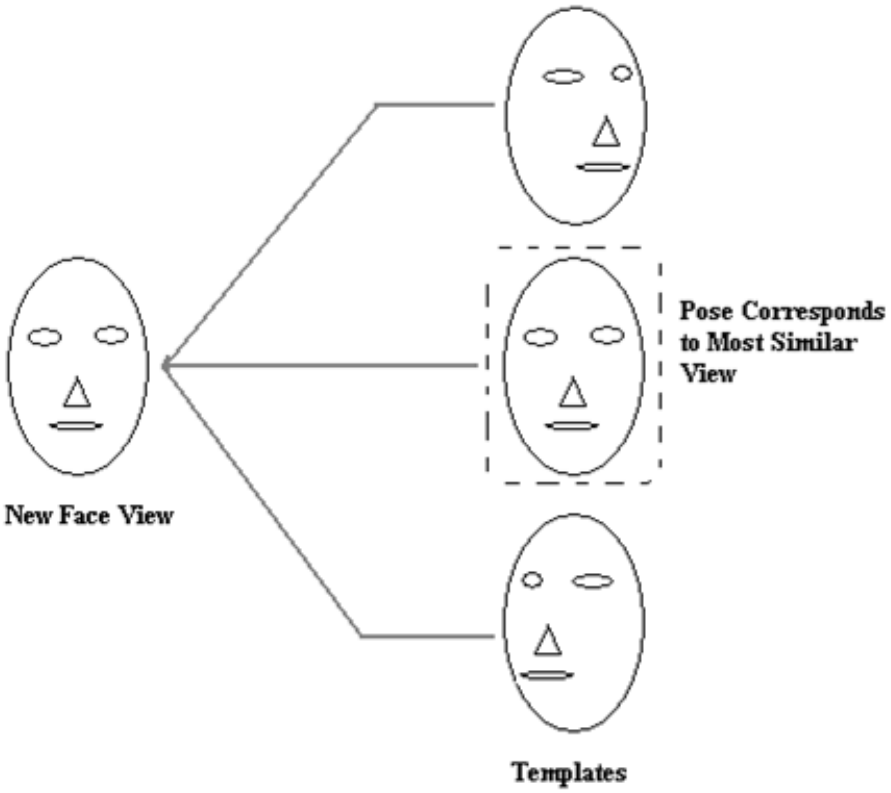


Figure 2-1 Appearance Template Methods

2.3 DETECTOR ARRAY METHODS

Face detection has been the focus of research in the recent past in computer vision. Many face detectors have been proposed for frontal faces as well as side views. Detector array methods utilize these face detectors for different poses to estimate pose. In detector array methods, multiple face detectors (each for a discrete pose) are trained. All these detectors are applied to the input image and the pose of the detector that has the maximum success in detecting the face is estimated as pose of the input image. For continuous sequence, pose is estimated by the detector with greatest support. Detectors array methods and appearance template both operate on image patch. However, in detector array methods, instead of comparing an image to a set of large templates, face detectors trained for different poses are used. Examples of this type of methods are: the use of three support vector machines (SVMs) for estimating three discrete yaw angles (Huang, Shao & Wechsler 1998) and the use of five trained FloatBoost classifiers (Zhang et al. 2007). Schematically these methods are similar to appearance template methods but processing is different as explained later.

Unlike appearance template methods, detector arrays do not require any separate face detection and localization step since each detector is capable of differentiating between face and non-face regions. Another advantage of these methods is that the face detectors can be trained to ignore appearance variations that do not correspond to pose change and thus avoiding the errors occurred due to the identity effect. Like appearance template methods, detector arrays are suited for both low and high resolution images.

Detector array methods also have disadvantages. It is burdensome and computationally expensive to train a large number of face detectors, and it is difficult to train a detector both as face detector and pose estimator because it requires to be trained for negative non-face examples as well, which needs a large set of negative training data. Similarly, increase in the number of detectors can be problematic. If two detectors are used for very similar poses, the images that are positive examples for one must be negative examples for the other so it is not always possible to train such detectors perfectly. That is the reason that detector array methods are limited to a single degree of freedom and 12 detectors in practice. Since the majority of face detectors have binary outputs, there is no reliable way to use them for continuous pose estimation which limits the use of these methods for coarse pose estimate only. The binary nature of the detectors results in ambiguities when multiple detectors simultaneously classify a positive image. Furthermore, the computational expense increases linearly with the number of detectors, making it difficult to implement real-time systems. Rowley, Baluja & Kanade (1998) suggested a solution to this problem by using a router classifier which selects a single subsequent detector for pose estimation. According to their proposed scheme, the router effectively determines the pose and then the subsequent detector verifies it. This technique sounds good in theory; however, it has only been demonstrated for rotation in the camera-plane, first using a neural network-based face detector (Rowley, Baluja & Kanade 1998), and later with cascaded AdaBoost detectors (Viola, Jones 2004).

Table 2-2 summarizes detector array methods, highlighting the approach, advantages, disadvantages and representative works.

Table 2-2 Detector Array Methods Summary

Approach	Train a series of face detectors (each for a discrete pose), apply these detectors to the input image and assign discrete pose to the detector with greatest support.
Advantages	• Separate face detection and localization step is not required • The face detectors used in the method can be trained to learn to ignore the appearance variations that do not correspond to pose change. • Suited for both low and high resolution images
Disadvantages	• It is burdensome and computationally expensive to train many detectors • It is difficult to train a detector both as face detector and pose estimator because it requires to be trained for negative non-face examples as well which needs a large set of negative training data • The binary nature of the detectors results in ambiguities when multiple detectors simultaneously classify a positive image • Computational expense increases linearly with the number of detectors
Representative Works	• The use of three support vector machines (SVMs) for estimating three discrete yaw angles (Huang, Shao & Wechsler 1998) • Trained five FloatBoost classifiers (Zhang et al. 2007)

2.4 NONLINEAR REGRESSION METHODS

In nonlinear regression methods, the image is mapped to one or more pose directions using a non-linear function. An illustration of nonlinear regression methods is presented in Figure 2-2. In nonlinear regression methods, a model is built with a set of labelled training data for discrete or continuous pose estimation. The problem with these methods is that it is hard to achieve an ideal mapping using regression. The high dimensional data of an image is a challenge for regression methods. Different solutions have been suggested. First some dimensionality reduction scheme is used and then nonlinear regression is applied. Li, Gong & Liddell (2000) and Li et al. (2004) use Support Vector Regressors (SVRs) after the dimensionality reduction by Principal Component Analysis while Murphy-Chutorian, Doshi & Trivedi (2007) use localized gradient orientation histogram for dimensionality reduction. The regression tools can be applied to relatively low dimensional feature data if the location of facial features is known in advance (Ma et al. 2006, Moon, Miller 2004).

Neural Networks have been widely used as a regression tool. A multi-layer perceptron (MLP), trained with back-propagation algorithm (Bishop 2005, Duda, Hart & Stork), is one of the most common type of neural network. A MLP consists of many feed-forward cells defined in multi-layers. The output of each layer serves as input to the next layer as shown in Figure 2-3. A MLP can be trained using different training algorithms. Back-propagation is a supervised training algorithm for MLP that propagates the error back through each of the hidden layers to adjust the weights and biases of these layers. Once the training is done, it can be used to classify new input data. In the case of pose estimation, the cropped face image serves as input to a MLP and each of the output cells corresponds to a discrete pose. A Gaussian kernel can be used to smooth the training poses to account for the similarity of nearby poses (Brown, Tian 2002, Schiele, Waibel 1995, Zhao, Pingali & Carlbom 2002). A MLP can be trained for fine face pose estimation. In this configuration, each of the output cells corresponds to a single degree of freedom and the activation of the output is proportional to its corresponding orientation (Brown, Tian 2002, Stiefelhagen, Yang & Waibel 2002, Stiefelhagen 2004, Voit, Nickel & Stiefelhagen 2007, Voit, Nickel & Stiefelhagen 2008). Another approach is to train a separate MLP with one output for each degree of freedom. This approach has been used by Voit, Nickel & Stiefelhagen (2006; 2007; 2008) and Tian et al. (2003). These approaches work for multiple far-field cameras in indoor environments. The facial region is detected by either a

background subtraction or colour filter, and Baysian filtering is used to fuse and smooth the estimate of each of the individual cameras.

A locally-linear map (LLM) is another type of neural network which can be used for pose estimation. An LLM consists of many linear maps and the input data is compared with the centroid of each map and used to learn a weight matrix. Face pose estimation consists of nearest neighbour search followed by linear regression with the corresponding map. This approach has been extended with the difference vectors, and dimensionality reduction (Bruske et al. 1998) and decomposition with Gabor-wavelets (Krueger, Sommer 2002). Similar to SVR, a neural network can be trained using the data from facial feature locations. Gourier, Hall & Crowley (2004a) and Gourier et al. (2007) use this approach with associative neural network.

Nonlinear regression pose estimation methods have several advantages. These systems are very fast and require only labelled cropped face images for training. Unlike detector array methods, these methods do not require any negative non-face data for training. These method works well for both near and far field images.

The main disadvantage of these methods is that they are prone to face localization error just like appearance template methods. Osadchy, Le Cun & Miller (2007) present a solution to this problem by using a convolutional network (Le Cun et al. 1998) that extends the classical MLP by modelling some shift, scale and distortion invariance.

Table 2-3 summarizes the nonlinear regression methods, highlighting the approach, advantages, disadvantages and representative works from the literature.

Table 2-3 Nonlinear Regression Methods Summary

Approach	The data is mapped from image space to one or more pose directions using nonlinear function mapping. The high dimensions of the image data is generally reduced first using some dimensionality reduction technique such as principal component analysis (PCA).
Advantages	• They are computationally less expensive and hence fast in processing • Unlike Detector Array Methods, no negative non-face data is required for training and only cropped labelled face images are enough • Work well for both near and far field images • Give some of the most accurate pose estimates in practice
Disadvantages	• The high dimensionality of image data presents a challenge to regression tools • They are prone to errors from poor face localization
Representative Works	• Extension of classical MLP by explicitly modelling shift, scale and distortion invariance (Osadchy, Le Cun & Miller 2007) • Use of associative neural networks using the data from facial features location (Gourier, Hall & Crowley 2004a, Gourier et al. 2007) • Use of Principal Component Analysis (PCA) and Support Vector Regressors (SVRs) (Li, Gong & Liddell 2000, Li et al. 2004)

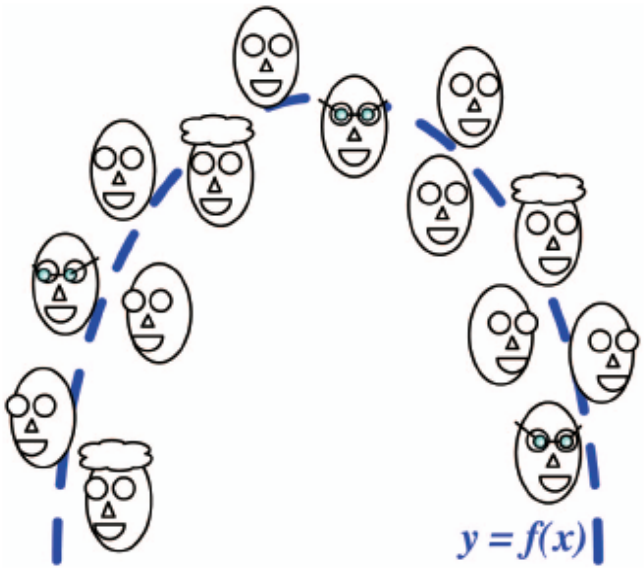


Figure 2-2 Nonlinear Regression Methods adapted from (Murphy-Chutorian, Trivedi 2009)

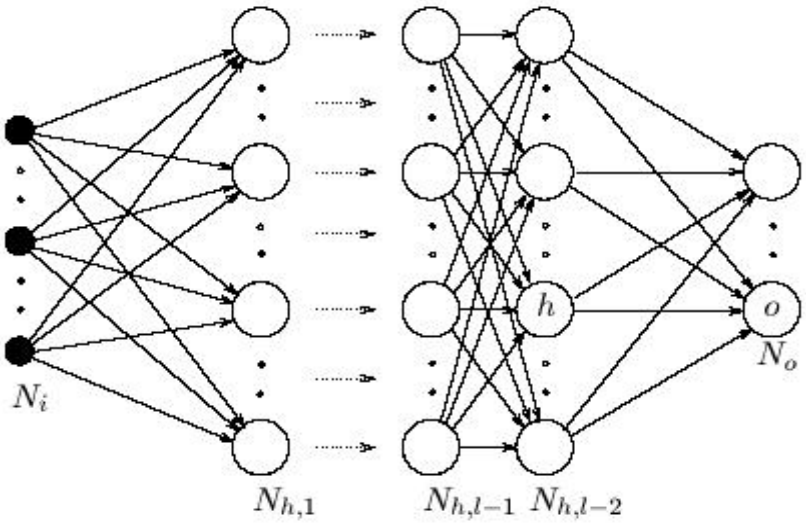


Figure 2-3 Typical Feed forward Multi-layer Perceptron

2.5 MANIFOLD EMBEDDING METHODS

Although a face image is data with a high dimensionality, there are inherently very few dimensions in which the face pose can vary. Therefore, it is possible to consider that each

high dimensional image lies on a low dimensional continuous manifold which can be used to estimate pose. This manifold must be properly modelled and embedding technique is used to project new data on this manifold. This low dimensional embedding can then be used for pose estimation along with the techniques such as regression or template matching in this low dimensional data. Any dimensionality reduction technique can be used as manifold embedding, but the challenge lies in finding a technique that can correctly classify face pose while ignoring other parameters such as illumination and interpersonal variations. The schematic representation of manifold embedding methods is depicted in Figure 2-4.

Two of the most popular dimensionality reduction techniques used for pose estimation are principal component analysis (PCA) and its nonlinear kernelized version KPCA. These techniques estimate the primary mode of variations from a set of data sample (Duda, Hart & Stork). McKenna and Gong (1998) estimate head pose by first projecting the face image into a PCA subspace and then comparing this reduced version with a set of embedded templates. It has been shown that similarity in this low PCA dimensions correlate more with pose similarity than appearance template matching with Gabor wavelet pre-processing (Sherrah, Gong & Ong, 1999, 2001). The use of PCA has some drawbacks, i.e. the linear limitations of standard PCA can not adequately represent the non linear variations in image caused by pose change and the unsupervised nature of this technique does not guarantee that the primary component will relate to pose variation rather than to appearance variation. Srinivasan and Boyer (2002) present a solution to these problems by decoupling the appearance information from the pose by splitting the training data into separate groups that each shares the same discrete pose. PCA and KPCA are then applied to create a separate projection matrix for each of these groups. These pose-eigenspaces represent the primary modes of appearance variations and provide decomposition that is independent of the pose variation for each group. The pose of the input image is then estimated by normalizing and projecting it into each of the pose-eigenspaces and then selecting pose-eigenspace with the highest projection energy. Wang et al. (2008) present a face pose estimator which is robust against large pose variation and uneven illumination conditions. They use non-linear dimension reduction techniques combined with a learned distance metric transformation. The learned distance metric provides better intra-class clustering and hence preserves a smooth low dimensional manifold in the presence of large variation in illumination conditions.

Another approach is to use the embedded sample as input to some classifier such as support vector machines (SVMs) (Li et al. 2001). Ma et al. (2006) show that improved head pose estimation can be achieved by using local Gabor binary patterns with multi-class SVMs while skipping KPCA. Similar to detector array methods, pose-eigen spaces can be used only for coarse pose estimation because the estimate is derived from discrete set of measurements. In case of coarse pose estimation, it is better to use linear dicriminent analysis (LDA) or kenalized discrimenent analysis (KLDA) because these techniques can be used to find the modes of variation that account for face pose (Chen et al., 2003; Wu and Trivedi 2008). Hu and Huang (2008) present a face estimator using different subspace learning including Principal Component Analysis (PCA), Linear Discriminent Analysis (LDA), Locality Preserving projection (LPP), and Pose Specific Subspace (PSS). The face region is first detected and cropped out from input image. The pose is then estimated using these subspace learning techniques. The Pose Specific Subspace performs better than the other three subspace techniques.

Other manifold embedding techniques include Isometric Feature Mapping (Isomap) (Raytchev, Yoda & Sakaue 2004, Tenenbaum, Silva & Langford 2000), Locally Linear embeeding (LLE) (Roweis, Saul 2000), and Laplacian Eigenmaps (LE) (Belkin, Niyogi 2003). In all these methods, the input data must be embedded into an existing manifold with an approximate technique, such as a Generalized Regression Neural Network (Balasubramanian, Ye & Panchanathan 2007).

Manifold embedding techniques have some weaknesses. Except LDA and KLDA, these techniques operate in an unsupervised fashion, ignoring the pose information that might be available for training data. As a result there is a high probability that the manifold will relate to appearance, identity and pose (Balasubramanian, Ye & Panchanathan 2007). As already discussed, the appearance problem can be solved by dividing the training data into groups each share the same discrete pose (Srinivasan, Boyer 2002). The identity problem can be solved in a similar way by creating separate manifold for each subject that can be aligned together. Another difficulty is the heterogeneity of the training data. To model identity, multiple subjects are needed during the training process and it is quite difficult to obtain regular samples of poses from different individual.

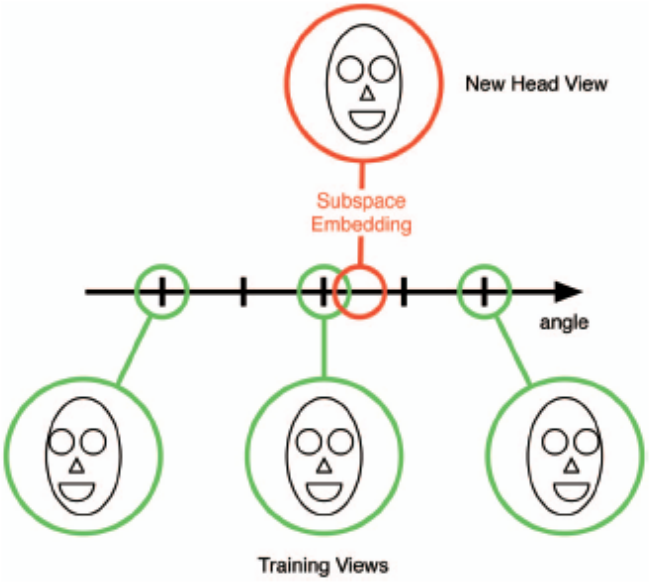


Figure 2-4 Manifold Embedding Methods, adapted from (Murphy-Chutorian, Trivedi 2009)

Table 2-4 summarizes manifold embedded methods, highlighting the approach, advantages, disadvantages and representative works from the literature.

Table 2-4 Manifold Embedding Techniques Summary

Approach	A low dimensional manifold is modelled and embedding technique is used to project new data into this manifold. This low dimensional embedding is then used to estimate face pose with techniques such as regression or template matching.
Advantages	• The manifold can be modelled to be invariant to appearance variation • No facial feature extraction step is required

	• The linear embedding techniques have the advantage of performing embedding by matrix multiplication
Disadvantages	• The linear limitations of PCA can not adequately represent the non-linear variation caused by pose change • Most of these techniques are unsupervised and do not incorporate the pose information that are available during the training process • Due to the unsupervised nature of these methods, there is tendency to build a manifold that corresponds to appearance and identity along with pose.
Representative Works	• PCA and template matching (McKenna, Gong 1998) • PCA and Splitting data into different groups each share the same discrete pose to cope with appearance problem (Srinivasan, Boyer 2002) • PCA and multi-class SVMs (Li et al. 2001) • LDA and KLDA (Chen et al. 2003, Wu, Trivedi 2008)

2.6 FLEXIBLE MODELS

In this approach, a flexible (non-rigid) model is fitted to a facial image such that it conforms to facial features. The face pose is then estimated from different properties of this deformed graph or model. Unlike appearance template methods where the rectangular image is overlaid on template to find similarity, a flexible model is deformed in such a way that it conforms to the individual facial features in flexible model pose estimation schemes. In case of appearance template methods, even with perfect registration there is a high probability that images of two different persons will not overlap completely due to the variations of facial features between different subjects. Flexible models on the other hand, do not suffer from this problem because instead of matching a rectangular region of image,

a flexible model is fitted onto individual facial features such as eyes corners, mouth, nostrils, eyebrows etc. In this case even if the individual features do not match exactly due to inter-subject variations, the shape of the deformed model is more similar for different individuals. The schematic representation of flexible models is depicted in Figure 2-5.

This type of methods requires a large set of training data which should contain not only pose labels but also annotated facial features. To train these systems, facial feature locations are manually labelled in the training data. Sometimes, along with facial feature locations, local features such as Gabor-jets can also be extracted at each location to be used for training. These features are extracted for different views, and invariance to inter-subject variations is achieved by training for a large set of data of different individuals. The information of these features is stored in a bunch of descriptors at each node. This representation is known as Elastic bunch graph (Lades et al. 1993). To compare an elastic bunch graph with new images, the graph is place over the image and is deformed in such a way that each node overlaps with its relevant feature point location. This process is called elastic graph matching (EGM). For face pose estimation, elastic bunch graph is created for each discrete pose image and the new image is matched with all these bunch graphs. The graph which gives maximum similarity is assigned as estimated pose for the input face image. Since the shape of a bunch graph does not vary much with inter-subject variations, these schemes are relatively invariant to inter-subject feature variations. To achieve fine pose estimation, many graphs should be created for different discrete poses. Unfortunately, comparing many bunch graphs, with elastic deformation, is computationally very expensive. Therefore, these methods are more suitable for coarse rather than fine pose estimation.

Another popular flexible model for pose estimation is Active Appearance Model (AAM) (Cootes, Edwards & Taylor 2001). AAM learns primary mode of variation in facial shape and texture from a 2D perspective. Consider N facial points such as eyes, mouth corners, nostrils, eyebrows etc, each having 2D coordinate in an image. These points can be arranged in a vector of length 2N based on the facial feature location. If these vectors are calculated for many faces including different individuals and different poses, they can be used for pose estimation by finding variation in facial shape. Using a dimensionality reduction scheme such as PCA on this data will result in an active shape model (ASM) (Cootes et al. 1995), which is capable of representing shape variation. By looking at the largest principal component obtained from this data, one can find the directions in the data

that correspond to pitch and yaw variations (Lanitis et al. 1995, Lanitis, Taylor & Cootes 1997). If locations of facial features are known, this data can be transformed to an active shape model and the pose can be estimated by evaluating the components responsible for a pose. The ASM can be augmented with the texture information and performing an iterative search to fit the new image to the model. Early work in this area includes extraction of the greyscale profile at each feature point and using a greedy search algorithm to fit the feature points in the model (Lanitis et al. 1995, Lanitis, Taylor & Cootes 1997). Later, a joint shape and texture AAM were introduced by Cootes, Edwards & Taylor (2001). To build a joint shape and texture AAM, first an ASM must be generated from training data. Next the face images are wrapped such that the feature points match those of the mean shape. The wrapped images are normalized and a shape-free texture model is generated. Finally the correlation between shape and texture is learned and used to generate a combined shape and texture model (Edwards et al. 1998). To estimate the pose of a new image, the combined shape and texture model is fitted to the image by iteratively rendering the model to observed image and adjusting the model parameters so that to minimize the distance between these two images. Once the model is converged onto the facial feature, face pose estimation is achieved by transforming the model parameters to a pose estimate. For example, Cootes et al. (2002) use linear regression to transform model parameters to pose estimate for yaw estimation. Morency et al. (2009) present a face pose estimator using a probabilistic framework called generalized adaptive view-based appearance model (GAVAM). The two main components of GAVAM are the view-based appearance model which is acquired and adapted over time, and a series of change-pose measurements.

AAMs have the advantage of being invariant to face localization error because they adapt to the image and find exact locations of feature points. The main limitation of AAMs is that facial features are required to be located first and facial feature localization error can degrade its accuracy. They are also limited to head orientation for which the outer corners of eyes are visible.

Table 2-5 summarizes flexible models for pose estimation, highlighting the approach, advantages, disadvantages and representative works from the literature.

Table 2-5 Flexible Models Summary

Approach	A non rigid model is transformed and fitted to the input image in such a way that it conforms to individual feature points. Face pose is then estimated either from model parameters or feature-level comparison.
Advantages	• They are invariant to face localization error because they adapt to the image and find exact locations of feature points • They are relatively invariant to inter-subject variation since the parameters of a model do not vary much with inter-subject variation
Disadvantages	• Facial feature annotation is required along with pose label for during training • Facial features need to be located first and the feature localization error can degrade the accuracy • They are limited to head orientation for which the outer corners of eyes are visible. • To achieve fine pose estimation, many graphs each for a discrete pose should be created. Unfortunately, comparing many bunch graphs, with elastic deformation, is computationally very expensive and that is the reason that these methods are more suitable for coarse pose rather than fine pose estimation
Representative Works	• Elastic Graph Matching (Krüger, Pötzsch & Von der Malsburg 1997) • Active Shape Model (Lanitis, Taylor & Cootes 1997) • Active Appearance Model (Cootes et al. 2002)

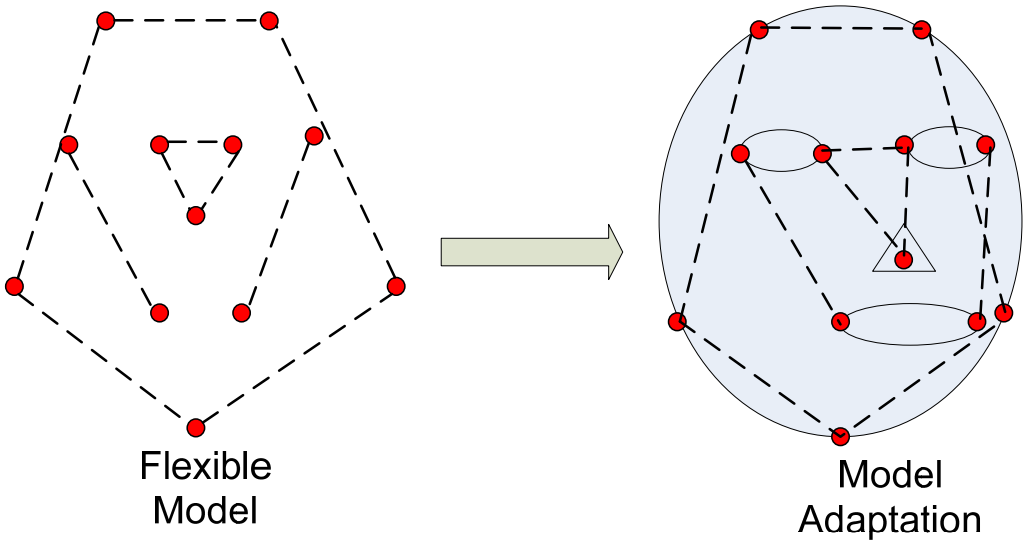


Figure 2-5 Flexible Model Adaptation

2.7 GEOMETRIC METHODS AND FACIAL FEATURES LOCALIZATION

The following section gives a review of geometric pose estimation methods followed by facial feature localization.

2.7.1 GEOMETRIC METHODS

Psychologically, it is believed that human perception about face pose is greatly influenced by the cues such as head-boundary, nose length, facial symmetry, facial features such as eyes and mouth location and orientation. Geometric methods use these cues to estimate face pose. These methods are relatively simple in nature and less expensive computationally but almost all of them require the precise locations of the facial features to be known in advance. Schematic representation of geometric methods is depicted in Figure 2-6.

Different geometric methods exploit the feature point locations and orientation in different ways to estimate face pose. Gee and Cipolla (1994) use five feature points (the outside corners of eyes, outside corners of mouth and tip of nose) to estimate pose. The face symmetry line is computed by joining the midpoint of two lines: one drawn between two eyes corners and the other drawn between mouth corners. Assuming a fixed ratio between these facial points and a fixed length of the nose, the facial direction is determined from 3D angle of the nose. Alternatively, face pose can be estimated from the normal to the plane which contains all these feature points, which can be found from planar skew-

symmetry and a coarse estimate of the nose position. Horprasert, Yacoob & Davis (1996) use five different feature points (the inner and outer corners of each eye and the tip of the nose) to estimate face pose. They assume that all these feature points are co-planner and yaw angle is estimated from difference in length of the two eyes. The line joining the two eyes are used to estimate roll and pitch is calculated by comparing the distance between the nose tip and the line joining two eyes to an anthropometric model. Yu et al. (2009) present a face pose estimator based on a face normal vector which incorporates both 2-D and 3-D face information. Face features such as eyes, nose, lips etc are obtained from 2-D from which the face normal vector is derived in 3-D space. The face pose is then estimated from this face normal vector. The process is iterated until a given accuracy of estimation is satisfied. Ebisawa (2008) presents a face pose estimations schemes based on the locations of pupils and nostrils. Two rings of infrared LEDs are used to obtain dark and bright pupil images. The pupils are then detected by subtracting dark image from the bright image. The nostrils are detected as darker areas in dark or bright image. Two normal vectors (one from two eyes and right nostril plane, and other from two eyes and left nostril plane) are then calculated. The average of these two vectors is estimated as face direction.

Another geometric pose estimation method is presented by Wang and Sung (2007), which uses six feature points i.e. the inner and outer corners of the two eyes and the outer corners of the mouth. This scheme is based on the observation that the line joining the inner eyes corners, the line joining the outer eyes corners and the line joining the mouth corners are parallel for a neutral pose (zero degree angle). Any deviation from parallel is the result of perspective which can be used to estimate pose. The points where these lines intersect (vanishing points) are computed using least square solution to minimize over-determining these three lines. The vanishing points along with the ratio of the line lengths are used to estimate the 3D orientation of these lines which are then used to compute the 3D positions of feature points. The EM algorithm with Gaussian mixture is used to adapt the facial parameters of different individuals to minimize the back-projection error. This method assumes the three lines to be visible which makes this method applicable to only frontal or near frontal images. Fitting an ellipse around the skin colour of a face can lead to a single DOF pose estimation (Cordea et al. 2001). With multiple cameras surrounded the head, yaw can also be estimated as orientation with the most skin colour (Canton-Ferrer, Casas & Pardas 2007, Canton-Ferrer, Casas & Pardas 2008). Martinez et al. (2009) present a face pose estimator for virtual reality applications. They place several markers on the frame of

eye-glasses. They use IR illumination in conjunction with IR reflective markers in order not to distract the user. The face pose is then estimated from the orientation of these markers. The problem with this method is that it is applicable to only their specific virtual reality environment.

Geometric methods have the advantage of being simple and fast. Only a few points are used to estimate pose which makes these methods computationally effective. However, these methods require the precise facial feature locations to be known and hence error in feature point localization can degrade the performance of these methods. Far-field imagery or images with low resolution or bad illumination are problematic in this context. Similarly, images with occlusions such as wearing glasses etc are also not feasible for this kind of methods. These methods are generally more sensitive to occlusions than appearance-based methods.

Table 2-6 summarizes geometric methods for pose estimation, highlighting the approach, advantages, disadvantages and representative works from the literature.

Table 2-6 Geometric Methods Summary

Approach	These methods use the location of facial feature such as eyes corners, mouth corners, nostrils, nose tip etc and their configuration to estimate pose.
Advantages	• These methods are relatively simple • These methods are quite fast. Only few feature points are used to estimate pose which makes these methods computationally very effective
Disadvantages	• These methods assume that feature points are precisely detected in advance. Feature localization error degrades the performance of these methods • These methods are more sensitive to occlusions such as wearing glasses because precise feature localization becomes difficult in this case

	• Far-filed imagery or images with low resolution or bad illumination are problematic for these methods
Representative Works	• Vanishing Points (Wang, Sung 2007) • Ellipse Fitting (Cordea et al. 2001) • Use of five feature points (Gee, Cipolla 1994) • Skin colour region (Canton-Ferrer, Casas & Pardas 2007, Canton-Ferrer, Casas & Pardas 2008)

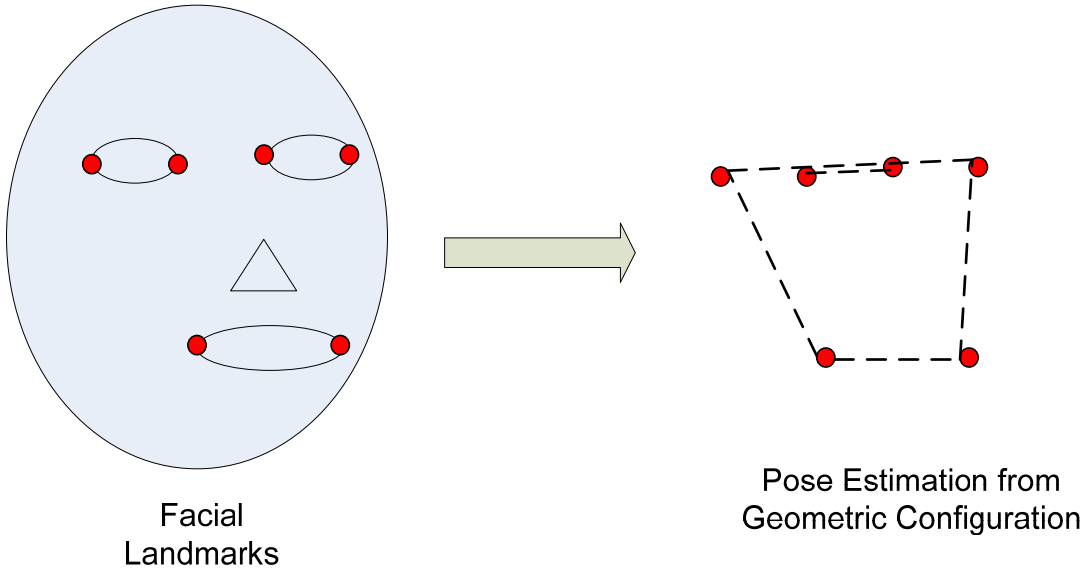


Figure 2-6 Geometric Methods

Since geometric methods depend on the accurate extraction of facial features, a review of facial features (especially eyes and mouth) extraction methods are discussed briefly in the following section.

2.7.2 FACIAL FEATURES LOCALIZATION

Facial feature localization has applications in human-computer interaction, face detection, face recognition, facial expression analysis, pose estimation, surveillance etc. Facial feature detection has been a focus of research in computer vision community for almost

two decades, and numerous papers have been published on this topic. There are still some challenges in this field i.e. feature localization under different illumination and lighting conditions, feature localization under different varying pose, feature localization under occlusions, etc. Facial feature detection techniques can be classified into two major categories. In the first category, it is assumed that the face has already been detected and localized. The second category searches for facial features across the whole image. The schematic representations of these two categories are depicted in Figures 2-7 and 2-8.

The following is a brief review of some of the representative works on facial feature localization from the literature.

Figure 2-7 Facial Feature Localization Category 1 i.e. Face Detection Followed by Feature Localization

Figure 2-8 Facial Feature Localization Category 2 i.e. Feature Localization Followed by Face Detection

Hsu, Abdel-Mottaleb & Jain (2002) present a robust scheme for eyes and mouth detection in colour images. According to their proposed method, the input RGB image is converted first to YCbCr. Two separate eye maps, one from chrominance component of the image and the other from luminance, are built. The eye map from the Chroma is based on the observation that high C_b and low C_r values are present around the eyes. The eye map from luma is based on the observation that eye regions contain both dark and bright pixels in luma. So the morphological operators (e.g. erosion and dilation) can be designed to emphasize brighter and dark pixels in the luma component in the eye regions. The eye map from the luma component is calculated by dividing the dilated version of luma on eroded luma. Their mouth detection algorithm is based on the observation that the mouth region contains stronger red component and weaker blue component as compared to other facial regions, which means that C_r is greater than C_b for the mouth region. A mouth also has relatively low response in C_r/C_b and high response in C_r^2. Bhuiyan et al. (2003) proposed a scheme for six facial features (two eyes, eyebrows, nose and mouth). In their proposed scheme, face is detected using skin colour segmentation in YIQ colour space. The face image obtained is then enhanced through histogram equalization and noise reduction. An optimum threshold value is then found to detect facial features since these features are darker than other parts of the face (forehead, cheeks etc.). Morphological operators are used to fill small holes and remove the unwanted regions. Tags are then assigned to each of the six facial features based on its position in face region. Shih and Chuang (2004) present a facial feature localization scheme to detect eyes, nose, mouth, and chin region. Two thresholds are applied to extract the face and head boundaries. The projections along x- and y-axes are used to locate facial features in the face region. In cases where the above procedure results in very close facial features which can not be distinguished, Gabor filters are used to detect the eyes. The rest of the features are then localized using heuristic rules which are based on face geometry. Perlibakas (2003) presents a facial feature localization scheme that uses morphological operators to find dark regions which correspond to eyes and mouth in the elliptical face region. Further heuristic rules are applied to remove the false positive and refine the detected features. Han et al. (2000) and Chan & Lewis (1999) also use a set of heuristic rules to find the features once the candidate regions are obtained. If these regions conform to the rules, each one of them is labelled as a potential eye or a potential lip. All possible triplets of the candidate eyes and lip regions are then established.

Further refinement is done by applying further rules on these triplets e.g. eyes and mouth triplets that are far from the centre of facial region, are removed. Similarly, triplets where the line joining the two eyes deviates significantly from the minor axis of the facial ellipse are removed.

Gourier, Hall & Crowley (2004b) present a facial feature localization scheme following face detection in colour images. Red and Green components, divided by the intensity are used to generate luminance invariant vector for each pixel. These vectors are then analyzed to find the probability that each pixel belongs to the face or not. The facial image obtained is then normalized in size and intensity to locate facial features. The first and second derivatives of a Gaussian in x and y directions are convolved with facial image that form a 5-D local appearance descriptor for each pixel. These Gaussian derivatives are then clustered using K-means algorithm. The obtained clusters are then exploited to detect facial features. Further geometrical analysis is done to refine the detected regions and remove false positives. Cristinacce, Cootes & Scott (2004) present an algorithm for detecting 17 features around the eyes, nose and chin, using a multi-stage approach. They detect the face first using the Viola-Jone face detector (Viola and Jones, 2004). The same classifier is trained using facial features and thus different detector is constructed for each feature. The pair wise reinforcement of feature responses (PRFR) is used to increase the localization accuracy. In more detail, the distribution of true location feature *i* given the best match of feature detector *j* in the reference coordinate systems defined by the face region is calculated. Relative histograms, which approximate these distributions, are calculated using the true feature locations and detector matches on a set of training images. Active Appearance model (AAM) is used to do further refinement while using four values for each pixel i.e. the normalized gradients in the x and y directions, its "cornerness" and "edgeness". Jesorsky, Kirchberg & Frischholz (2001) present a three-stage approach for eye centre localization. The Hausdorff distance between edges of the image, and an edge model of the face, is used to detect the face. The Hausdorff distance between the edge map of the input image and a more refined model of the areas around the eyes is used for more accurate localization of the upper part of a face in the second stage. In the third stage, a Multi-layer Perceptron (MLP) is used to detect exact pupil location. Zhou& Geng (2004) present a method for eye centre localization using generalized projection functions (GPFs). GPFs are linear combinations of functions which use the mean and variance of intensity along rows and columns of an image. The bounding box of an eye is located using GPFs

and the centre is found by taking the centre of this box. Wang & Sung (1999) present a method for the detection of eyes, nostrils and mouth. The face is detected first, using skin colour segmentation. Face contour is calculated by sequence of morphological operators and ellipse fitting. Eyes and nostrils are then detected by searching for the darkest regions in certain area of the face. Geometrical constraints are applied to search for the best pair of eyes or nostrils. Integral projection and edge operators are then applied to an estimated region to detect mouth.

Ma et al. (2004) propose a three-stage algorithm for eyes localization. Viola-Jones face detector is used to localize the face in the first step. In the second stage, the face region is vertically divided into two (by looking at the vertical projection) and eye detector is applied to each of the half regions to detect the eyes. The threshold value is kept low so that the eye is definitely obtained in the connected regions. The eye detections are sub-sampled according to proximity (by exploiting the face size) and to their probability. This step is necessary to remove false positives. The final stage works in a top-down fashion in which an eye-pair classifier is applied on all possible pairs of the candidate eye regions. The method is declared to work on upright frontal faces with in and out of the plane rotations up to 25 degrees. Tang et al. (2005) present a three-stage technique for eyes localization which uses cascaded Adaboost along with Support Vector machines (SVMs). Adaboost face detector is applied first to extract the face region. In the second stage, an Adaboost eye detector is applied to detect all possible eye candidate regions. Finally, a SVM post classifier is applied to remove non eye regions and extract both eyes. The method is declared to work for 10 degrees head rotation both in and out of plane. (Hamouz et al. 2005) present a four-stage face detection algorithm while extracting facial features first. In the first step, 10 facial points are searched for by applying Gabor filter and comparing the response matrix with that of the already created model for each feature. The next step is face hypothesis generation which considers all possible triplets of the feature points obtained and discards the one with improbable geometrical configurations. The third step uses affine transformations to normalize the remaining candidate points with respect to scale and in-plane rotation. Finally, a verification step is applied which uses two SVMs which work in coarse-to-fine fashion. Everingham & Zisserman (2006) present three approaches to eye localization and compare all the three approaches. The three approaches are: a regression methods which trying to minimize the prediction error, a Baysian approach which builds probabilistic models for eye and non-eye appearance, and a single

classifier which is trained using Adaboost. All the three approaches were applied to a facial region obtained through Viola-Jone face detector (Viola, Jones 2004). Results show that simple Baysian model outperforms the other approaches.

More recently, a facial feature localization scheme has been proposed by Asteriadis, Nikolaidis & Pitas (2009). Their proposed scheme utilizes distance vector fields (DVFs) to locate facial features. The Adaboost face detector Viola & Jones (2004) is used first to locate the face. The distance vector fields of the face are calculated by assigning to every facial image pixel a vector pointing to the closest edge pixel. Figure 2-9 shows how distance vector fields are calculated for a facial image. The x and y components of these vectors are used to detect the coarse locations of the eyes and mouth. The eyes and mouth regions are detected by finding the regions inside the face, their distance vector fields resemble more with that of the DVFs of eyes and mouth templates, which are extracted from a training set of eye and mouth images. Luminance information is used for eye centre localization while hue channel is used for mouth detection. Since DVFs are relatively invariant to colour, illumination and intensity variations, this method is robust and works well under various colour, illumination and intensity conditions.

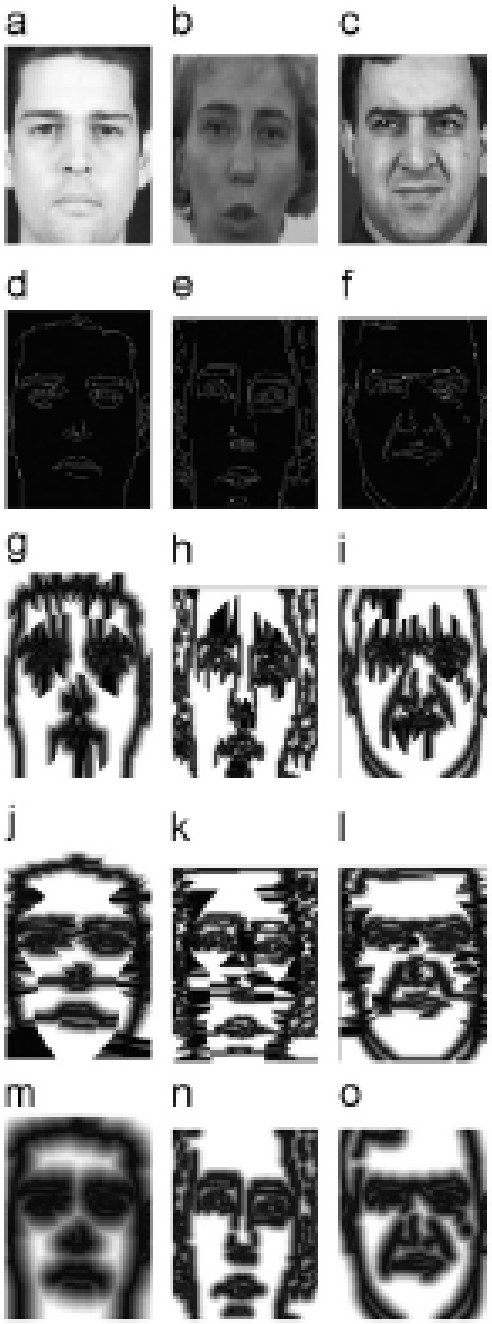


Figure 2-9 Distance Vector Fields, adapted from (Asteriadis, Nikolaidis & Pitas 2009)

2.8 HYBRID METHODS

Hybrid pose estimation methods combine two or more of the aforementioned techniques to estimate face pose. Multiple pose estimation techniques are combined in such a way to overcome the weaknesses of each of the individual methods. Most of the hybrid pose estimation methods combine one of the static pose estimation schemes with pose tracking to achieve pose tracking in videos. The pose is generally estimated in the first frame of a

video and is then tracked using some tracking technique. Many methods have been presented by combining geometric pose estimation methods with point tracking (Heinzmann, Zelinsky 1998, Horprasert, Yacoob & Davis 1997, Hu et al. 2004, Jebara, Pentland 1997, Newman et al. 2000). Zhu & Fujimura (2004) combine PCA template matching with optical flow and Huang & Trivedi (2004) combine PCA template matching with continuous density hidden Markov models. Morency, Rahimi & Darrell (2003) combine PCA embedded template key-frame matching with stereo tracking, and Ba & Odobez (2004) combines texture appearance template with particle filtering.

A few of the hybrid methods combine pose estimation methods for pose estimation in static images without any tracking. These methods gather information from multiple cues which results in increased accuracy. Sherrah & Gong (2001) present a hybrid pose estimation scheme which combines appearance template with geometric cues. Wu & Trivedi (2008) and Wu et al. (2004) proposed a scheme which is a hybrid of manifold embedding and flexible models. In this scheme the Elastic graph matching is used to refine the pose estimated by manifold embedding.

The biggest advantage of hybrid methods is that they overcome the weaknesses of each of the individual methods and thus increase the accuracy of pose estimation. However, the fusion of two or more methods results in increased complexity. Therefore, hybrid methods are usually more expensive computationally and slower than each of the individual methods.

Table 2-7 gives a summary of hybrid methods, highlighting the approach, advantages, disadvantages and representative works. Figure 2-10 shows the schematic representation of hybrid methods.

Table 2-7 Hybird Methods Summary

Approach	These methods combine two or more than two aforementioned pose estimation techniques to overcome the limitations of each of the individual techniques
Advantages	• These methods overcome the limitations of the individual

	schemes which are combined • Usually, these methods are more accurate because they are based on multiple pose cues
Disadvantages	• Since these methods combine two or more than two individual methods, they are more complex • These methods are usually more expensive computationally and slower than each of the individual methods
Representative Works	• Appearance template and geometric cues (Sherrah, Gong 2001) • Manifold embedding and Elastic Graph Matching (Wu, Trivedi 2008, Wu et al. 2004)

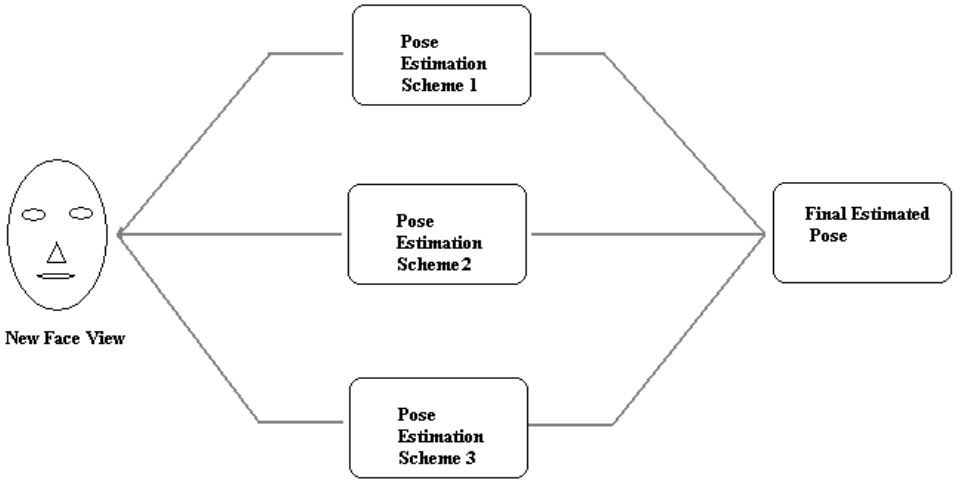


Figure 2-10 Hybrid Methods

2.9 COMPARISON OF PUBLISHED RESULTS

The mean absolute angular error is a common metric for evaluation of fine face pose estimation systems. For coarse pose estimation, classification error (i.e. how often an image at specific discrete pose was classified accurately) is the most common evaluation

metric. Since the proposed face pose estimation schemes (discussed in chapter 5 and 6 in detail) are for coarse face pose estimation, we present a comparison of coarse pose estimation systems only in Table 2-8. The results are presented for standard pose databases which are discussed in detail in appendix A.

Table 2-8 Mean Absolute Error of Coarse Pose Estimation, adapted from (Murphy-Chutorian, Trivedi 2009)

Dataset Publication	Mean Absolute Error Yaw	 Pitch	Classification Accuracy	# of Discrete Poses	Notes
Pointing '04					
(Wu et al. 2004)	-	-	[90%]*	93	-
(Stiefelhagen 2004)	9.5°	9.7°	{52%,66.3%}+	{13,9}+	1
Human Performance (Gourier et al. 2007)	11.8°	9.4°	{40.7%59.0%}+	{13,9}+	2
Associative Memories (Gourier et al. 2007)	10.1°	15.9°	{50.0%,43.9%}+	{13,9}+	3
(Voit, Nickel & Stiefelhagen 2007)	12.3°	12.77°	-	93	-
CHIl-CLEAR06					
(Voit, Nickel & Stiefelhagen 2006)	-	-	39.40%	8	-
(Voit, Nickel & Stiefelhagen 2007)	49.2°	-	34.90%	8	-
(Canton-Ferrer, Casas & Pardas 2007)	73.63°	-	19.67%	8	-
(Zhang et al. 2007)	33.56°	-	[87%]*	5	-
CMU-PIE					
Probabilistic Method (Brown, Tian 2002)	3.6°	-	-	9	-
Neural Network (Brown, Tian	4.6°	-	-	9	-

2002)	-	-	91%	9	3
(Brown, Tian 2002)	-	-	71.20%	9	-
(Ba, Odobez 2004)	-	-	89%	9	6
(Tian et al. 2003)					
Softopia HOIP					
PCA (Raytchev, Yoda & Sakaue 2004)	15.9°	12.4°	-	91	7
LPP (Raytchev, Yoda & Sakaue 2004)	15.9°	12.8°	-	91	7
Isomap (Raytchev, Yoda & Sakaue 2004)	11.5°	11.9°	-	91	8
SVR (Ma et al. 2006)	-	-	81.50%	45	3,4
SBL (Ma et al. 2006)	-	-	81.70%	45	3,4
CVPR-86					
PCA (Wu, Trivedi 2008)	-	-	36.40%	86	-
LDA (Wu, Trivedi 2008)	-	-	40.10%	86	-
KPCA (Wu, Trivedi 2008)	-	-	42%	86	-
KLDA+EGM (Wu, Trivedi 2008)	-	-	75.40%	86	-
CAS-PEAL					
(Ma et al. 2006)	-	-	97.14%	7	3
Other Datasets					
(Niyogi, Freeman 1996)	-	-	48%	15	9
(Krüger, Pötzsch & Von der Malsburg 1997)	-	-	92%	5	10
(Li et al. 2001)	-	-	[96.8%]*	10	-
(Tian et al. 2003)	-	-	84.30%	12	11
Notes * Any neighboring pose considered correct classification as well					

+ DOF classified separately {yaw, pitch}

1. Used 80% of Pointing ’04 images for training and 10% for evaluation
2. Human performance with training
3. best results over different reported methods
4. better results have been obtained with manual localization
6. Best single camera results
7. Results for 100-dim embedding
8. Results for 8-dim embedding
9. Dataset: 15 images for each of 11 subjects
10. Dataset: 413 images of varying people
11. Dataset: far-field images with wide-baseline stereo

2.10 SUMMARY AND CONCLUSIONS

This chapter presented a detailed review for face pose estimation methods from the literature. The face pose estimation schemes for static images were categorized into seven classes: Appearance Template Methods, Detector Array Methods, Non-linear Regression Methods, Manifold Embedding Methods, Flexible Models, Geometric Methods, and Hybrid Methods. Since geometric methods assume that facial features are already localized, a brief review of facial features localization was also presented in section 2.7. Finally, detailed comparison of the published pose estimation methods was presented in section 2.9.

All the pose estimation techniques discussed in this chapter have some shortcomings and limitations in terms of accuracy, applicability to monocular images, being autonomous, identity and lighting variations, image resolution variations, range of face motion, computational expense, presence of facial hairs, presence of accessories like glasses and hats, etc. These shortcomings of existing face pose estimation techniques motivated the research work presented in this thesis. The main focus of this research is to design and develop novel face pose estimation algorithms that improve automatic face pose estimation in terms of processing time, computational expense, and invariance to different conditions as already mentioned.

3 FUNDAMENTAL TECHNIQUES FOR IMAGE PROCESSING

3.1 INTRODUCTION

This chapter discusses some of the fundamental image processing techniques which have been utilized in the proposed algorithms in chapters 4, 5 and 6. Colour spaces, edge detection techniques, morphological operators (i.e. erosion and dilation), distance transform and normalized cross-correlation have been discussed briefly. Sample examples and applications have been provided in each section.

3.2 COLOUR SPACES

Colour spaces are three dimensional arrangements of colour sensations. Each colour is specified by a point in these spaces. The following are some of the popular colour spaces used in image processing.

3.2.1 RGB

RGB is the most widely used colour space in image processing. This colour space is the basic one, and, if necessary, can be transformed to other colour spaces. RGB colour space is based on the principle that red, green and blue being the primary colours and can be mixed in different proportions to form different colours. The main disadvantage of RGB colour space is a high correlation between its components: about 0.78 for B-R, 0.98 for R-G and 0.94 for G-B. This property makes the RGB colour space unsuitable for compression. Also, it is not possible to estimate the perceived distance between two colours on the basis of distance between them in RGB colour space. A variant of RGB is normalized rgb, which is calculated as:

$$r = \frac{R}{R+G+B} \tag{3-1}$$

$$g = \frac{G}{R+G+B} \tag{3-2}$$

$$b = \frac{B}{R+G+B} = 1 - r - g \tag{3-3}$$

Values of rgb space are more stable with changes in illumination level than RGB space. Figure 3-1 shows a RGB colour space cube.

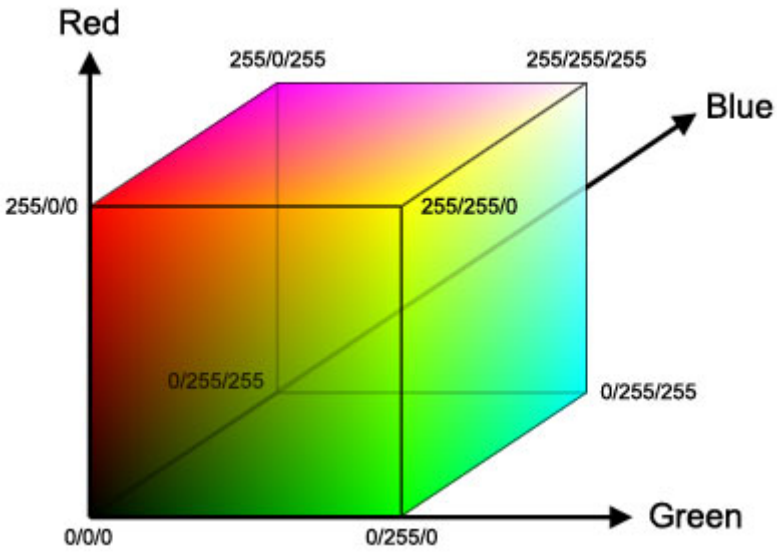


Figure 3-1 RGB Colours Cube

3.2.2 HSV

Similar to RGB, HSV colour space has also three channels: Hue, Saturation and Intensity. The Hue channel represents the "colour", Saturation represents the "Amount" of colour (e.g., there is a different amount of colour in pale green and pure green), and Intensity represents the brightness of the colour (e.g. light green and dark green are the same colours having different intensities). Figure 3-2 shows a HSV colours cone.

3.2.3 YCBCR

YCbCr also has three channels. Y is the luma channel which represents brightness.Cb and Cr are the chroma components that represent blue-difference and red-difference respectively. RGB colours can be converted to YCbCr using the following equations:

$$Y = 0.299R + 0.587G + 0.114B \quad \textbf{3-4}$$

$$Cb = -0.169R - 0.331G + 0.500B \quad \textbf{3-5}$$

$$Cr = 0.500R - 0.418G - 0.081B \quad \textbf{3-6}$$

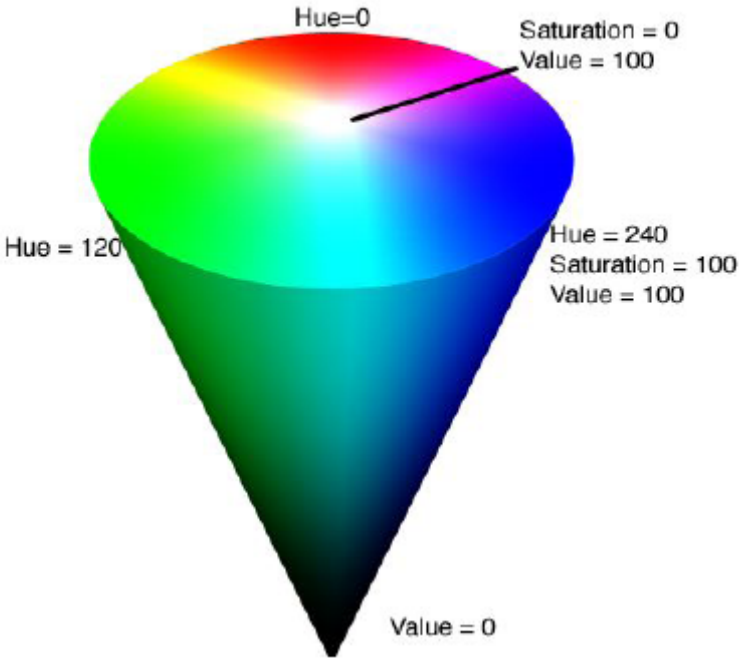


Figure 3-2 HSV Colour Space Cone

3.3 EDGE DETECTION

Edge detection refers to the process of identifying sharp discontinuities in an image. These continuities are abrupt changes in the pixel intensity values which usually characterize the boundaries of objects in an image. The detection of edges in an image is done by convolving the image with an edge operator (a 2-D filter, also known as a "mask"). There are several edge operators available and the selection of an operator depends on the following factors:

- **Edge Orientation:** The geometry of an operator determines the direction in which the operator is most sensitive to edges. Operators are generally optimized to look for edges in horizontal, vertical or diagonal directions.
- **Noise:** Some edge operators are more robust to image noise while others are very sensitive.
- **Edge Structure:** Not all edges involve a step change in intensity. Some edges involve a gradual change instead of an abrupt change in pixels intensity values.

The edge detection schemes can be classified into two major classes: Gradient and Laplacian. The gradient methods detect edges by looking for maximum and minimum in the first derivative of an image while the laplacian methods search for zero crossings in the second derivative of an image to detect edges. Some of the most commonly used edge operators are described below.

3.3.1 SOBEL OPERATOR

Sobel operator comprises two 3*3 masks: one for horizontal edges and one for vertical edges (see Figure 3-3). These two masks can be applied separately to produce separate measurement of the gradient component in each direction. These two can also be combined to get the absolute magnitude and direction of the gradient. The magnitude of the gradient is given by:

$$|G| = \sqrt{Gx^2 + Gy^2} \qquad \text{3-7}$$

Orientation of the edge can be calculated as:

$$\theta = \arctan(Gy / Gx) \qquad \text{3-8}$$

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figure 3-3 Sobel Masks

3.3.2 ROBERT'S CROSS OPERATOR

Robert's cross operator consists of two 2*2 convolution masks (see Figure 3-4). These two operators are applied separately to an image and the combined gradient magnitude can be calculated as:

$$|G| = \sqrt{Gx^2 + Gy^2} \quad \text{3-9}$$

The angle of orientation of an edge giving rise to the gradient can be calculated as:

$$\theta = \arctan(Gy/Gx) - 2\pi/4 \quad \text{3-10}$$

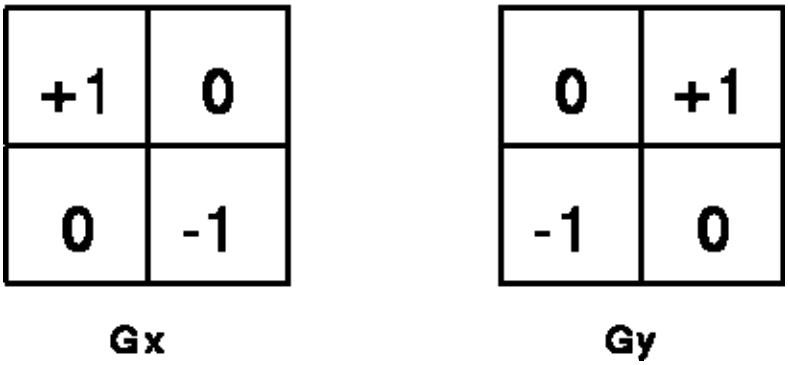


Figure 3-4 Robert's Cross Masks

3.3.3 PREWITT'S OPERATOR

Prewitt's operator is quite similar to Sobel's operator with the convolution masks shown in Figure 3-5.

$$h1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \qquad h3 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

Figure 3-5 Prewitt's Masks

3.3.4 LAPLACIAN OF GAUSSIAN

Laplacian is a 2-D isotropic measure of the second derivative of an image. Laplacian highlights the regions of an image with rapid intensity change and hence it is useful for edge detection. The Laplacian *L(x, y)* of an image with pixel intensity values *I(x, y)* is given by:

$$L(x,y)=\frac{\partial^2 I}{\partial x^2}+\frac{\partial^2 I}{\partial y^2} \qquad \text{3-11}$$

Laplacian is very sensitive to noise. Therefore, it is usually applied to an image that has already been smoothed with a Gaussian filter. Three commonly used laplacian masks for edge detection are shown in Figure 3-6.

0	−1	0
−1	4	−1
0	−1	0

−1	−1	−1
−1	8	−1
−1	−1	−1

Figure 3-6 Laplacian Masks

3.3.5 CANNY EDGE DETECTOR

Canny edge detector is also known as optimal edge detector. In the canny edge detection scheme, an image is first smoothed to reduce noise. The image gradient is then calculated to find the regions with high spatial derivatives. These high spatial derivative regions are then tracked and any of the pixels that is not at maximum is set as non-edge pixel. Two threshold values are then used for processing the remaining high spatial derivative regions. The pixels in these regions, which are above the high threshold value, are set to one (edge) and the pixels whose intensity values are less than the low threshold value are set to zero (non-edge). For each of the remaining pixels, a pixel which is between the two threshold values is set to zero unless there is path from this pixel to a pixel with a gradient above the high threshold value.

Figure 3-7 shows the edges detected using the same image by the Sobel, Robert, Prewitt, Laplacian and Canny operators. Matlab edge detection function, with default parameters, was used to detect the edges.

Original Image

Sobel

Robert's

Prewitt

Laplacian

Canny

Figure 3-7 Edge Detection Example

The performance of edge edector is highly dependent on the image quality and relative performance of the edge detectors vary significantly across the images. An edge detector might perform better for one image and worse for a second image. So the selection of an edge detector is based on the image quality and nature of application. The classical operators such as Sobel, Robert's cross and Prewitt use first derivatives to detect edges and their orientations. The advantage of these operators is that their calculation is very simple but these operators are very sensitive to noise in the image. The Laplacian of Gaussian (LOG) operator uses second derivative of the image to detect edges. The disadvantage of Laplacian of Gausian is that it can not find the orientation of edge because of using the laplacian filter. The Canny edge detector which is also known as optimal edge detector, uses probability for finding error rate and localization. It reduces the noise by smoothing the image and hence it performs better in noisy conditions. The disadvantage of Canny edge detector is that it has relatively a complex computation. In our work, we use Canny edge detector with a threshold value for the computation of distance transform. We need the edges of eyes and mouth which are sharp enough to be detected with a Canny edge detector with a threshold value.

3.4 MORPHOLOGICAL OPERATORS

Morphological Dilation and Erosion are the most important and primitive morphological operators which are widely used in image processing. The following is a brief description of morphological dilation and erosion.

3.4.1 MORPHOLOGICAL DILATION

Let A and B be two sets in Z^2, the dilation of A by B is defined as:

$$A \oplus B = \left\{ z \middle| (\hat{B})_z \cap A \neq \emptyset \right\} \qquad \text{3-12}$$

Where $A \oplus B$ is the dilation of A by B. B is generally known as the structuring element. The above equation is based on getting the reflection of B about its origin and shifting this reflection by z. The set of all displacements z such that A and $\hat{B}$ overlaps by at least one element is called dilation.

In the case of image processing, the structuring element is a matrix of 1s and 0s which is usually much smaller than the size of the image being processed. The centre pixel of the structuring element identifies the pixel of interest (pixel being processed). The pixels

containing 1s define the neighbourhood of structuring element. In the case of dilation, the structuring element is shifted through all pixels of the image and the value of the output pixel is the maximum value of all the pixels in the input pixel's neighbourhood. In the case of binary image, if any of the pixels in the neighbourhood is 1 then the output is set to 1. Figures 3-8 and 3-9 show the schematic representation of Grey level and binary dilation. Figure 3-10 shows an example of binary dilation.

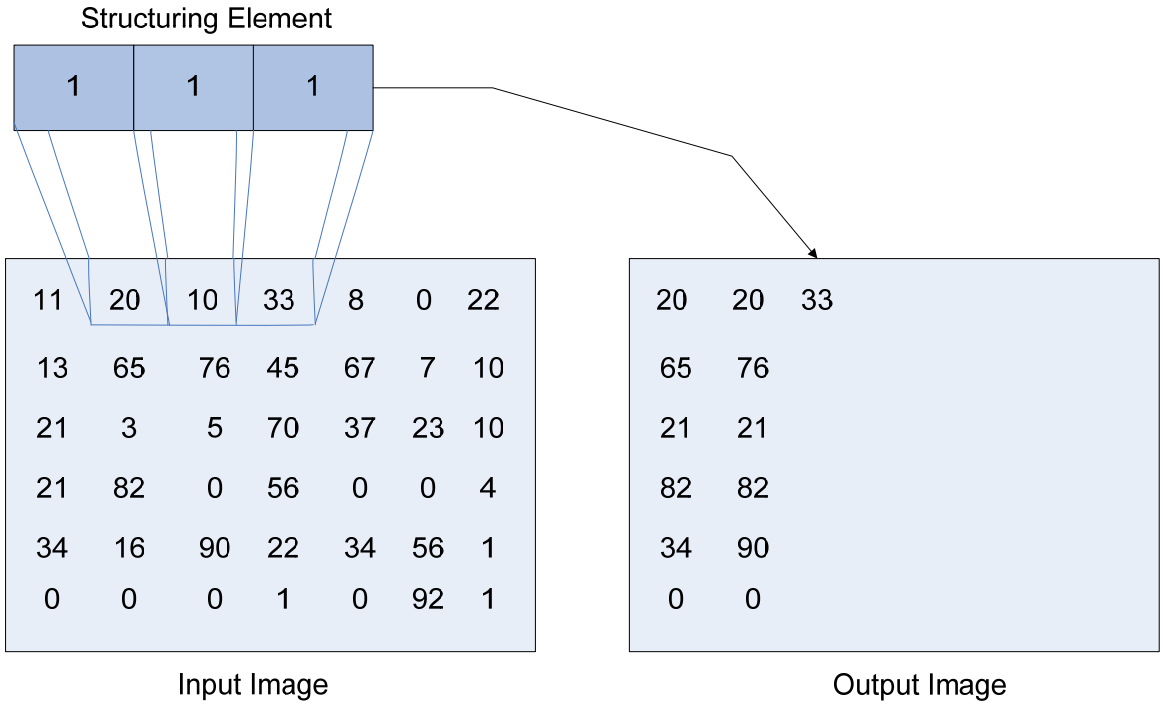


Figure 3-8 Morphological Dilation of Greyscale Image

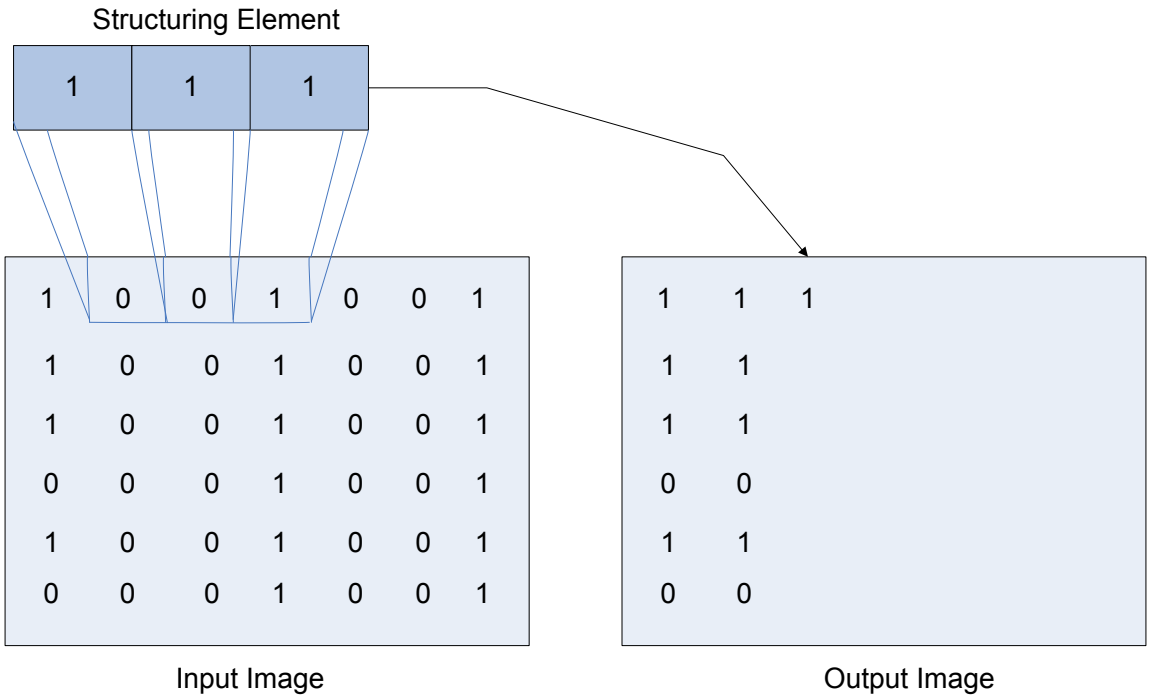


Figure 3-9 Morphological Dilation of Binary Image

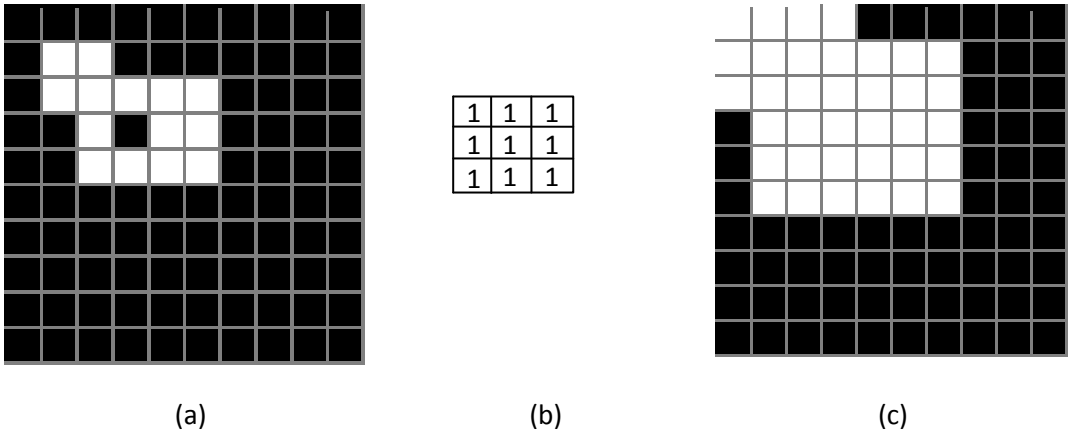


Figure 3-10 Binary Dilation Example (a) Original Image (b) Structuring Element (c) Dilated Image

One of the simplest applications of dilation is for bridging gaps in binary mode as shown in Figure 3-11.

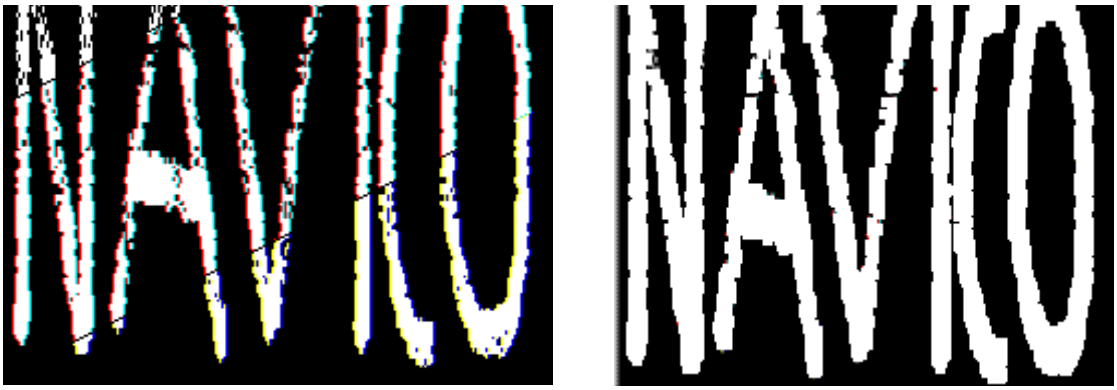


Image before dilation Image after two pixels dilation

Figure 3-11 Dilation Application

3.4.2 MORPHOLOGICAL EROSION

Let A and B be two sets in Z^2, the erosion of A by B is defined as:

$$A \Theta B = \{z \mid (B)_z \cap A^c = \phi\} \quad \text{3-13}$$

Where A⊖B is erosion of A by B, A^c is the complement of A and ϕ is the empty set. In other words, the erosion of A by B means the set of all points z such that B, translated by z, is contained in A.

In the case of image processing, the structuring element is shifted through all pixels of the image and the value of the output pixel is the minimum value of all the pixels in the input pixel's neighbourhood in erosion. In the case of binary image, if any of the pixels in the neighbourhood is 0, the output is set to 0. Figures 3-12 and 3-13 show the schematic representation of Grey level and binary erosion. Figure 3-14 shows an example of binary dilation.

One of the simplest applications of erosion is for thinning and increasing the space between gaps as shown in Figure 3-15.

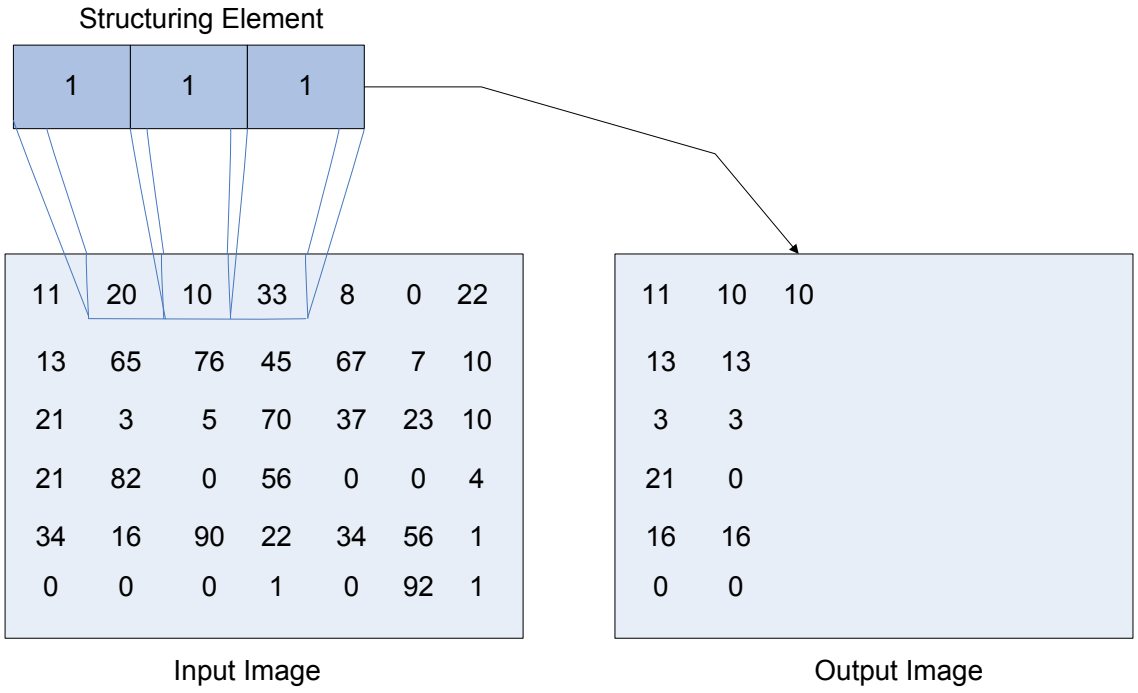


Figure 3-12 Morphological Erosion of Greyscale Image

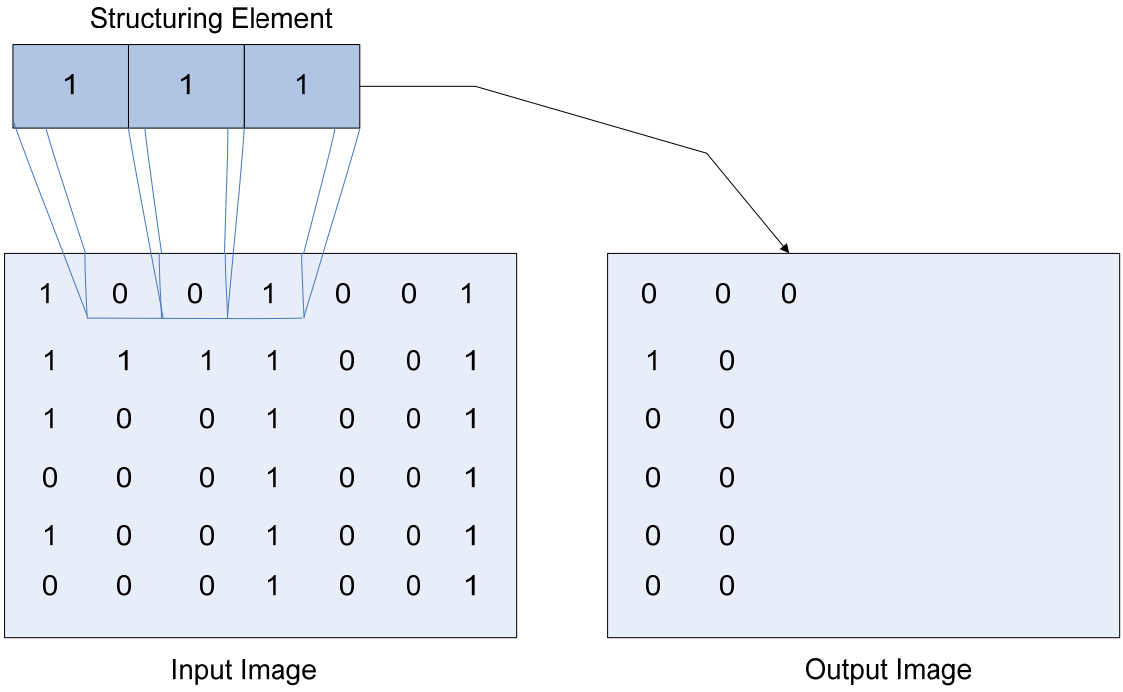


Figure 3-13 Morphological Erosion of Binary Image

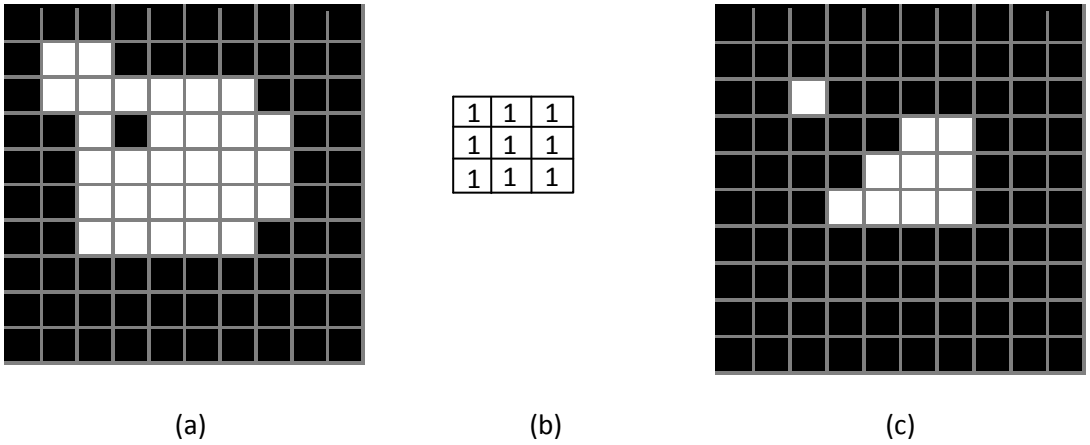


Figure 3-14 Binary Erosion Example (a) Original Image (b) Structuring Element (c) Eroded Image

Figure 3-15 Erosion Application

3.5 DISTANCE TRANSFORM

Distance transform is also known as distance map or distance field. Distance transform is normally applied to binary images. The result of distance transform is a gray level image, in which the gray level intensities of pixels are changed to show the distance to the closest object pixel from each pixel. For a 2-D binary image containing "object" and "background" pixels, distance transform is calculated by assigning to each background pixel (i,j) the distance to the closest object pixel (k,l).

Distance transform has the property of being relatively invariant to gray level intensities and lighting conditions. As shown in Figure 3-16, the two rectangles have different intensities but their distance transforms are similar. This property is particularly useful in building systems that are invariant to gray level intensities, e.g. Face Recognition under different skin colours, pose estimation under different skin colours and lighting conditions. Also, distance transform is more suitable for template matching compared to edge image because the resulting similarity measure is smoother as a function of the template transformation parameters. There are different variants of distance transforms depending on the distance metric used. The most common distance metrics used in calculating distance transform are Euclidian distance, Cityblock distance and Chessboard distance. Following is a brief description of these distance metrics:

- **Euclidian Distance:** Consider two pixels having coordinates (x1,y1) and (x2,y2), the Euclidian distance between these two pixels is calculated as:

$$D_{Euclid} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad \textbf{3-14}$$

  Figure 3-17 shows an example of Euclidian distance transform.

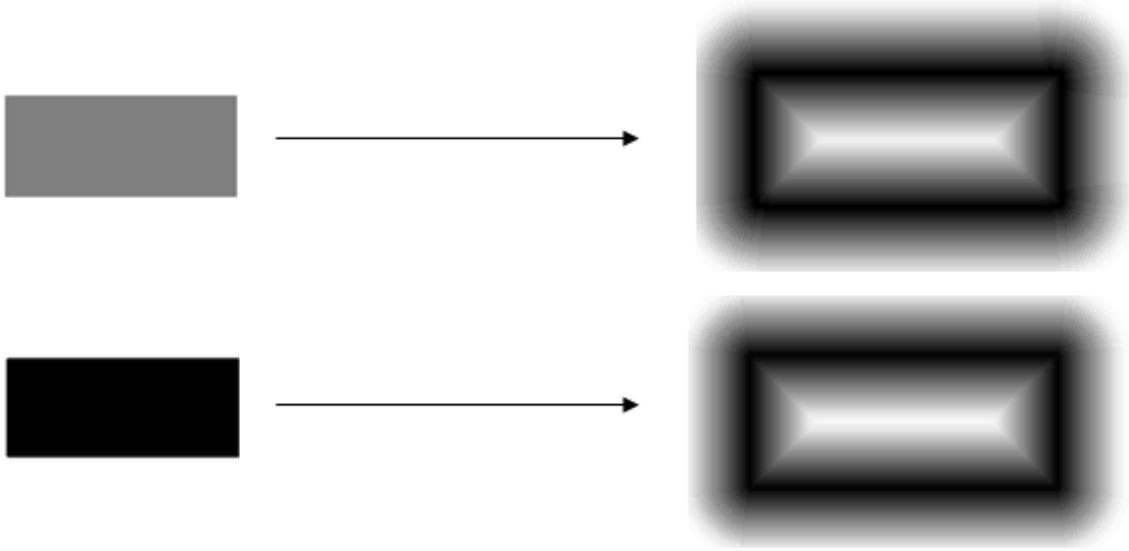
- **Cityblock Distance:** The Cityblock distance assumes that it is only possible to go from one pixel to another pixel along the gridline and no diagonal moves are allowed. The Cityblock distance between two pixels having coordinates (x1,y1) and (x2,y2) is calculated as:

$$D_{City} = | x_2 - x_1 | + | y_2 - y_1 | \quad \textbf{3-15}$$

  Cityblock distance is also known as Manhattan distance.

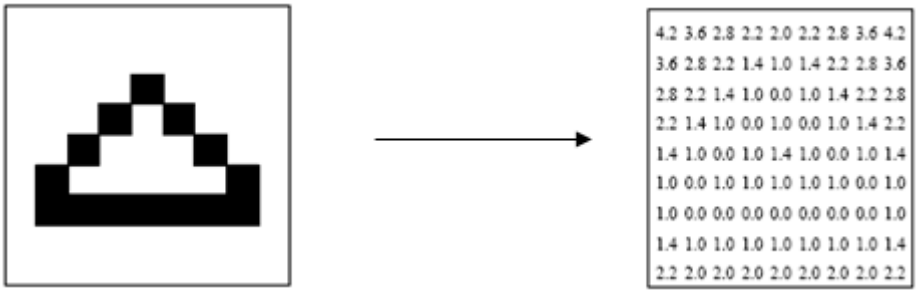
- **Chessboard Distance: In** Chessboard distance calculation each pixel is treated as the king in a chess game. A diagonal move counts the same as a horizontal or vertical move. Chessboard distance between two pixels having coordinates (x1,y1) and (x2,y2) is calculated as:

$$D_{Chess} = \max(| x_2 - x_1 |, | y_2 - y_1 |) \quad \textbf{3-16}$$



Original Images Distance Transform

Figure 3-16 Distance Transform Invariance to Intensity



Binary Image Euclidian Distance Transform

Figure 3-17 Euclidian Distance Transform

3.6 NORMALIZED CROSS CORRELATION

Template matching is usually done by finding cross-correlation between the input image and a template. Cross-correlation is based on squared Euclidian distance which is calculated as:

$$d^2_{f,t}(u,v) = \sum_{x,y}[f(x,y) - t(x-u, y-v)]^2 \qquad \textbf{3-17}$$

where *f*(x, y) is the image and the sum is over *x,y* under the window containing the feature *t* positioned at *u,v*. Equation 3-17 ca be expanded as:

$$d^2_{f,t}(u,v) = \sum_{x,y}[f^2(x,y) - 2f(x,y)t(x-u,y-v) + t^2(x-u,y-v)] \quad \textbf{3-18}$$

where the term $t^2(x-u,y-v)$ is constant and the term $f^2(x,y)$ is approximately constant. So the similarity measure between the image and template can be calculated as:

$$c(u,v) = \sum_{x,y} f(x,y)t(x-u,y-v) \quad \textbf{3-19}$$

However, using equation 3-19 for template matching has the following disadvantages:

- The range of c(*u,v*) is dependent on the size of the template.
- If the image energy function $\sum f^2(x,y)$ changes with position, equation 3-19 can fail. For example, the correlation between the template and exactly matching a region in the image may be less than the correlation between the template and a bright spot in the image.
- Equation 3-19 is not invariant to changes in image intensities, which may be caused by changing lighting condition for example.

These difficulties can be overcome by normalizing the image and template vector to unit length:

$$\gamma(u,v) = \frac{\sum_{x,y}[f(x,y) - \bar{f}_{u,v}][t(x-u,y-v) - \bar{t}]}{\{\sum_{x,y}[f(x,y) - \bar{f}_{u,v}]^2 \sum_{x,y}[t(x-u,y-v) - \bar{t}]^2\}^{0.5}} \quad \textbf{3-20}$$

where $\bar{t}$ is the mean of the template and $\bar{f}_{u,v}$ is the mean of image *f*(x,y) under the template. Equation 3-20 is known as normalized cross-correlation.

3.7 SUMMARY AND CONCLUSIONS

Fundamental concepts of image processing utilized in the proposed algorithms (i.e. chapters 4, 5 and 6) are briefly discussed in this chapter so that the remaining chapters are more easily understood. Three basic colour spaces i.e. RGB, HSV and YCbCr, and their mutual conversions are discussed briefly. The mouth region has a stronger red component as compared to its near regions. The eye regions have high C_b and low C_r values contain both dark and bright pixels near to each other in luma. So the morphological operators (e.g. erosion and dilation) can be designed to emphasize brighter and dark pixels in the luma component in the eye regions. Since all these information of eyes and mouth can be effectively exploited in YCbCr colour space, we use YCbCr colour space in our facial features detection schemes. Five of the most commonly used edge detectors i.e. Sobel, Robert's, Prewitt's, Laplacian and Canny, are discussed in detail in this chapter. While there is no single edge detector which performs well in all contexts and the performance of an edge detector is highly dependent on the quality of image, we use the Canny edge detector with a threshold value for the calculation of distance transform in our experiments because it detects the eyes and mouth edges (which are sharp enough) better as compared to other edge detectors. Fundamentals of the two basic morphological operators i.e. erosion and dilation, and their applications are discussed briefly. Distance transform, its different types and relative invariance to intensity, is demonstrated. Normalized cross-correlation and its relative invariance to image intensities, is also briefly discussed. Sample examples are provided to demonstrate all these techniques. Also, possible applications of these techniques are discussed briefly.

4 FACIAL FEATURES LOCALIZATION

4.1 INTRODUCTION

Facial feature extraction plays an important role in many applications, e.g. face recognition, facial expression recognition and behaviour analysis. Some face pose estimation methods are also based on the successful extraction of facial feature extraction. Our proposed pose estimation methods (which are discussed in detail in the following chapters) also rely on facial features, such as the eyes and mouth, have been successfully identified. Eyes and mouth are the most salient features of a face. Automatic facial feature extraction has been a focus of research in computer vision community for over decades. An overview of the existing feature extraction methods has already been presented in section 2.7. This chapter discusses two novel methods for eyes localization and one for mouth detection that resulted from this research project. One of the proposed eyes detection methods is based on edge-density information. This is based on the observation that edge-density is high around the eye regions compared to other regions in a face. The second method is a hybrid method which combines colour, edge and illumination information to detect the eyes. The mouth detection method is a modified version of Hsu et al. (2002) which uses adaptive thresholding to extract the exact corners of mouth. Once the corners of the eyes and mouth are detected then their centres are calculated using a set of equations which are based on the ratio of the length of two eyes. The proposed methods are discussed in detail in the following section, followed by experimental results and conclusions.

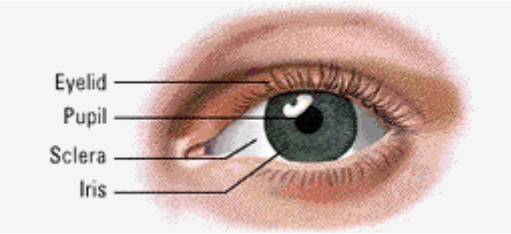


Figure 4-1 Typical Eye Structure

4.2 EYES DETECTION

The following two eyes detection methods assume that the face region is already detected and extracted using a face detector like the one described in Viola and Jones (2004).

4.2.1 EDGE-DENSITY BASED METHOD

The positions of the eyes can be extracted from an image using different eye features such as colour, illumination, edge, shape and geometrical information. Most colour variations occur in eye regions in facial images due to colour differences between the eyelids and skin, skin and sclera, sclera and iris, and iris and pupil (see Figure 4-1). The colours of the rest of the face are more uniform when compared to the eyes. Therefore, if edge detection was applied to a facial image then the eye regions would have maximum edge densities. The proposed method uses this edge density information to extract the eyes from a facial image. After edge detection, morphological dilation and erosion are used to enhance the identified blobs and unnecessary small edges are removed. Shape and geometry information is then used to extract and verify the eyes. The proposed scheme was tested on PICS (http://pics.psych.stir.ac.uk/) images database with the results presented in section 4.2.4. The overall algorithm is given below then followed by a more detailed description of the important steps in the algorithm. The flow diagram is shown in Figure 4-3.

ALGORITHM

Input: *head shoulder colour image*

Output: *head shoulders colour image with rectangles drawn around the two eyes*

Start

Step A: *Lighting Compensation*

Convert the Colour Image to Grey Level

Do histogram Equalization

Step B: *Edge Detection*

Detect Edges Using Sobel Operator

Step C: *Dilation, Holes Filling and Erosion*

Dilate the Image Twice Successively

Fill the Small Holes inside Connected Regions

Erode the Image Three Times Successively

Step D: *Eyes Extraction and Verification*

Apply the Shape and Geometry Based Rules and Reject the Regions those don't Satisfy These Rules

Draw Rectangles around the Remaining Connected Regions

End

Step A: Lighting Compensation

Some of the images in the PICS (http://pics.psych.stir.ac.uk/) database have very poor contrast due to poor lighting condition. If edge detection was applied directly to these images then the sharp edges might not be identified properly. Therefore, the contrast of the images is adjusted first to get better results. Simple histogram equalization has been used in our experiments for contrast adjustment. Figure 4-2 shows the result of histogram equalization.

(a) (b)

Figure 4-2 Lighting Compensation (a) Original Image (b) Light-Compensated Image

Step B. Edge detection

Sobel operator is applied to detect the edges. Since the edges in the eye regions are sufficiently sharp, so Sobel edge detector is used. Figure 4-4 shows the result of edge detection.

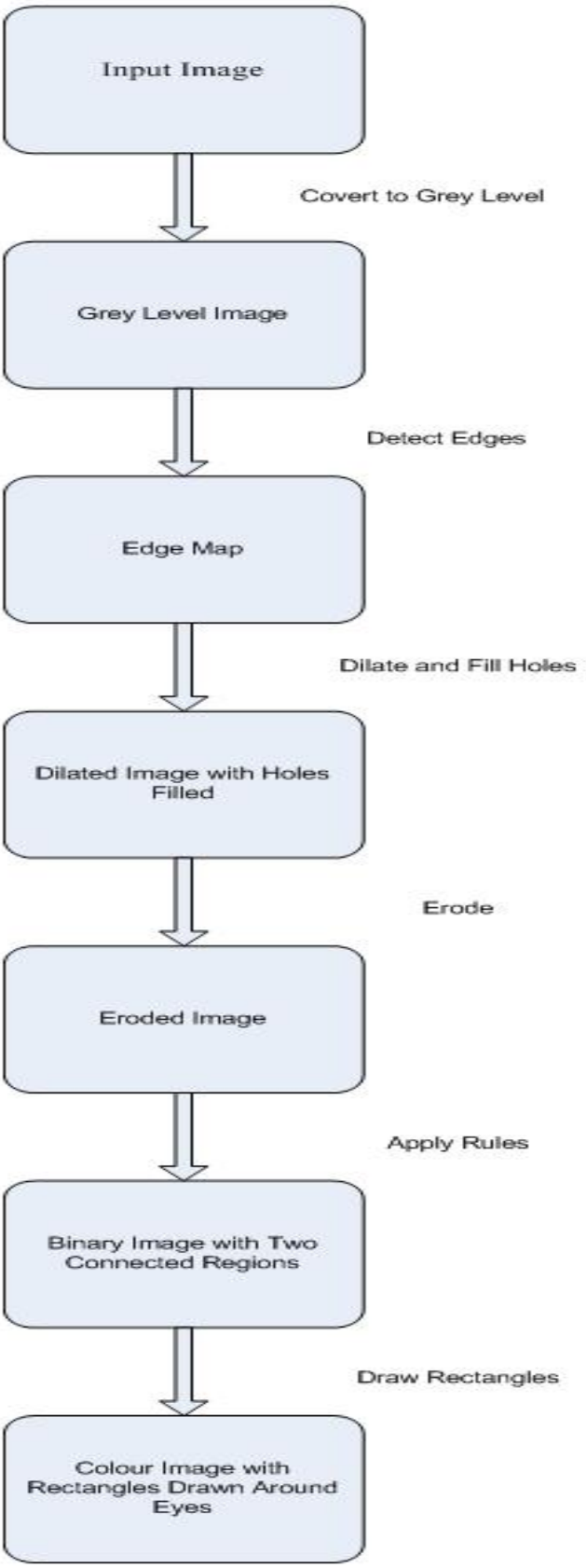


Figure 4-3 Flow Diagram of the Algorithm

(a) (b)

Figure 4-4 Edge Detection (a) Grey Level Image (b) Detected Edges

Step C. Dilation, holes filling and Erosion

Morphological dilation is used to enhance the edges around the eye regions. The image is dilated twice to fill the holes in the eyes. Disk shape structuring element with radius 3 has been used for dilation in our experiments. After this step the eyes become filled regions. However, sometimes small holes are left inside an eye region which may hinder the algorithm during erosion. Therefore, any small hole inside an eye region needs to be filled before erosion. To fill the small holes the negative of a dilated image is taken and regions that have very small areas are discarded, as the small regions are very likely to be small holes. The negative of the image is taken again to get the image without any of the small holes. The unwanted edges are also removed by eroding the image three times. Figures 4-5, 4-6 and 4-7 show hole filling, dilation and erosion respectively.

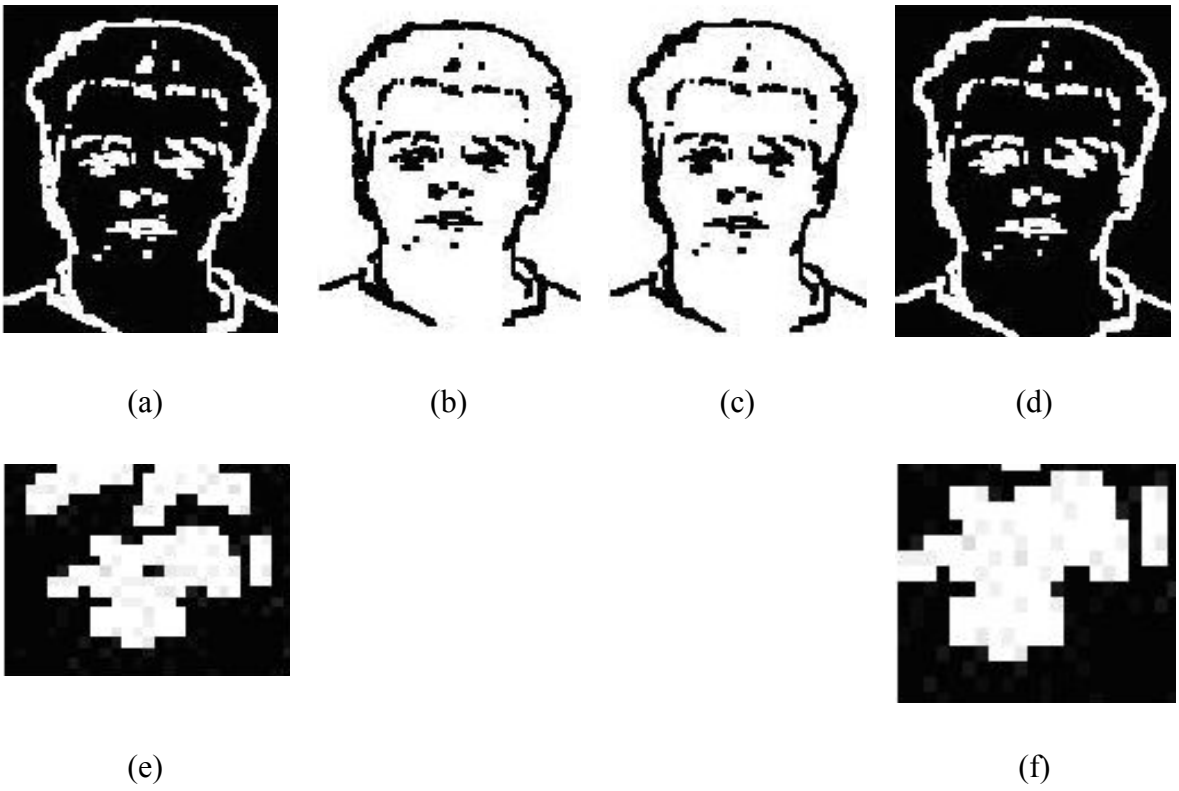


Figure 4-5 Holes Filling (a) Image With Small Holes (b) Negative Image With Holes (c) Negative Image With Holes Filled (d) Final Image With Holes Filled (e) Enlarged Sample Hole (f) Enlarged Filled Hole

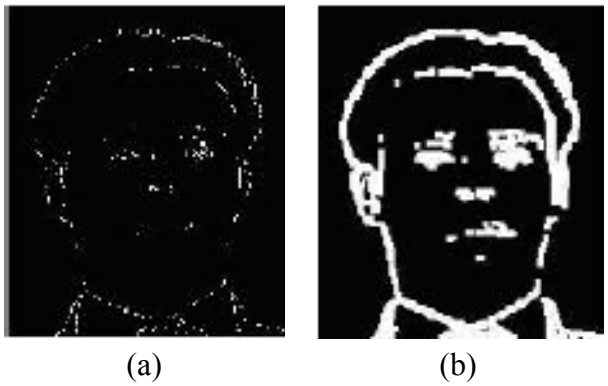


Figure 4-6 Morphological Dilation (a) Before Dilation (b) After Dilation

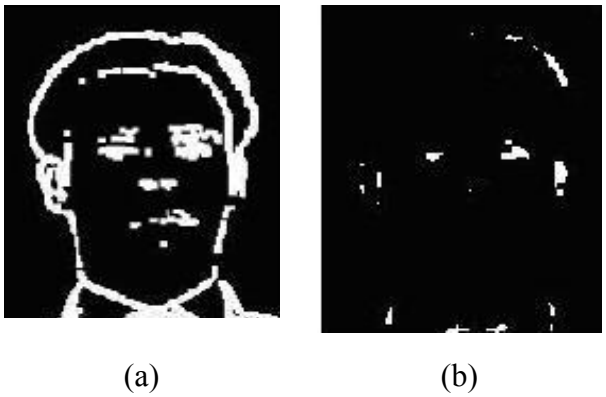


Figure 4-7 Morphological Erosion (a) Dilated Image (b) Eroded Image

Step D. Eyes Extraction and Verification

After the above steps, the image contains several blobs. In order to find which of them are eyes, the following rules based on the shape and geometry of eyes are applied.

- The aspect ratios (width/height) of the eyes regions are between 0.8 and 4.0.
- The orientation angle of eyes is not greater than 45 degrees.
- Size of one eye shouldn't be greater than twice of the other eye and shouldn't be smaller than half of the other.
- The orientation of the two eyes shouldn't differ by more than 30 degrees.
- The line joining two eyes shouldn't have slope greater than 45 degree.
- The eyes are not too close to the borders of the image.

The above rules are applied successively. The region which does not conform to any of the rules is removed and thus is not considered for further rule tests. This successive testing scheme speeds up the process.

4.2.2 HYBRID METHOD

The proposed hybrid method combines edge, intensity and illumination cues to detect eyes. The edge based method is already discussed in detail in the previous section. The illumination and colour based methods are discussed in detail in the following section followed by the description of the proposed hybrid method.

4.2.2.1 Illumination-based Method

Illumination based method has been proposed by Hsu et al. (2002) which works for colour images. In their method, the input RGB image is converted first to YCbCr. Two separate eye maps, one from chrominance component of the image and the other from luminance, are built. The eye map from the Chroma is based on the observation that high C_b and low C_r values are present around the eyes. It is calculated using the equation:

$$EyeMapC = \frac{1}{3}\left\{C_b^2 + \bar{C}_r^2 + {C_b}/{C_r}\right\} \quad \text{4-1}$$

Where C_b, C_r and $\overline{C}_r$ are the blue, red and negative of red chroma components respectively. These values are normalized to the range [0, 255]. The eye map from luma is based on the observation that eye regions contain both dark and bright pixels near to each other in luma. So the morphological operators (e.g. erosion and dilation) can be designed to emphasize brighter and dark pixels in the luma component in the eye regions. The eye map from the luma component is calculated as follow:

$$EyeMapL = \frac{Y(x,y) \oplus g_\sigma(x,y)}{Y(x,y) \ominus g_\sigma(x,y)} \quad \textbf{4-2}$$

where Y(x,y) is the luma component of the image and $g_\sigma(x,y)$ is the structuring element, and $\oplus$ and $\ominus$ are morphological dilation and erosion respectively. These two maps are then combined into a single eye map using the following formula:

$$EyeMap = (EyeMapC) AND (EyeMapL) \quad \textbf{4-3}$$

Figure 4-8 shows the step by step output of this method.

4.2.2.2 Intensity-based Method

This method proposed by Chiang et al. (2003) is based on the observation that the eyes are the darkest regions of the face. In this method the colour image is converted to a histogram equalized grey level image. The eyes regions are then extracted from this histogram equalized image using a threshold operation with a threshold value of 20. Apart from the eyes, some other dark regions are also extracted as connected regions when this threshold operation is applied. Unwanted regions are then removed using a component verification process. This method is very fast and simple but it does not work for people with dark skin. Figure 4-9 shows the step by step output of this method.

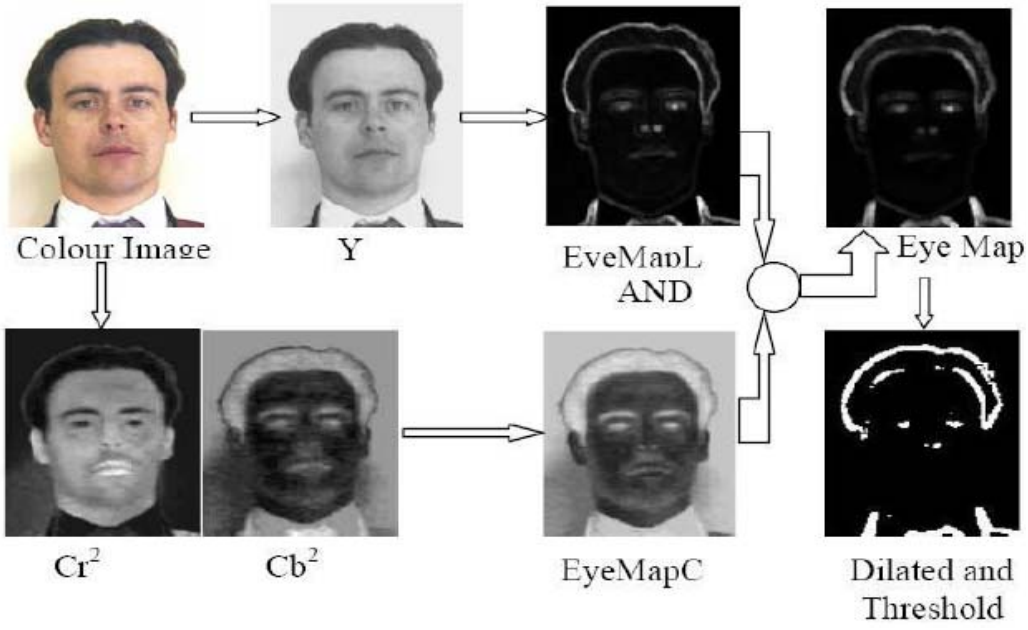


Figure 4-8 Illumination-based Method

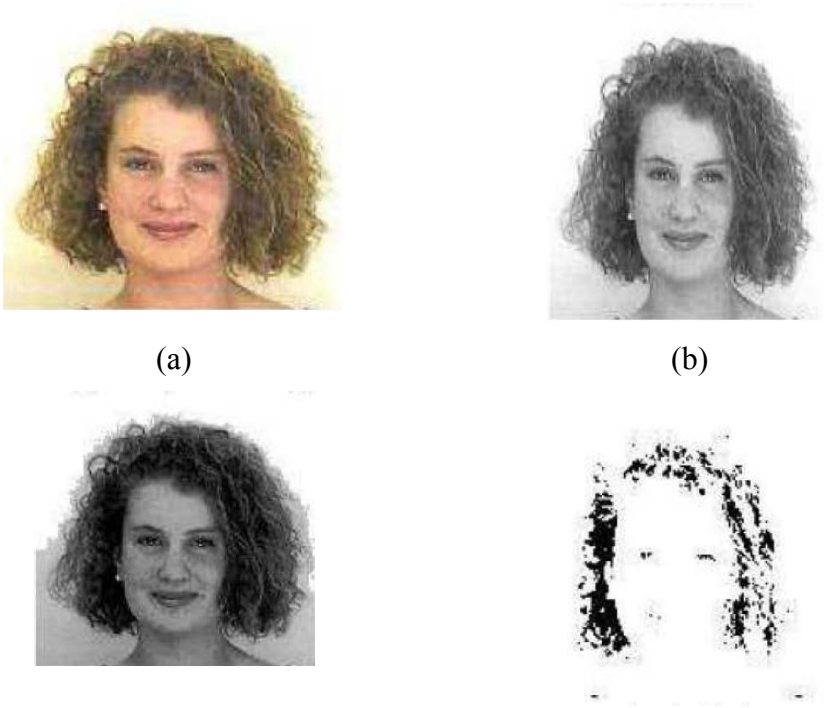


Figure 4-9 Intensity-Based Method (a) Colour Image (b) Grey Level Image (c) Histogram Equalized Image (d) Thresholded Image

4.2.2.3 Proposed Hybrid Method

For the methods discussed above, when they fail to locate the eyes they are for different reasons. Therefore, the proposed method is to combine the three methods in a systematic way to overcome the weaknesses of the individual methods and result in a method that is more accurate and robust than each of the individual methods. In the proposed method, blobs are detected using each of the methods. Each of the blobs obtained from the three methods are then tested using the following rules. If a blob that does not satisfy any of the rules then it is removed.

- The solidity is of the region is greater than 0.5
- The aspect ratio is between 0.8 and 4.0
- The connected region is not touching the border
- The orientation of the connected component is between -45 and +45 degrees

Given $Image_{Illu}$, $Image_{Col}$ and $Image_{Edge}$ which are the binary images containing the blobs obtained through the illumination-based, intensity-based and edge-density-based methods respectively, all the possible pairs of these images are made and images in each pair are combined through a bitwise (pixel by pixel) AND operators as follow:

$$Image_{IlluCol} = Image_{Illu}\ AND\ Image_{Col} \quad \textbf{4-4}$$

$$Image_{ColEdge} = Image_{Col}\ AND\ Image_{Edge} \quad \textbf{4-5}$$

$$Image_{IlluEdge} = Image_{Illu}\ AND\ Image_{Edge} \quad \textbf{4-6}$$

The unwanted connected regions which are detected by one method but not by others are automatically removed due to this bitwise AND operation. The next step is to combine the three images, i.e. $Image_{IlluCol}$, $Image_{ColEdge}$ and $Image_{IlluEdge}$, using a bitwise (pixel by pixel) OR operation:

$$Image_{Hybrid} = Image_{IlluCol}\ \mathbf{OR}\ Image_{ColEdge}\ OR\ Image_{IlluEdge} \quad \textbf{4-7}$$

Shape and geometry based rules are again applied to $Image_{Hybrid}$ to remove candidates that are not likely to be eyes. Figure 4-10 shows the schematic representation of the proposed method.

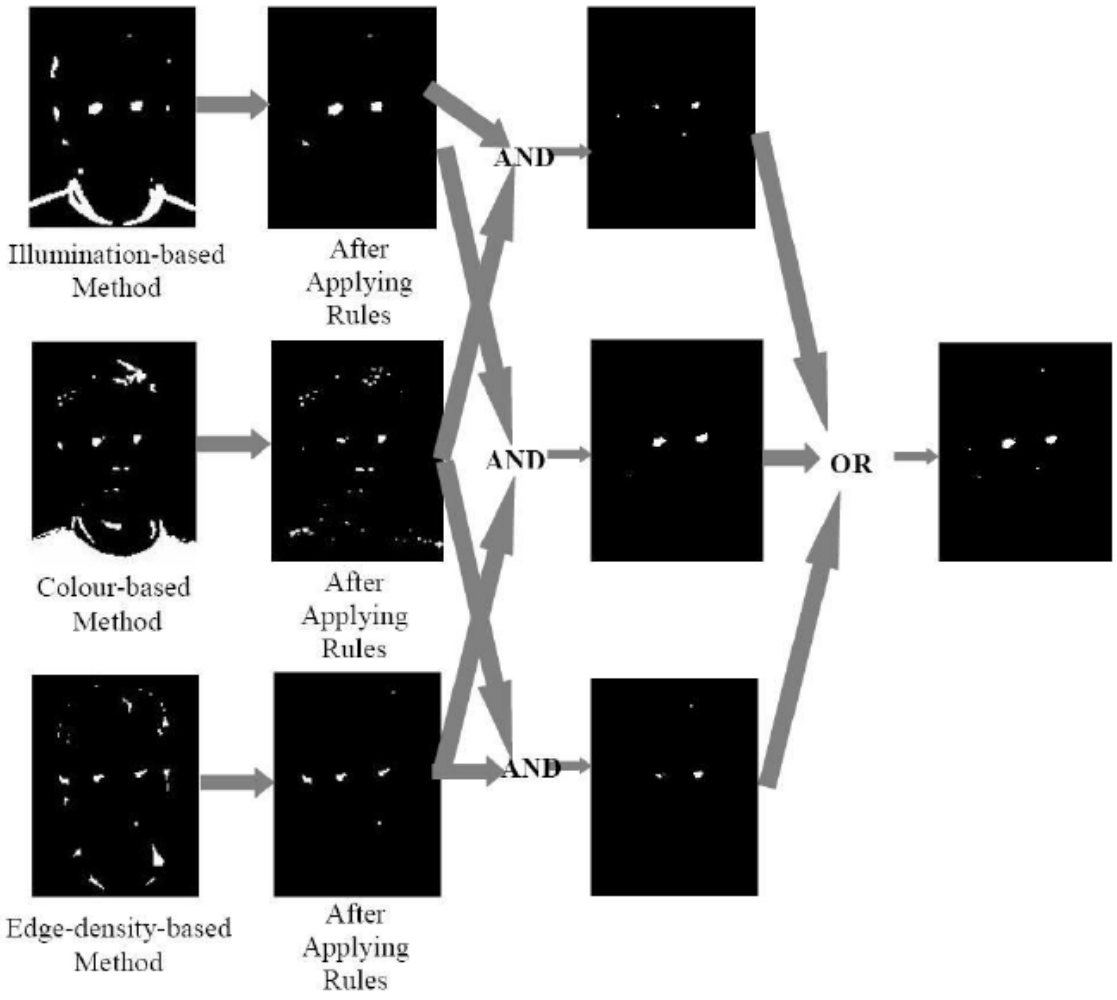


Figure 4-10 Proposed Hybrid Method

4.2.3 EYE CENTRES LOCALIZATION

Eye centre localization is important for our geometric pose estimation method (discussed in chapter 5). Once the corners of an eye are detected then the midpoint of the two corners gives a good approximation for the centre of the eye if it is a frontal image. However, the error increases if the face is directed further away from the front. Therefore, instead of using simple midpoint, the following equations based on the ratio of the lengths of the two eyes on a facial image are used to find the centre of the eyes.

$$Centre_{LeftEye_x} = \frac{d_{Left}^2}{\sqrt{1 + \left(\frac{y_{LR} - y_{LL}}{x_{LR} - x_{LL}}\right)^2}} + x_{LL} \qquad 4\text{-}8$$

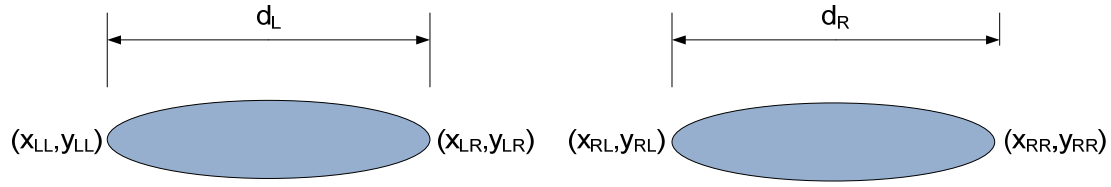


Figure 4-11 Eyes Model

$$Centre_{LeftEye_Y} = \left(\frac{y_{LR} - y_{LL}}{x_{LR} - x_{LL}}\right) \times \left[Centre_{LeftEye_X} - x_{LL}\right] + y_{LL} \quad \textbf{4-9}$$

$$Centre_{RightEye_X} = \sqrt{\frac{d_{Right}^2}{1 + \left(\frac{y_{RR} - y_{RL}}{x_{RR} - x_{RL}}\right)^2}} + x_{RL} \quad \textbf{4-10}$$

$$Centre_{RightEye_Y} = \left(\frac{y_{RR} - y_{RL}}{x_{RR} - x_{RL}}\right) \times \left[Centre_{RightEye_X} - x_{RL}\right] + y_{RL} \quad \textbf{4-11}$$

where (x_{LL}, y_{LL}), (x_{LR}, y_{LR}), (x_{RL}, y_{RL}), (x_{RR}, y_{RR}) are the corners of eyes, and d_L/d_R is the ratio of the lengths of left and right eyes as shown in Figure 4-11 and d_{Left} and d_{Right} are calculated as:

$$d_{Left} = \frac{d_L}{d_R} \times \frac{\sqrt{(x_{LR} - x_{LL})^2 + (y_{LR} - y_{LL})^2}}{2} \quad \textbf{4-12}$$

$$d_{Right} = \left(1 - 0.1 \times \frac{d_L}{d_R}\right) \times \frac{\sqrt{(x_{RR} - x_{RL})^2 + (y_{RR} - y_{RL})^2}}{2} \quad \textbf{4-13}$$

The above equations are used when d_L/d_R is less than or equal to 1. When d_L/d_R is greater than 1, then the left corner coordinates are replaced by the right corner coordinates in these equations.

4.2.4 Experimental Results and Analysis

Experimental results and analysis for all the three methods i.e. edge-density method, hybrid method and eyes centres localization are discussed in the following section.

4.2.4.1 Edge-Density Based method

The proposed idea was implemented in Matlab using the image processing toolbox and was tested using the PICS (http://pics.psych.stir.ac.uk/) facial images database. Eighty photos of fifty individuals, including both males and females of different ages and ethnicities, were randomly selected. Most of them were frontal faces. However, some faces were tilted to the left or to the right. These are coloured images with widths vary from 360 to 480 pixels while their heights vary from 480 to 540 pixels. Using an Intel core duo 1.60 GHz Processor, the average processing time of an image using this method was one second. Table 4-1 shows the results and percentage accuracy of the proposed algorithm. The table shows that the algorithm is very accurate in initial blobs extraction. From the table it is also clear that accuracy increases with good illumination images.

In Table 4-2, the proposed algorithm is compared with two other methods based on colour (Chiang et al. (2003)) and illumination (Hsu et al. (2002)) respectively. Since these methods do not mention their eyes verification steps explicitly comparison is done only on blob extraction and not on final eye detection. Figure 4-12 shows some of the example images in which eyes were detected accurately. Figure 4-13 shows an example image where the proposed method performed better than the other two methods.

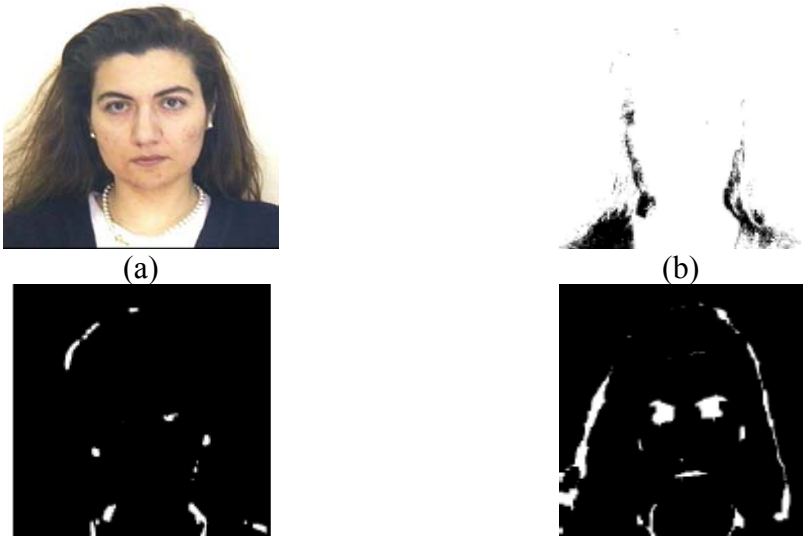
Table 4-1 Results of Proposed Edge-density-Based Method

Step	Images Type	Total Images	Correct Detection	Incorrect Detection	% Accuracy
Blobs Extraction	All Images	80	76	4	95%
	Good Illumination Images	75	72	3	96%
Final Eyes Detection	All Images	80	58	22	72.5%
	Good Illumination Images	75	58	17	77.33%

Table 4-2 Comparison of Proposed Edge-Density Based Method with Existing Methods

Method	Images Type	Total Images	Correct Detection	Incorrect Detection	% Accuracy
Colour Based Method	All Images	80	76	4	95%
	Good Illumination Images	75	72	3	96%
Illumination Based Method	All Images	80	72	8	90%
	Good Illumination Images	75	66	9	88%
Proposed Edge Based Method	All Images	80	72	8	90%
	Good Illumination Images	75	72	3	96%

Figure 4-12 Sample Images in which Eyes were correctly Detected



(a) (b)

(c) (d)

Figure 4-13 Sample Image for which Proposed Edge Density Based Method Performs better (a) Original Image (b) Blobs Detected by Colour Based Method (c) Blobs Detected by Illumination Based Method (d) Blobs Detected by Proposed Method

4.2.4.2 Hybrid Method

The proposed hybrid method described in a previous section was implemented in Matlab using the image processing toolbox and was tested using the PICS (http://pics.psych.stir.ac.uk/) facial images database. One hundred and sixty photos including both males and females of different ages and ethnicities were randomly selected from the database. Most of them were frontal faces. However, some faces were tilted to the left or to the right. These are colour images with widths vary from 360 to 480 pixels while their heights vary from 480 to 540 pixels. The accuracy of eye detection of the proposed method is 87%. In Table 4-3, the proposed hybrid algorithm is compared with the three existing methods operating on their own. Since the methods by Hsu et al. (2002) and Chiang et al.(2003) do not mention their eyes verification steps explicitly comparison is done only on the initial blobs extraction and not the final eye detection. The qualitative comparison is summarized in Table 4-4. Figure 4-14 shows some of the output images from the hybrid method. Figure 4-15 shows example images where the hybrid method performed better than individual methods.

Table 4-3 Comparison of Proposed Hybrid Method with Existing Methods

Method	Total Images	Correct Detection	Incorrect Detection	Percentage Accuracy
Illumination-based Method	160	144	16	90.00%
Colour-Based Method	160	134	26	83.75%
Edge-Density-Based Method	160	128	32	80.00%
Proposed Hybrid Method	160	150	5	93.75%

Table 4-4 Comparison Summary

Method	Closed Eyes	Dark Skin Colour	Bad Illumination
Illumination-based Method	✓	✓	✖
Colour-Based Method	✓	✖	✓
Edge-Density-Based Method	✖	✓	✓
Proposed Hybrid Method	✓	✓	✓

Figure 4-14 Examples of Eyes Detection by Hybrid Method

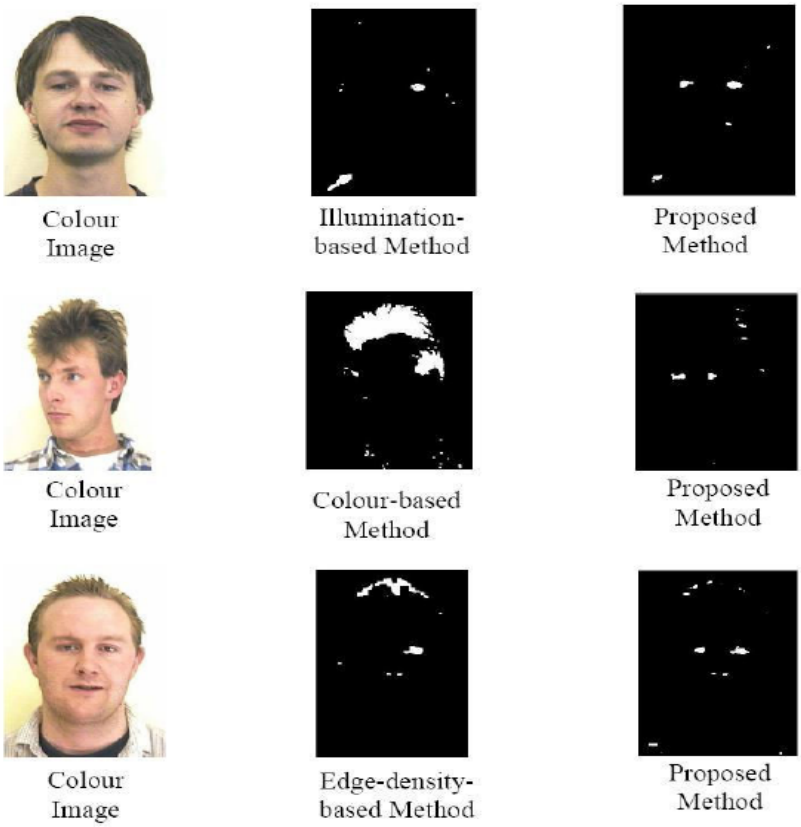


Figure 4-15 Example Images for which the Proposed Hybrid Method Performed Better

4.2.4.3 Eye Centre Localization

To test the equations derived in section 4.2.3 for eyes centre estimation, an in-house database which contains images of 6 people with varying horizontal poses, was used. It should be noted that the centre of an eye means the projection of midpoint between two corners of the eye in three dimensional space and not the pupil's centre. Figure 4-16 shows some of the sample results from the tests. In Figure 4-16, the black dots show the centres of eyes detected by the simple midpoint method and the white dots show the results of the proposed method. The actual centres are almost completely overlapped by the white dots, so they cannot be seen in Figure 4-16. In detecting the centre of an eye only the horizontal accuracy has been tested, which is the requirement of the proposed geometric pose estimation scheme discussed in chapter 5. Figure 4-17 shows the error comparison for a sample sequence of images from the in-house database, which is calculated as distance between the actual centre and detected centre divided by the total length of the eye where the actual centre was marked manually to compare the results. There are many techniques for coarse and fine eyes localization (some of them are discussed in literature review chapter) in the literature but to the best of author's knowledge, no one has published the eyes centres calculation once the corners are detected. The obvious reason for this could be that they did not feel the need of centres calculation in their work. That is the reason that the proposed scheme is compared with a simple mid-point estimation only.

Figure 4-16 Eyes Centre Detection Examples

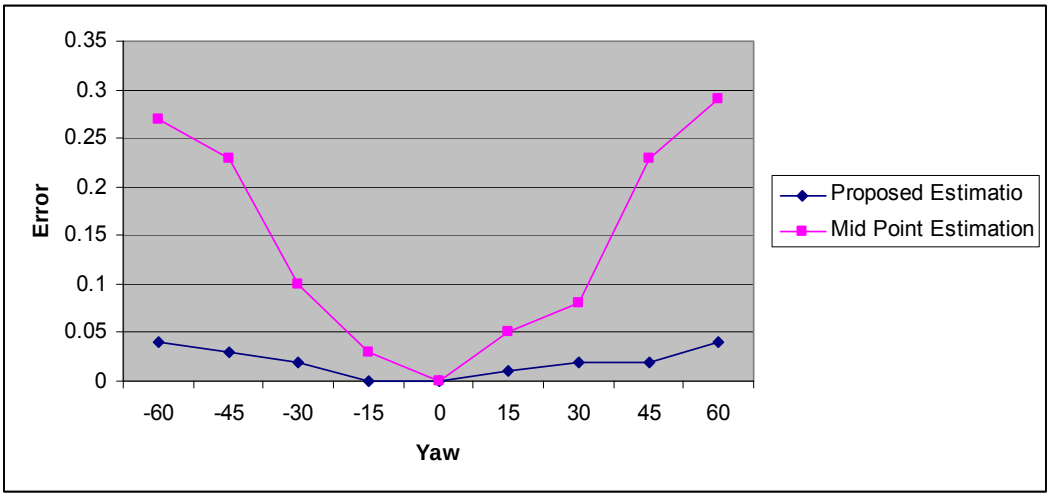


Figure 4-17 Error in Eye Centre Detection

4.3 MOUTH DETECTION

The following section describes the mouth detection algorithms followed by experimental results and analysis.

4.3.1 METHODS

Once the centres of both eyes are identified, the image is rotated so that the line joining two eyes becomes horizontal. This rotation operation makes the mouth detection step quicker because the processing time of an image is reduced when the targeted features are aligned with the horizontal or vertical axis. The lower half of the rotated face image is the target location for the mouth. Figure 4-19 shows how this rotation operation works. The mouth detection algorithm presented in Hsu et al. (2002) is used to detect the coarse mouth location. Their method is based on the observation that mouth region contains stronger red component and weaker blue component as compared to other facial regions, which means that C_r is greater than C_b for mouth region. The mouth region also has relatively low response in C_r/C_b and high response in C_r^2. The mouth map is calculated as:

$$\mathrm{MouthMap} = C_r^2 \times \left(C_r^2 - \eta \times {C_r}/{C_b}\right)^2 \qquad \text{4-14}$$

$$\eta = 0.95 \times \frac{\frac{1}{n}\sum_{(x,y)\in FG} C_r(x,y)^2}{\frac{1}{n}\sum_{(x,y)\in FG} C_r(x,y)/C_b(x,y)} \qquad \text{4-15}$$

where both C_r/C_b and C_r^2 are normalized to the range [0, 255], and n is the number of pixels within the face mask I^G. The mouth map construction is shown in Figure 4-18.

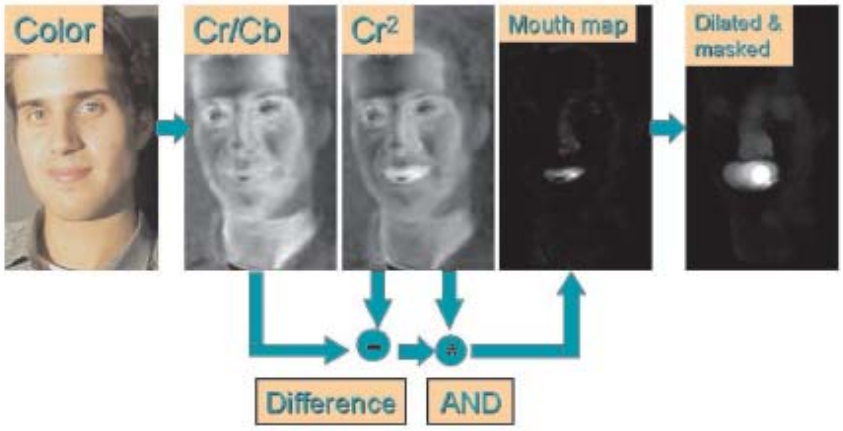


Figure 4-18 Mouth Map Construction adapted from Hsu et al. (2002)

After applying the mouth map, a simple threshold operation is sufficient to identify the mouth and its two corners. Figure 4-20 shows how the mouth corners are detected. Again, simple midpoint is not a good approximation for the centre of the mouth. Therefore, the following equations, which are based on the ratio of the lengths of the two eyes, are used to find the centre of the mouth.

$$Centre_{Mouth_x} = \sqrt{\frac{d_{Mouth}^2}{1 + \left(\frac{y_{MR} - y_{ML}}{x_{MR} - x_{ML}}\right)^2}} + x_{ML} \qquad \textbf{4-16}$$

$$Centre_{Mouth_y} = \left(\frac{y_{MR} - y_{ML}}{x_{MR} - x_{ML}}\right) \times [Centre]_{Mouth_x} - x_{ML}) + y_{ML} \qquad \textbf{4-17}$$

where x_{ML}, y_{ML}, x_{MR} and y_{MR} are the mouth corners as shown in Figure 4-21 and d_{Mouth} as calculated as:

$$d_{Mouth} = \frac{d_L}{d_R} \times \frac{\sqrt{(x_{MR} - x_{ML})^2 + (y_{MR} - y_{ML})^2}}{2} \qquad \textbf{4-18}$$

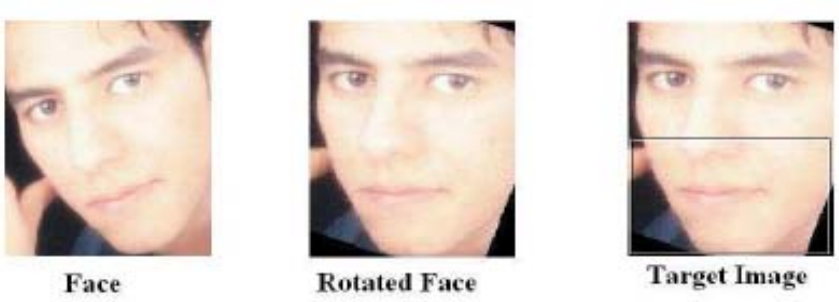
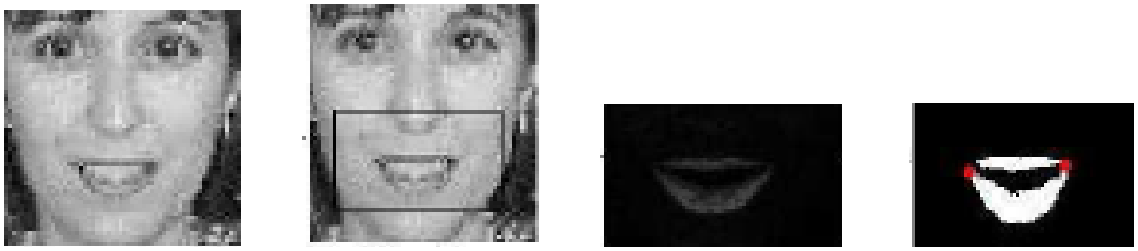


Figure 4-19 Face Rotation



(a) (b) (c) (d)

Figure 4-20 Mouth Corners Detection (a) Face Image (b) Mouth Search Area (c) Mouth Map (d) Mouth Corners

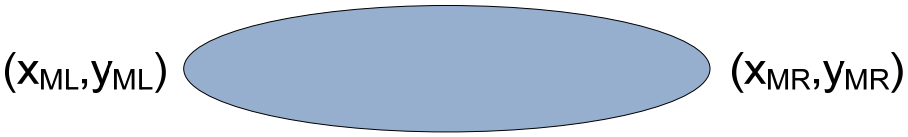


Figure 4-21 mouth Model

4.3.2 EXPERIMENTAL RESULTS AND ANALYSIS

To test the equations derived in section 4.3 for mouth centre estimation, an in-house database which contains images of 6 people with varying horizontal poses, was used. Figure 4-22 shows some of the sample results from the tests. In detecting the centre of a mouth only the horizontal accuracy is tested, which is the requirement of the proposed geometric pose estimation scheme discussed in detail in chapter 5. Figure 4-23 shows the error comparison against the mid point method using a sample sequence of images from

the in-house database. An error is calculated as the distance between the actual centre and detected centre divided by the total length of the mouth where the actual centre was marked manually. . There are many techniques for coarse and fine mouth localization (some of them are discussed in literature review chapter) in the literature but to the best of author's knowledge, no one has published the mouth centre calculation once the corners are detected. The obvious reason for this could be that they did not feel the need of centres calculation in their work. That is the reason that the proposed scheme is compared with a simple mid-point estimation only.

Figure 4-22 Mouth Centre Detection Examples

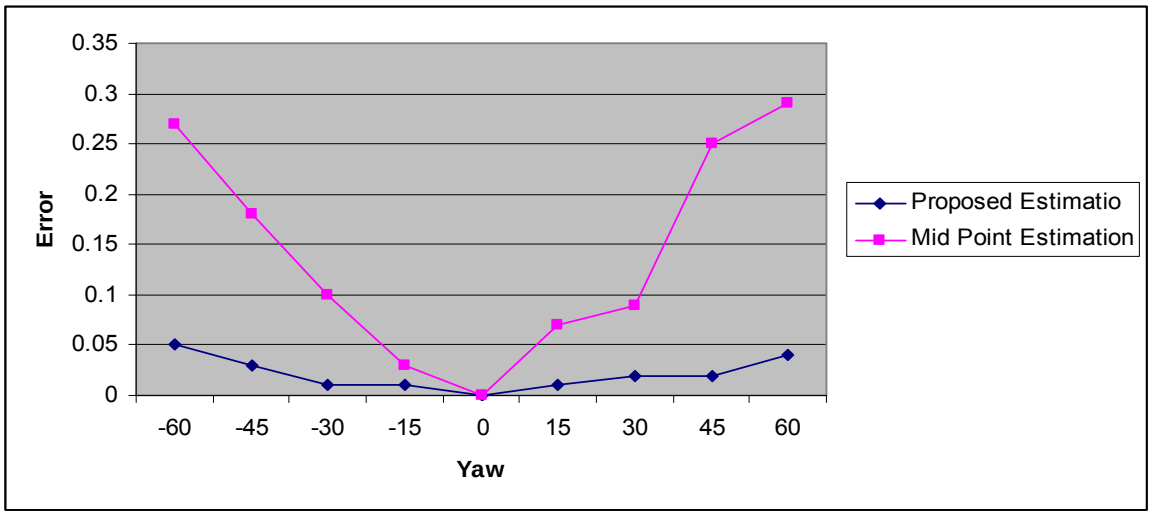


Figure 4-23 Error in Mouth Centre Detection

4.4 SUMMARY AND CONCLUSIONS

This chapter discussed the proposed facial feature extraction methods. Two novel eyes detection methods were presented. The first method is based on the observation that edge density is high around the eye regions as compared to other regions of face. The edge density based method has the advantage of being applicable to different skin colours because it is relatively insensitive to skin colour. The second is a hybrid method which combines edge, colour and illumination cues to extract the eyes. The hybrid method overcomes the weaknesses of each of the individual methods and thus giving more accurate results as compared to each of the individual methods. The two methods were tested using the PICS (http://pics.psych.stir.ac.uk/) image database with very good result. Once the eyes are localized, a set of equations for finding the centres of the eyes is also presented which utilize the ratio of the lengths of two eyes. A modified version of Hsu et al. (2002) mouth detection scheme has also been discussed in detail in this chapter. A set of equations which utilizes the ratio of the lengths of the two eyes was also presented for localizing the mouth centre. The proposed method was tested using an in-house facial image database and the results are significantly better than the mid point method.

5 FACE POSE ESTIMATION USING EYES-MOUTH TRIANGLE

5.1 INTRODUCTION

Facial pose estimation plays an important role in applications such as pose-independent face recognition, gaze estimation, virtual reality applications and human computer interaction. As suggested by Murphy-Chutorian and Trivedi (2009), an ideal pose estimation system should be able to:

- provide a reasonable estimate of the face pose with minimum error.
- work with images taken with a monocular camera because they are more suitable for real life applications.
- work autonomously. There should not be any manual initialization, feature selection, etc.
- work across all identities.
- estimate continuous range of head orientation in real time.
- work for a whole range of poses, even if the face is directed away from the camera.
- apply to both low and high resolution images.

In this chapter, a novel geometric yaw estimation method for near-field images is proposed which is based on the observation that eyes-mouth triangle has a distinct shape for distinct yaw angle. The method uses the shape of an eyes-mouth triangle to estimate face pose. The eyes and mouth centres are localized first using the methods discussed in chapter 4. To minimize the inter-subject variations, the triangle obtained from the eyes and mouth centres is rotated and normalized, i.e. resized to a standard size. The shape of this normalized triangle is then used to estimate face pose. The proposed method has the following advantages over other pose estimation systems in literature:

- It is based on mouth and eyes which are the most stable features of face according to Gourier et al. (2004).

- It is based only on three points which makes it less computationally intensive and hence suitable for real-time applications. Existing pose estimation methods in the literature, e.g. Gee and Cipolla (1994), Ho and Huang (1998) and Wang and Sung (2001), use more than three feature points.

- Unlike Gee and Cipolla (1994), Ho and Huang (1998) and Wang and Sung (2001), the proposed method uses centres instead of corners of eyes and mouth. Since centres can be calculated without the explicit exposure of eyes and mouth corners, this increases the range of pose estimation.

Typical pose estimation comprises Yaw, Pitch and Roll estimation of a face. The main target of the proposed method is Yaw estimation which is the requirement of most real life applications, e.g. interactive games, virtual reality applications and driver vigilance assessment. However, to be suitable for real life applications, the proposed method is capable of working within -15 to 15 degrees of pitch variations.

The rest of the chapter is organized as follow: section 5.2 discusses the proposed scheme, section 5.3 presents the experimental results and analysis and section 5.4 concludes the chapter.

5.2 PROPOSED METHOD

The face in the input image is detected using the adaboost face detector by Viola and Jones (2004). The coarse eyes and mouth are first localized and then fine corners are detected using the facial features localization methods discussed in chapter 4. It should be noted here that by corners, here we mean the visible extremes of eyes and mouth regions and not the actual corners which may not be visible in the image.

The triangle made by the centres of the two eyes and mouth has a distinct shape for distinct face pose. As shown in Figure 5-7, the shape of the triangle changes when the pose changes and becomes an isosceles triangle when the pose is frontal i.e. yaw is zero degree. In order to estimate the pose from a triangular shape, the following steps are followed:

- Detect the centres of the eyes and mouth using the methods described in chapter 4 (see Figure 5-1).

- Rotate the triangle so that the line joining centres of the eyes become horizontal.

- Normalize the triangle i.e. resize to a standard size.
- Use d_{LM}/d_{RM} to estimate the pose. Where d_{LM} and d_{RM} are the distances between left eye and mouth, and right eye and mouth in the normalized triangle respectively as shown in Figure 5-4.

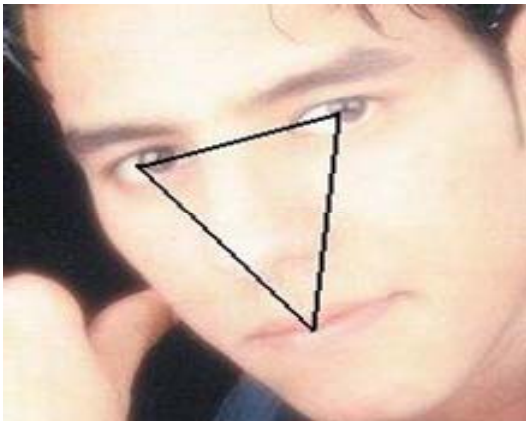


Figure 5-1 Eyes-Mouth Triangle

The triangle rotation and normalization is discussed in detail in the following sections.

5.2.1 TRIANGLE ROTATION

Let (x_1,y_1), (x_2,y_2) and (x_3,y_3) be the vertices of the eyes-mouth triangle as shown in Figure 5-2. To rotate the triangle in such a way that the line joining the two eyes coordinates becomes horizontal, the rotation angle θ need to be calculated first which can be calculated as:

$$\theta = \text{atan}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \quad \textbf{5-1}$$

The eyes-mouth triangle is then rotated clockwise through angle θ using 2D affine rotation using the following equations;

$$x_i^{'} = x_i \cos\theta + y_i \sin\theta \quad \textbf{5-2}$$

$$y_i^{'} = -x_i \sin\theta + y_i \cos\theta \quad \textbf{5-3}$$

In matrix form

$$\begin{bmatrix} x_i' \\ y_i' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad \textbf{5-4}$$

where i=1,2,3 and (x_i', y_i') are the vertices of rotated triangle. Figure 5-3 shows this rotation operation.

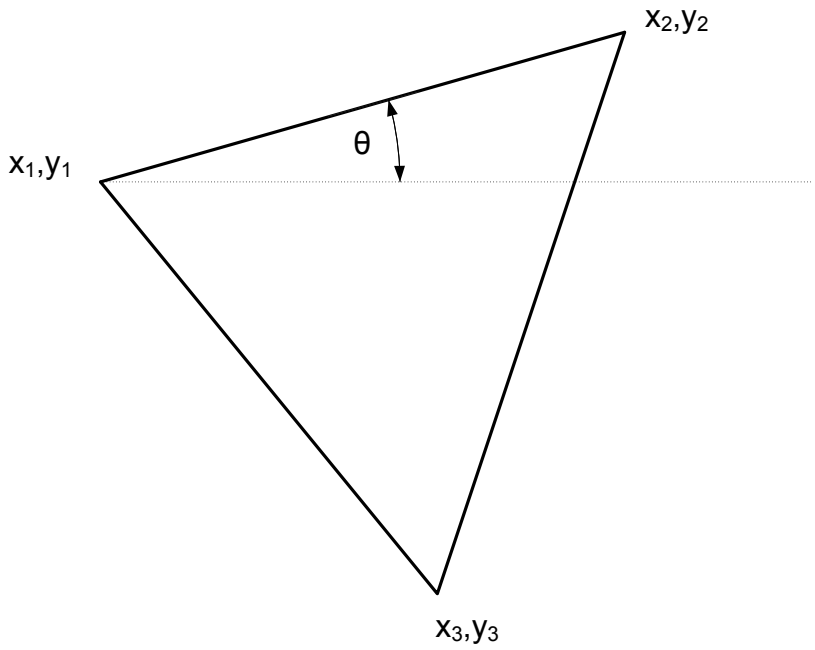


Figure 5-2 Eyes Mouth Coordinates

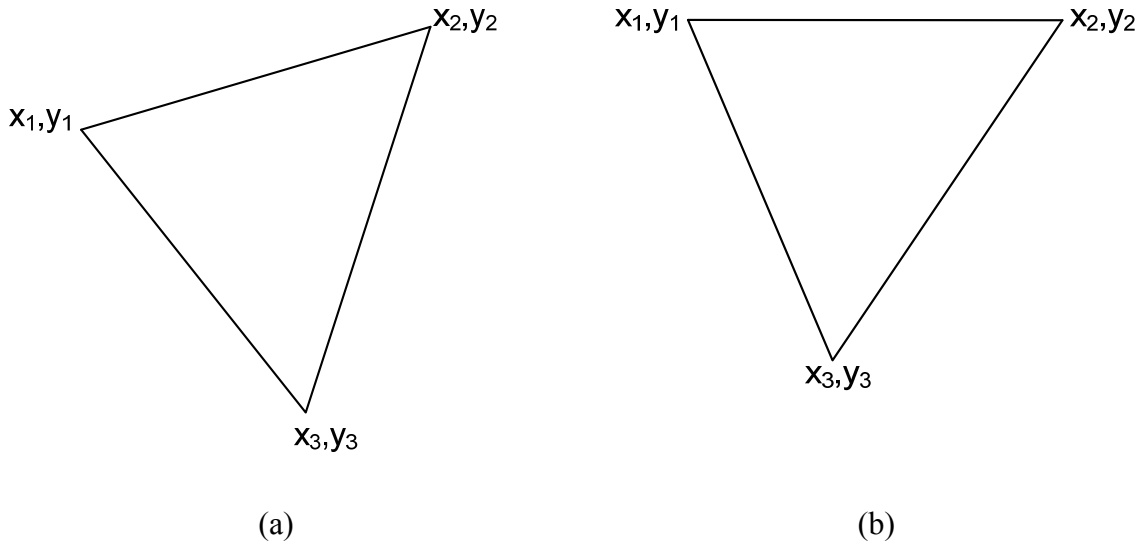


Figure 5-3 Triangle Rotation (a) Original Triangle (b) Rotated Triangle

5.2.2 TRIANGLE NORMALIZATION

Let $(x_1^{'}, y_1^{'})$, $(x_2^{'}, y_2^{'})$ and $(x_3^{'}, y_3^{'})$ be the vertices of rotated triangle obtained in the previous step. This triangle is resized to a fixed width W and height H, using the following equations:

$$x_{\min} = \min(x_1^{'}, x_2^{'}, x_3^{'}) \quad \text{5-5}$$

$$x_{\max} = \max(x_1^{'}, x_2^{'}, x_3^{'}) \quad \text{5-6}$$

$$y_{\min} = \min(y_1^{'}, y_2^{'}, y_3^{'}) \quad \text{5-7}$$

$$y_{\max} = \max(y_1^{'}, y_2^{'}, y_3^{'}) \quad \text{5-8}$$

$$x_{Final\_i} = (x_i^{'} - x_{\min}) \times (\frac{W}{x_{\max} - x_{\min}}) \quad \text{5-9}$$

$$y_{Final\_i} = (y_i^{'} - y_{\min}) \times (\frac{H}{y_{\max} - y_{\min}}) \quad \text{5-10}$$

where i=1,2,3 and (x_{Final_i}, y_{Final_i}) are the vertices of final resized triangle as shown in Figure 5-4.

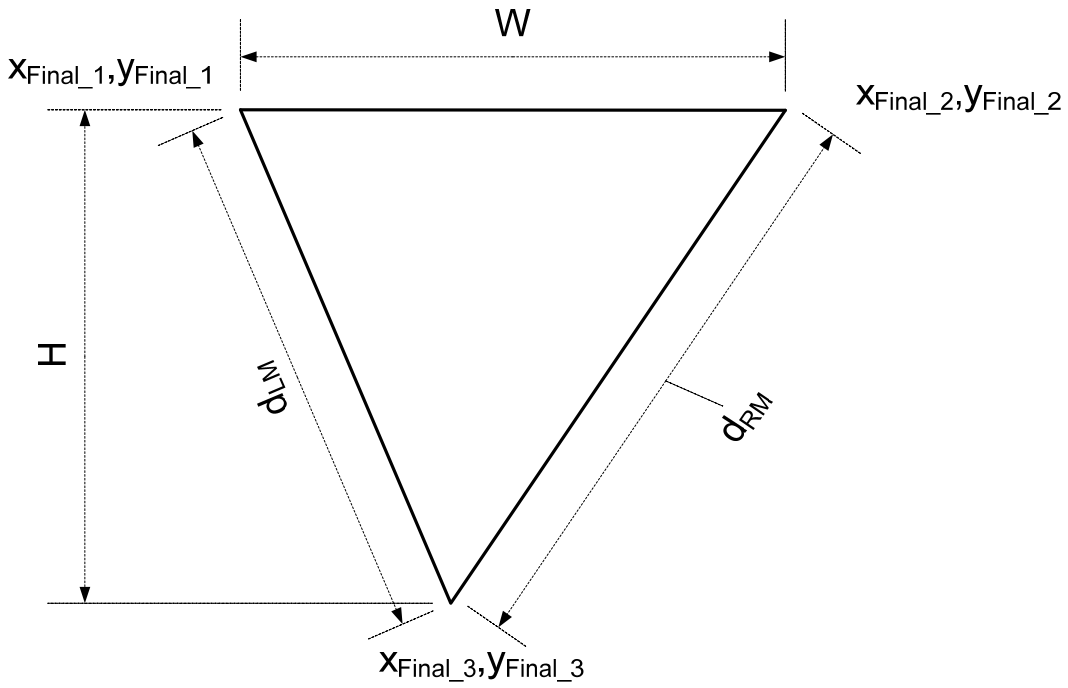


Figure 5-4 Sample Rotated and Normalized Triangle

Figures 5-5, 5-6 and 5-7 show the original triangles, rotated triangles and normalized triangles respectively for a series of images where the Yaw angle changes from -60 to 60 degrees with an interval of 15 degrees. The size of circles on the vertices represents the relative yaw angle. Figure 5-8 shows sample normalized triangles with different yaw angles.

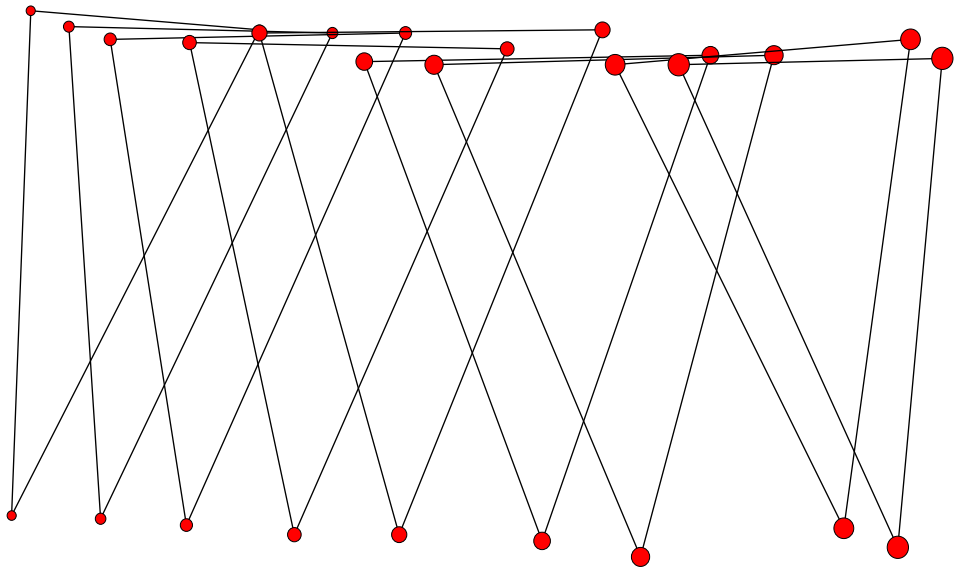


Figure 5-5 Original Positions of Eyes and Mouth

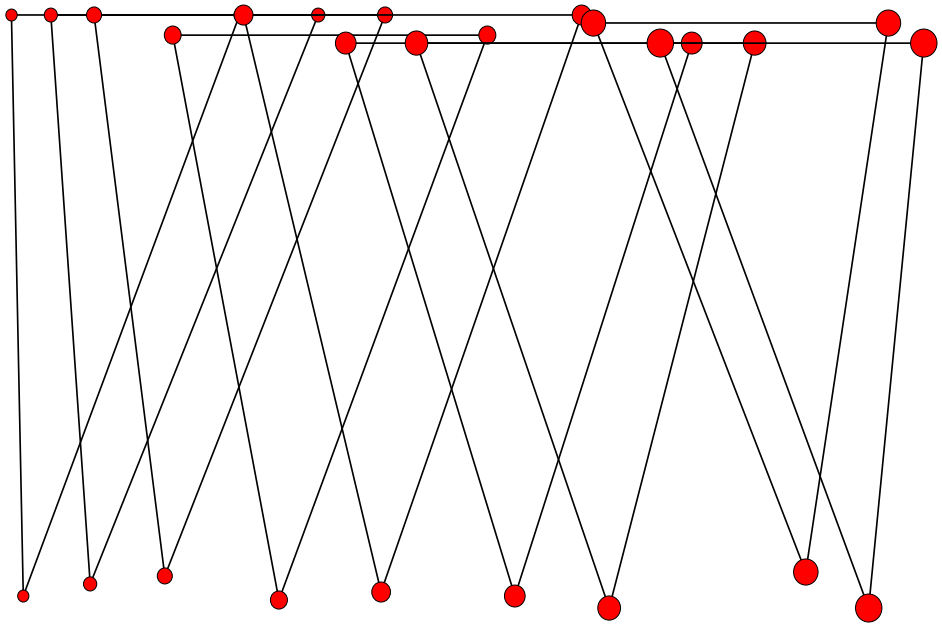


Figure 5-6 Rotated Triangles

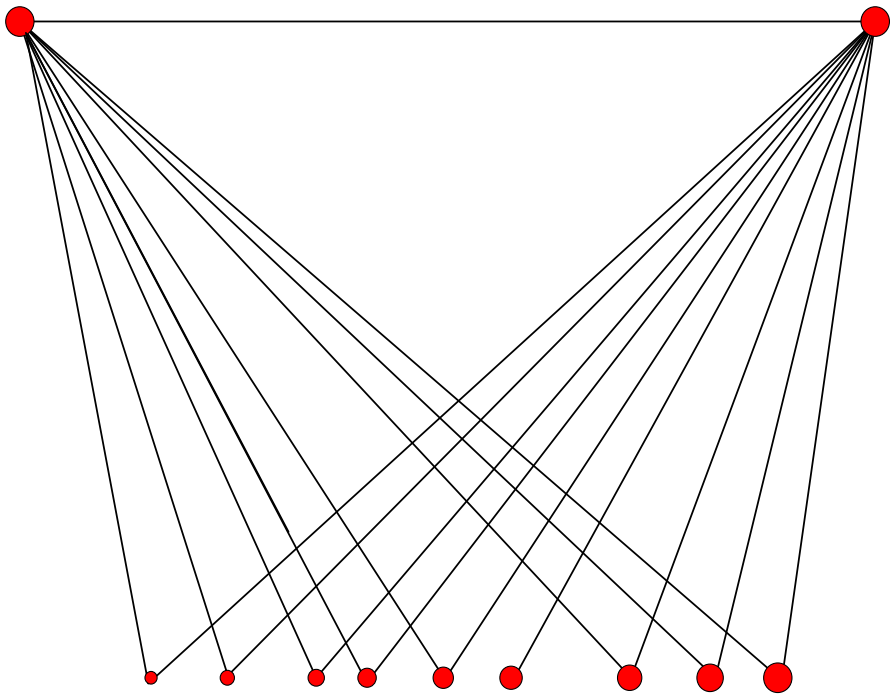


Figure 5-7 Normalized Triangles

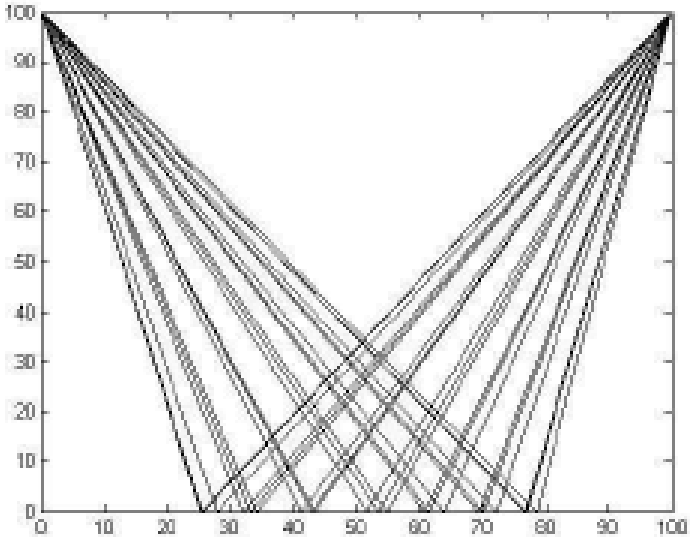


Figure 5-8 Sample Normalized Triangles

5.3 EXPERIMENTAL RESULTS AND ANALYSIS

The proposed method was tested using the following databases: Pointing ’04 (Gourier et al. (2004)), CAS-PEAL (Gao et al. 2008) and an in-house database discussed in detail in appendix A. Since the Pointing ’04 database contains badly illuminated and low resolution images, therefore, eyes and mouth centres were selected manually. The testing was carried out on images of all the 15 subjects in the dataset with 9 discrete yaws (-60° to 60°) and 3 discrete pitch values (-15°, 0° and 15°). From the CAS-PEAL database, images of first 50 subjects were selected with 7 discrete yaws (-67.5° to 67.5°) and one pitch value, i.e. 0°. As the facial feature methods presented in chapter 4 work for colour images only and CAS-PEAL comprises gray-level images, so eyes and mouth centres were selected manually. Images of 10 subjects with 9 yaws (-60° to 60°) and 3 pitch values (-15°, 0° and 15°) were selected from the in-house pose database for testing. Eyes and mouth centres were detected using the methods discussed in chapter 4. Figures 5-9 and 5-10 show sample images from Pointing ’04 and CAS-PEAL databases respectively.

Table 5-1 shows the experimental results for images from all the three databases. The accuracy shown is for pose estimation only and does not account for errors in feature points detection (i.e. the eyes and mouth centres were corrected manually in case they were detected inaccurately by the feature detection methods). The classification was considered accurate if correct yaw label was assigned to the input image. The accuracy is relatively lower for the in-house database. This is because of the errors in image capturing for the database. As directional suggestion (discussed in appendix A) was used, which is not a very accurate method.

Table 5-2 compares the proposed method with an existing pose estimation method i.e. Ma et al. (2006) on CAS-PEAL dataset. Ma et al. (2006) have reported the maximum face pose estimation accuracy on PEAL-CAS dataset. The comparison was made with method presented by Ma et al. (2006) because with the pose estimation range of the proposed method, their method is the most accurate in literature.

Figure 5-9 Sample Images from Pointing' 04 Database

Figure 5-10 Sample Images from CAS-PEAL Database

Table 5-1 Experimental Results

Data Set	Total Images	# of discrete Yaw values	# of discrete Pitch Values	Classification Accuracy
Pointing '04	405	9 (from -60° to 60° with 15° interval)	3 (-15°,0°,15°)	98.76%
CAS-PEAL	350	7 (from -67.5° to 67.5° with interval of 22.5°)	1 (0°)	99.14%
In-house Database (discussed in appendix A)	270	9 (from -60° to 60° with 15° interval)	3 (-15°,0°,15°)	94.44%

Table 5-2 Comparison with Existing Method

Dataset	# of Discrete Yaws	Accuracy	
		LGBP (Ma et al. (2006))	Proposed Method
CAS-PEAL	7	97.14%	99.14%

5.4 SUMMARY AND CONCLUSIONS

This chapter presented a novel geometric-based face pose estimation scheme which is based on the eyes-mouth triangle. According to the proposed methods, facial features i.e. eyes and mouth, are first detected and then the triangle obtained is rotated and normalized. The ratio of the lengths between left eye and mouth and right eye and mouth, in this normalized triangle is then used to estimate the pose. The method uses only eyes and mouth which are the most salient features of face. Also, it uses centres instead of corners of eyes and mouth which increases the working range of the scheme. It should be noted here that although the eyes and mouth centres calculation are based on corners, however, here the corners mean the visible extremes and not the actual corners of the features. Unlike other methods in the literature, the proposed method uses only three points for pose estimation. The use of only three points makes the method less expensive computationally and simplifies the equations used to estimate the pose.

6 FACE POSE ESTIMATION USING DISTANCE TRANSFORM AND NORMALIZED CROSS-CORRELATION

6.1 INTRODUCTION

This chapter discusses in detail an appearance template based method for face pose estimation. The proposed method uses distance transform instead of intensity images or edge-map for template matching. For a binary 2-D image with "object" and "background" pixels, the distance transform is calculated by assigning the distance to the nearest "object" pixel to each pixel. In our case, we first convert grey level image to edge-map and then calculate distance transform while treating edge pixels as object pixels. Figure 6-1 shows the schematic representation of distance transform for a face image. The length of the arrows shows the magnitude which is assigned to each pixel while the arrow heads are drawn to show the direction of nearest object pixel. The use of distance transform has two main advantages: firstly, unlike intensity image, the distance transform is relatively invariant to intensity and illumination of image. Secondly, unlike edge-map the distance transform has a smoother distribution. The distance transform property of being invariant to intensity makes the proposed method suitable for different skin colour because the skin colour variation appears as intensity variation in grey level images. Also, since the distance transform is relatively invariant to illumination, the proposed method is suitable for different illumination conditions. The proposed method is also relatively invariant to facial hairs and whether the subject is wearing glasses.

A face pose distance transform template is first created for each of the discrete yaw angle. The proposed method has been tested extensively on CAS-PEAL pose database (Gao et al. (2008)) which contains seven different yaw angles. Therefore, seven templates were created. To estimate the pose for an input image, the image is first converted to grey level if it is in colour. Distance transform is calculated and then it is resized and normalized so that the facial feature locations coincide with the already created templates. The pose is then estimated using a normalized cross-correlation between this resized distance transform image and the set of templates.

To show that the method works for different skin colours, the method has also been tested using an in-house database which contains darker colour faces as compared to standard pose databases (discussed in appendix A). The rest of the chapter is organized as follow: section 6.2 gives details of the pose templates, section 6.3 discusses the template matching and pose estimation using normalized cross correlation, section 6.4 presents the experimental results and analysis, and section 6.5 concludes the chapter.

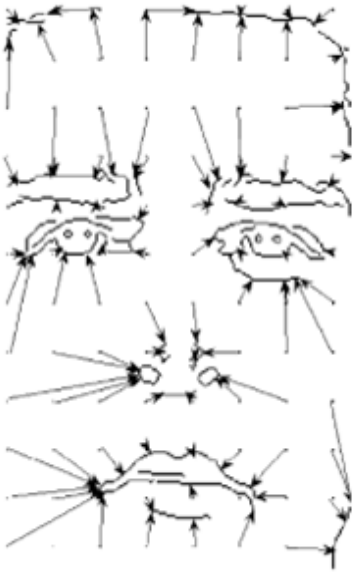


Figure 6-1 Schematic Representation of the Distance Transform of a Face, adapted from Asteriadis et al. (2009)

6.2 DISTANCE TRANSFORM POSE TEMPLATES

Since the CAS-PEAL pose database (Gao et al. (2008)) contains images with seven different yaw angles, seven templates were created. Grey level images with pitch angle of zero degrees of the first ten persons in CAS-PEAL database were used to compute templates for each yaw angle. In order to compute the distance transform template, the face area was manually cropped in the images. A Gaussian smoothing filter was applied to these images to remove the noise. Edges were detected using canny edge detector with a threshold value of 0.3 which gives the best results in our experiments. Canny edge detector with a threshold value is used because it ensures that only strong edges (eyes and mouth edges) are selected. Canny edge detector performed better as compared to sobel and prewitt operators for eyes and mouth edges detection in our experiments. Distance transform was then calculated for each of the images. The distance transform images were

then rotated so that the line joining the two eyes becomes horizontal. This rotation ensures that each of the images has a zero degree roll angle. All of these rotated distance transform images were then normalized so that the eyes and mouth centres have the same xy positions. All these images were then resized to a standard size of 265 by 265 pixels. Averages of these images were then taken for each of the discrete yaw value which resulted in a pose template. Histogram equalization was then applied to all the templates to increase the contrast. Only an elliptical facial region was then cropped in these templates which resulted in the final seven distance transform pose templates. Figure 6-2 and 6-3 show the process of template creation and the final templates respectively. The following is the pseudo code of templates computation algorithm.

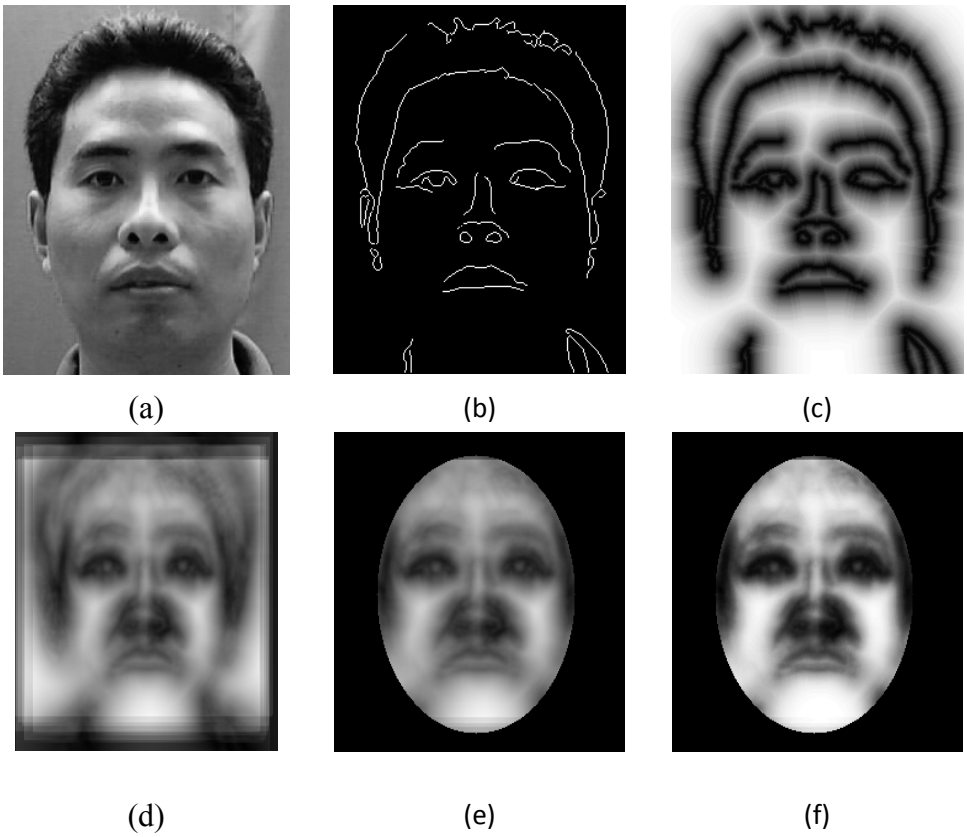


Figure 6-2 Distance Transform Templates Computation (a) Original Image (b) Edge Map (c) Distance Transform (d) Average Image (e) Cropped Image (f) Histogram Equalized Template

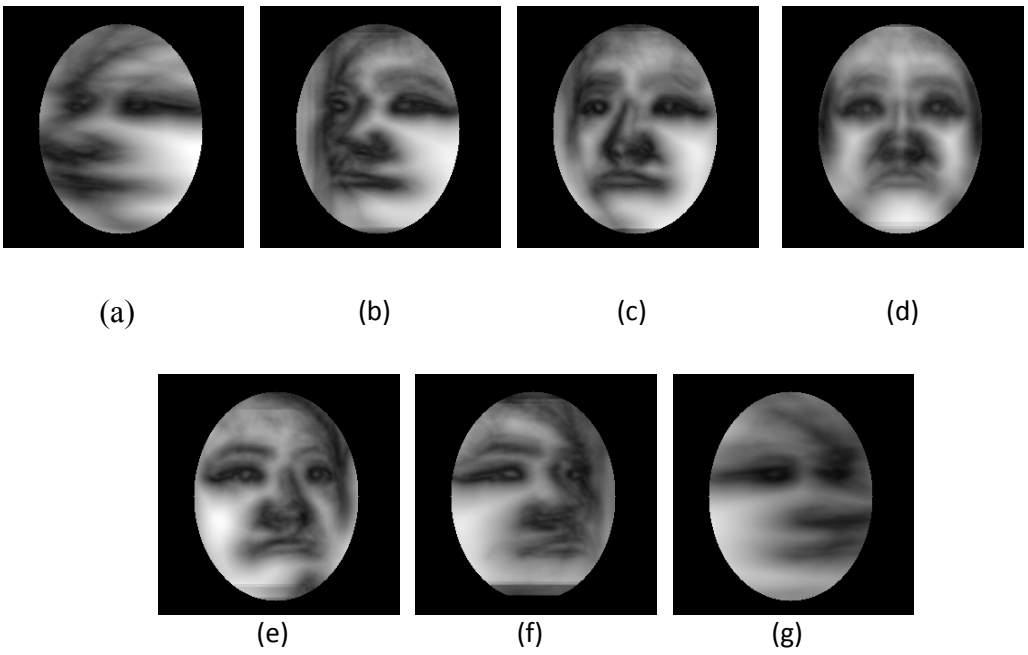


Figure 6-3 Distance Transform Templates with yaw angle of (a) -67° (b) -45° (c) -22.5° (d) 0° (e) 22.5° (f) 45° (g) 67°

Input: CAS-PEAL database pose images

Output: pose templates

Start

For each yaw angle

Select images of first ten persons from CAS-PEAL database

For each image

Crop face area

Remove noise using Gaussian filter

Detect edges using canny edge detector

Calculate distance transform

Rotate the distance image so that roll angle becomes zero

Resize and Normalize

End for

Calculate average of all ten images obtained in the for loop

Apply histogram equalization

Crop an elliptical face area to obtain final template

End for

End

6.3 TEMPLATE MATCHING USING NORMALIZED CROSS-CORRELATION

Normalized cross correlation was used to match an input image to the set of distance transform pose templates. A face was detected using the adaboost face detector by Viola and Jones (2004) and the face region was cropped. Coarse location of eyes and mouth were then extracted. These coarse eyes and mouth locations ensure that the image overlaps properly with the template in the process of normalized cross correlation calculation. Coarse eyes and mouth locations for colour images can be detected using the methods discussed in chapter 4. However, the CAS-PEAL database contains only grey level images, therefore, of the method by Hsu et al. (2002) was slightly modified and used. For coarse eyes localization in gray level images, the eyemap was calculated by dividing the eroded image by the dilated image. A threshold was applied to the eyemap to obtain the coarse eye locations. The lower half of the face image was kept as the target for the mouth. Adaptive thresholding was then applied, starting with a threshold value of 1. The threshold value was decreased in steps and the region that changes to black first, was considered as the mouth since the mouth is darker colour than its near regions in a grey level image. To calculate the distance transform of the face region, edges were extracted using canny edge detector. A threshold value of 0.3 was used in the canny edge detection which gave better results based in our experiments. Distance transform was then calculated from the edge-map. The distance transform image obtained was rotated so that the line joining the two eyes became horizontal. This rotation minimizes the effect of roll angle variation. The image was then resized based on the coarse locations of the eyes and mouth already obtained so that the eyes and mouth locations coincided with that of the templates. This resizing operation has two advantages: first it minimizes the pitch variation which enables the proposed method to be invariant to pitch variations. Secondly, it enables the image to overlap the templates properly during the process of correlation calculation. A 265 by 265

size image was cropped which was used to calculate normalized cross-correlation. Normalized cross correlation was calculated for each of the templates with this image and the template which gave the highest value was selected as the pose of the input image. Normalized cross correlation performs better than simple cross correlation in image-processing applications in which the brightness of the image and template can vary due to lighting and exposure conditions. Normalized cross correlation is calculated as:

$$r = \frac{\langle TI \rangle - \langle T \rangle \langle I \rangle}{\sigma(T)\sigma(I)} \quad \text{6-1}$$

where I is the input image, T is template, TI is the image obtained by multiplying I with T, <> is mean operator. σ is the standard deviation and is calculated as:

$$\sigma = \sqrt{E[(X - \mu)^2]} \quad \text{6-2}$$

where μ is the mean value and E is the expected value. Figure 6-4 shows the flow diagram of the proposed method.

6.4 EXPERIMENTAL RESULTS AND ANALYSIS

The proposed method was tested using the CAS-PEAL pose database which contains images of 1042 persons with 21 discrete poses for each subject. The yaw angle varies from -67° to 67° with interval of 22.5° while the pitch varies from -30° to 30° with interval of 30°. Nine cameras were placed, each with 15 degree difference to capture the yaw angles. The camera setup is shown in Figure 6-5. Apart from the pose images with normal conditions, i.e. normal lighting, neutral expressions etc, there are also some frontal images with different illumination conditions and facial expressions. Figure 6-6 shows how 20 different illumination conditions were generated in CAS-PEAL database by placing the lamps in specific positions.

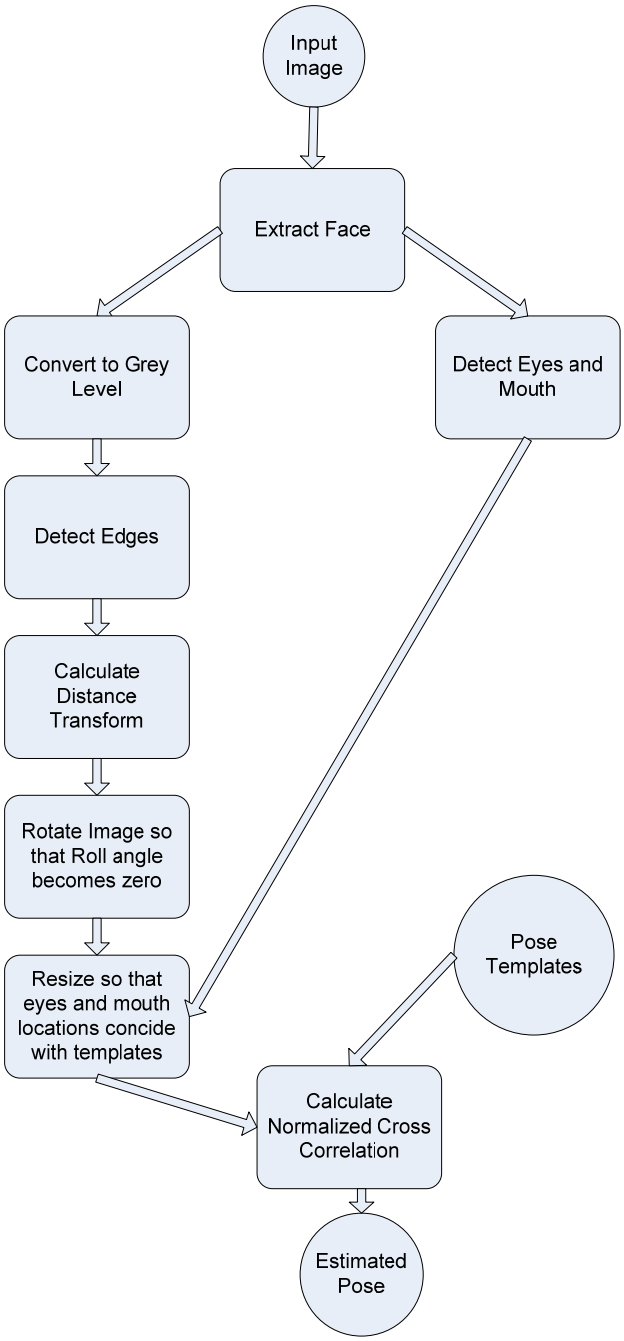


Figure 6-4 Flow Diagram

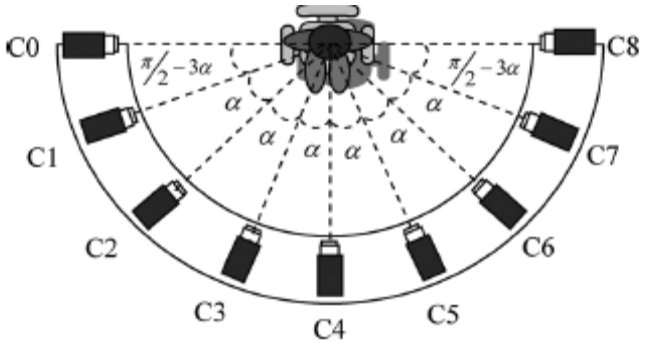


Figure 6-5 Cameras Positions in CAS-PEAL Database Gao et al. (2008)

The proposed method was tested by selecting a total of 1050 images of the first fifty persons from the CAS-PEAL database. The template which gave maximum normalized cross correlation value with the input image was assigned as the estimated pose. To test that the proposed scheme is invariant to lighting conditions, the method was tested on a total of 100 images of 5 persons with twenty different illumination conditions. For all five persons, the method worked well for 12 out of 20 illumination conditions. The 12 illumination conditions are the ones for which the light falls almost equally on the left and right half of the face even if the face is badly illuminated. Figure 6-7 shows an example of various illumination conditions from the CAS-PEAL database. The tick sign shows the illumination conditions for which the proposed method works, while the cross sign shows the lighting conditions for which the proposed methods fails either completely or partially.

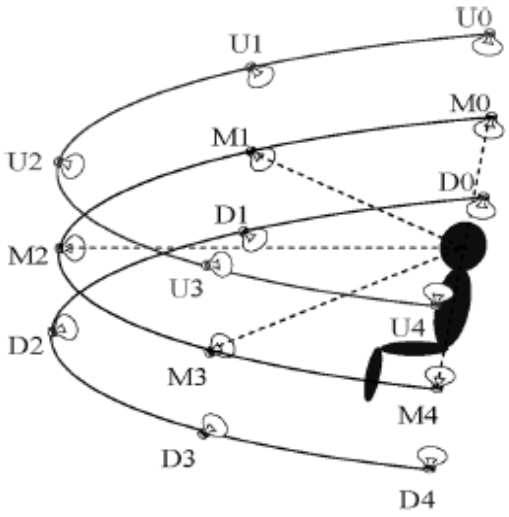


Figure 6-6 Light Positions for CAS-PEAL Database Gao et al. (2008)

Figure 6-7 Different Illumination Conditions in CAS-PEAL Database

Table 6-1 shows experimental results of the proposed scheme for the CAS-PEAL face pose database.

Table 6-1 Experimental Results for CAS-PEAL Pose Database

Pitch	Total images	Correct	Incorrect	%age accuracy
-30°	350	283	67	80.85%
0°	350	342	08	97.71%
30°	350	287	63	82.00%
Total	1050	912	138	86.85%

Table 6-2 shows the comparison of the proposed method with the LGBP pose estimation schemes by Ma et al. (2006) and eyes-mouth triangle (chapter 5). To the best of the author's knowledge, their method has the best reported results in the literature for the CAS-PEAL database. The LGBP based method has been tested only on images with zero pitch value that is why the comparison has been made only for those images. Although the eyes-mouth triangle method (discussed in chapter 5) has more accuracy, the proposed method is relatively invariant to skin colour variations, bad lighting condition, facial hairs, wearing glasses and exact localization of facial features.

Table 6-2 Comparison with Existing Methods

Dataset	# of Discrete Yaws	Accuracy		
		LGBP (Ma et al. (2006))	Eyes-Mouth Triangle Method (chapter 5)	Proposed Method
CAS-PEAL	7	97.14%	99.14%	97.71%

Figure 6-8 shows the percentage accuracy for different yaw angles at zero degree pitch and Figure 6-9 shows how the value of normalized cross correlation varies for different images.

To test the robustness for different skin colours and for subjects with facial hairs or glasses, the proposed method was tested using 90 images from an in-house pose database (discussed in detail in appendix A). The 90 images are of 10 subjects all having skin colours darker than that of standard pose databases, 3 having facial hairs and one wearing glasses, where yaw varies from -60 to +60 degrees. The poses were accurately estimated in 88 images with a percentage accuracy of 97.77%. An estimation was considered accurate if the right pose label was assigned to the input image. Interestingly, the accuracy of the proposed algorithm is higher for the in-house database (which contains images with glasses and facial hairs) than the CAS-PEAL database (which contains images without any glasses or facial hairs). Since the in-house database comprises images of good resolution and better lighting conditions, which means that distance transform will be computed with high accuracy and that is the reason that the proposed algorithm gives better results for the in-house database.

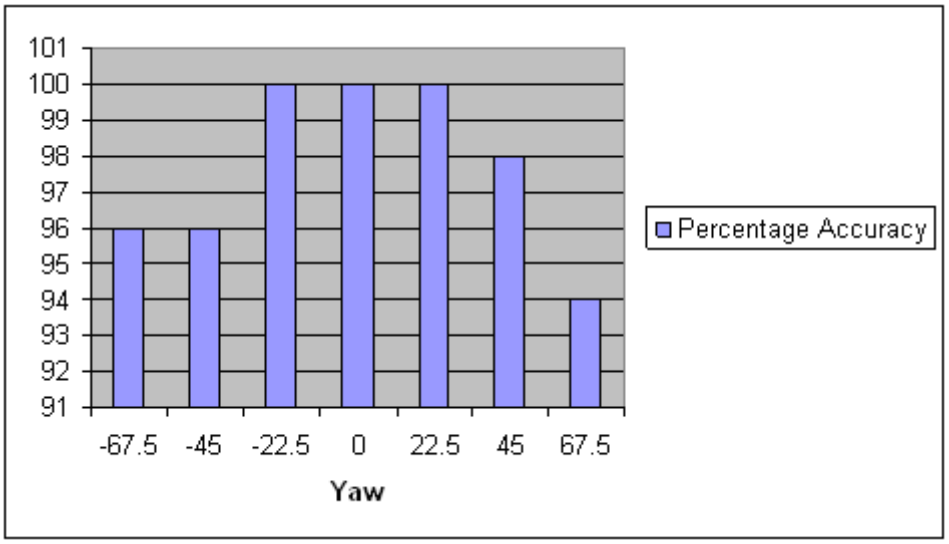


Figure 6-8 Accuracy for Different Yaw Angles

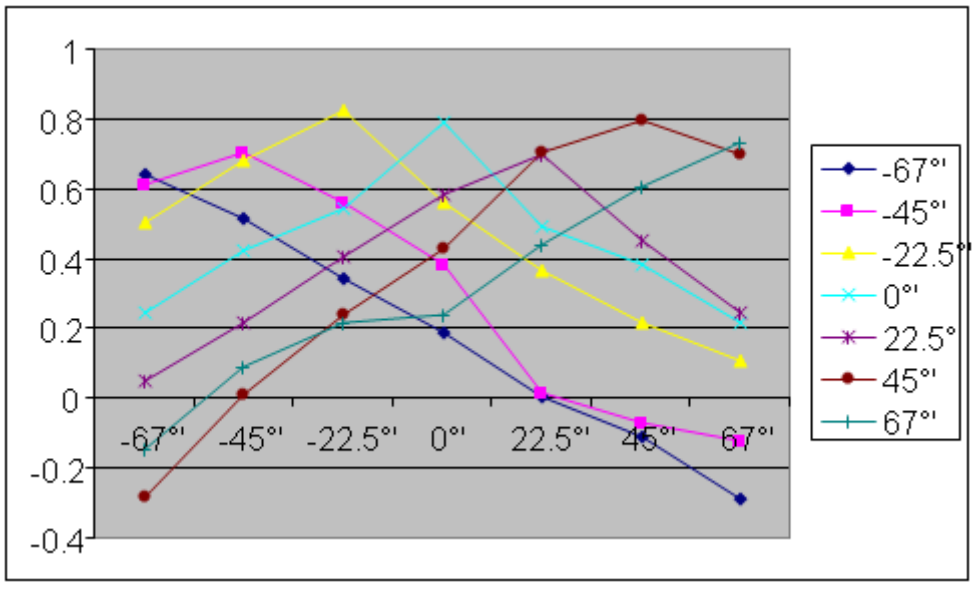


Figure 6-9 Typical Normalized Cross-Correlation Values, where yaw angles of input images are shown on vertical axis and yaw angles of templates are shown on horizontal axis

6.5 SUMMARY AND CONCLUSIONS

This chapter presented a novel appearance template based method for yaw estimation in near-field images. The method is based on distance transform and template matching using normalized cross correlation. A set of seven distance transform templates each for a discrete yaw value was created. The input image is converted to distance transform and is then compared with the set of templates using normalized cross correlation to estimate the pose. The distance transform property of being invariant to lighting and illumination conditions makes the method applicable to different skin colours and illumination conditions. The method was tested on CAS-PAEL pose database and the results show that it is very robust. The robustness is further demonstrated using images from an in-house database of subjects with dark skin colours, and some with facial hairs or glasses. Badly illuminated images from the in-house database and from the CAS-PEAL pose database were also tested with very good results.

7 CONCLUSIONS AND FUTURE WORK

7.1 INTRODUCTION

The main theme in this thesis has been improvements to existing face pose (specifically Yaw) estimation in terms of processing time, and invariance to different conditions e.g. bad illumination, different skin colours, facial hairs, and presence of glasses, etc. Since facial features localization is generally used as an intermediate step in face pose estimation, it has also been considered as a subtopic in this thesis. This chapter summarizes the proposed algorithms presented in chapters 4, 5 and 6, and highlights the key contributions made by the research presented in this thesis. It also discusses the future research directions, particularly with the intention of further extending the functionality and efficiency of the proposed algorithms.

7.2 SUMMARY OF CONTRIBUTIONS

This thesis presented four original contributions which are summarized below.

7.2.1 USE OF EDGE-DENSITY FOR EYES LOCALIZATION IN FACIAL IMAGES

The novel edge-density based eyes localization method (discussed in chapter 4) is based on the observation that edge-density is high in eye regions as compared to other regions in a face. Most colour variations occur in eye regions in facial images are due to colour difference between eyelids and skin, skin and sclera, sclera and iris, and iris and pupil. The colours of the rest of the face are more uniform when compared to the eyes. Therefore, if edge detection was applied to facial images then the eye regions would have maximum edge densities. After edge detection, morphological erosion and dilation are applied to obtain the candidate eyes regions. Shape and geometry information is then used to extract and verify the eyes. The proposed scheme was tested on PICS (http://pics.psych.stir.ac.uk/) images database with very good results.

7.2.2 USE OF A COMBINATION OF COLOUR, EDGE AND ILLUMINATION CUES FOR EYES LOCALIZATION IN FACIAL IMAGES

The proposed hybrid method (discussed in chapter 4) combines intensity, edge and illumination cues to form a more accurate eyes localization system. The system was tested

using the PICS (http://pics.psych.stir.ac.uk/) image database. The result demonstrates that the proposed method overcomes the weaknesses of each of the individual methods. .

7.2.3 USES OF EYES-MOUTH TRIANGLE SHAPE FOR FACE POSE ESTIMATION

This proposed method (discussed in chapter 5) is based on the observation that eyes-mouth triangle has a distinct shape for distinct yaw angle. The method uses the shape of eyes-mouth triangle to estimate face pose. Eyes and mouth centres are localized first using the methods discussed in chapter 4. To minimize inter-subject variations, the triangle obtained from the eyes and mouth centres is rotated and normalized, i.e. resized to a standard size. The shape of this normalized triangle is then used to estimate face pose. The proposed method has a number of advantages over existing methods. It is based on eyes and mouth which are the most salient features of a face. It is based on only three feature points, instead of five, which makes it faster computationally and hence suitable for real-time applications. Because it uses centres instead of corners of eyes and mouth, the range of pose estimation is increased. The proposed method was tested using Pointing '04 (Gourier et al. 2004), CAS-PEAL (Gao et al. 2008) and an in-house (discussed in detail in Appendices) databases with very good results.

7.2.4 USE OF DISTANCE TRANSFORM AND NORMALIZED CROSS-CORRELATION FOR POSE ESTIMATION IN NEAR-FIELD IMAGES

The proposed method (discussed in chapter 6) uses distance transform and normalized cross-correlation for face pose estimation. Distance transform face pose template is first computed for each discrete yaw angle. To estimate the pose for an input image, the image is first converted to grey level if it is in colour. Distance transform is calculated and then it is resized and normalized so that the facial feature locations coincide with the already created templates. The pose is then estimated using a normalized cross-correlation between this resized distance transform image and the set of templates. The distance transform property of being invariant to intensity makes the proposed method suitable for different skin colour because the skin colour variation appears as intensity variation in grey level images. Also, since the distance transform is relatively invariant to illumination, the proposed method is suitable for different illumination conditions. The proposed method is also relatively invariant to facial hairs and whether the subject is wearing glasses. The proposed method was tested using the CAS-PEAL pose database (Gao et al. (2008)). To show that the method works for different skin colours, the method was also tested using an

in-house database which contains darker colour faces as compared to other publically available pose databases.

7.3 FUTURE WORK

Although a number of novel contributions in face pose estimation were presented in this thesis, there is always room for improvements. The following are some of the future research directions.

All the implementations and testing of the proposed algorithms have been carried out in Matlab. One desirable task is to implement the proposed algorithms in some lower level language such as C or C++. This will increase the execution speed.

The face pose templates computed in chapter 6 can be further refined by correcting the detected edges manually during the template creation process. More accurate edges mean more accurate distance transform and hence more accurate templates. It is likely that these refinements to face pose templates will result in more accurate face pose estimation method.

The face pose estimation algorithm proposed in chapter 6 uses distance transform. Distance transform contains only magnitude information while Distance Vector Fields (a variant of distance transform) contain magnitude as well as direction information. Thus the use of distance vector fields instead of simple distance transform is likely to increase accuracy of the proposed algorithm.

Almost all the standard face pose databases contain images of light skin colour subjects. Although, the newly created in-house database (discussed in Appendix A) contains images with darker skin colour subjects as compared to standard databases, there is still a need of including more subjects with various skin colours (particularly darker skin colour).

Main focus of this thesis was face pose estimation in still images. More advanced face pose estimation and tracking in videos can be designed by combining the proposed face pose estimation with pose tracking techniques. The proposed methods can be used to estimate the face pose in the first frame of the video and tracking can be used to track the estimated pose in the following frames.

REFERENCES

Asteriadis, S., Nikolaidis, N. & Pitas, I. 2009, "Facial feature detection using distance vector fields", *Pattern Recognition,* vol. 42, no. 7, pp. 1388-1398.

Ba, S. & Odobez, J.M. 2004, "A probabilistic framework for joint head tracking and pose estimation", *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004,* pp. 264-267.

Balasubramanian, V.N., Ye, J. & Panchanathan, S. 2007, "Biased manifold embedding: a framework for person-independent head pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pp. 1-7.

Belkin, M. & Niyogi, P. 2003, "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural computation,* vol. 15, no. 6, pp. 1373-1396.

Beymer, D.J. 1993, "Face recognition under varying pose", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. CVPR '94,* pp. 756-761.

Bhuiyan, M.A., Ampornaramveth, V., Muto, S. & Ueno, H. 2003, "Face Detection and Facial Feature Localization for Human-machine Interface", *Natl Inst Inform,* vol. 5, pp. 25-39.

Bishop, C.M. 2005, *Neural networks for pattern recognition,* Oxford University Press, 2005.

Bowyer, D. & Sarkar, S. 2001, *USF HumanID 3D face dataset*.

Brown, L.M. & Tian, Y.L. 2002, "Comparative study of coarse head pose estimation", *Proceedings of IEEE Workshop on Motion and Video Computing, Orlando FL,* pp. 125-130.

Bruske, J., Abraham-Mumm, E., Pauli, J. & Sommer, G. 1998, "Head-pose estimation from facial images with subspace neural networks", *Proceedings of International Neural Network and Brain Conference*, pp. 528–531.

Canton-Ferrer, C., Casas, J.R. & Pardas, M. 2008, "Head Orientation Estimation Using Particle Filtering in Multiview Scenarios", *Multimodal Technologies for Perception of Humans: International Evaluation Workshops Clear 2007 and Rt 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers,* , pp. 317-327.

Canton-Ferrer, C., Casas, J.R. & Pardas, M. 2007, "Head pose detection based on fusion of multiple viewpoint information", *Lecture Notes in Computer Science,* vol. 4122, pp. 305-310.

Cascia, M.L., Sclaroff, S. & Athitsos, V. 2000, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3 D models", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 4, pp. 322-336.

Chan, S.C.Y. & Lewis, P.H. 1999, "A pre-filter enabling fast frontal face detection", *Lecture notes in computer science,* vol. 1614, pp. 781-789.

Chen, L., Zhang, L., Hu, Y., Li, M. & Zhang, H. 2003, "Head pose estimation using fisher manifold learning", *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures,* IEEE Computer Society Washington, DC, USA, pp. 203-207.

Chiang, C.C., Tai, W.K., Yang, M.T., Huang, Y.T. & Huang, C.J. 2003, "A novel method for detecting lips, eyes and faces in real time", *Real-Time Imaging,* vol. 9, no. 4, pp. 277-287.

Cootes, T.F., Edwards, G.J. & Taylor, C.J. 2001, "Active appearance models", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 6, pp. 681-685.

Cootes, T.F., Taylor, C.J., Cooper, D.H. & Graham, J. 1995, "Active shape models-their training and application", *Computer Vision and Image Understanding,* vol. 61, no. 1, pp. 38-59.

Cootes, T.F., Wheeler, G.V., Walker, K.N. & Taylor, C.J. 2002, "View-based active appearance models", *Image and Vision Computing,* vol. 20, no. 9-10, pp. 657-664.

Cordea, M.D., Petriu, E.M., Georganos, N.D., Petriu, D.C. & Whalen, T.E. 2001, "Real-time 2 (1/2)-D head pose recovery for model-based video-coding", *IEEE Transactions on Instrumentation and Measurement,* vol. 50, no. 4, pp. 1007-1013.

Cristinacce, D., Cootes, T. & Scott, I. 2004, "A multi-stage approach to facial feature detection", *Proceedings of the 15 th British Machine Vision Conference, London, England*, pp. 277–286.

Duda, R.O., Hart, P.E. & Stork, D.G. 2001 *Pattern classification,* 2nd edition, John Wiley & Sons, Inc, 2001.

Ebisawa, Y 2008, "Head Pose Detection with One Camera Based on Pupil and Nostril Detection Technique", *Proceedings of the IEEE International Conference on Virtual Envirements, Human Computer Interfaces, and Measurements Systems,* pp. 172-177.

Ebisawa, Y. & Satoh, S.I. 1993, "Effectiveness of pupil area detection technique using two light sources and image difference method", *Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1993*, pp. 1268-1269.

Edwards, G.J., Lanitis, A., Taylor, C.J. & Cootes, T.F. 1998, "Statistical models of face images—Improving specificity", *Image and Vision Computing,* vol. 16, no. 3, pp. 203-211.

Everingham, M. & Zisserman, A. 2006, "Regression and classification approaches to eye localization in face images", Proceedings of *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006*, pp. 441-446.

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X. & Zhao, D. 2008, "The CAS-PEAL large-scale Chinese face database and baseline evaluations", *IEEE Transactions on Systems Man and Cybernetics Part A Systems and Humans,* vol. 38, no. 1, pp. 149-161.

Gee, A.H. & Cipolla, R. 1994, "Determining the gaze of faces in images", *Image and Vision Computing,* vol. 12, no. 10, pp. 639-647.

Gonzalez, R. & Woods, R. 2002, "Digital Image Processing" in , 2nd edn, Prentice-Hall, Inc., pp. 582-584.

Gordon, G.G. 1998, "3D pose estimation of the face from video", *NATO ASI Series F Computer and Systems Sciences,* vol. 163, pp. 433-445.

Gourier, N., Hall, D. & Crowley, J.L. 2004a, "Estimating face orientation from robust detection of salient facial structures", *Proceedings of the Pointing 2004 Workshop on Visual Observation of Deictic Gestures*, pp. 17–25.

Gourier, N., Hall, D. & Crowley, J.L. 2004b, "Facial features detection robust to pose, illumination and identity", *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, pp. 617-622.

Gourier, N., Maisonnasse, J., Hall, D. & Crowley, J.L. 2007, "Head pose estimation on low resolution images", *Lecture Notes in Computer Science,* vol. 4122, pp. 270-280.

Hamouz, M., Kittler, J., Kamarainen, J.K., Paalanen, P., Kälviäinen, H. & Matas, J. 2005, "Feature-based affine-invariant localization of faces", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 9, pp. 1490-1495.

Han, C.C., Liao, H.Y.M., Yu, G.J. & Chen, L.H. 2000, "Fast face detection via morphology-based pre-processing", *Pattern Recognition,* vol. 33, no. 10, pp. 1701-1712.

Heinzmann, J. & Zelinsky, A. 1998, "3-D facial pose and gaze point estimation using a robust real-time tracking paradigm", *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 142-147.

Ho, S.Y. & Huang, H.L. 1998, "An analytic solution for the pose determination of human faces from a monocular image", *Pattern Recognition Letters,* vol. 19, no. 11, pp. 1045-1054.

Horprasert, T., Yacoob, Y. & Davis, L.S. 1997, "An anthropometric shape model for estimating head orientation", *Proceedings of the 3rd International Workshop on Visual Form*, *Capri, Italy*.

Horprasert, T., Yacoob, Y. & Davis, L.S. 1996, "Computing 3-D head orientation from a monocular image sequence", *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 242-247.

Hsu, R.L., Abdel-Mottaleb, M. & Jain, A.K. 2002, "Face detection in color images", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 5 , pp. 696-706.

Hu, Y., Chen, L., Zhou, Y. & Zhang, H. 2004, "Estimating face pose by facial asymmetry and geometry", *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR'04)*, pp. 651-656.

Hu, Y. & Huang, T. S. 2008, "Subspace Learning for Human Head Pose Estimation", *In Proceedings of the IEEE International Conference on Multimedia and Expo,* pp. 1585-1588.

Huang, J., Shao, X. & Wechsler, H. 1998, "Face pose discrimination using support vector machines (SVM)", *Proceedings of Fourteenth International Conference on Pattern Recognition, Brisbane, Qld., Australia,* pp. 154-156.

Huang, K.S. & Trivedi, M.M. 2004, "Robust real-time detection, tracking, and pose estimation of faces in video streams", *Proceedings of the 17th International Conference on Pattern Recognition 2004 (ICPR 2004),* vol. 3, pp. 965-968.

Jebara, T. & Pentland, A. 1997, "Parametrized structure from motion for 3d adaptive feedback tracking of faces", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 144-150.

Jesorsky, O., Kirchberg, K.J. & Frischholz, R.W. 2001, "Robust face detection using the hausdorff distance", *Lecture notes in computer science,* vol. 2091, pp. 90-95.

Krueger, V. & Sommer, G. 2002, "Gabor wavelet networks for efficient head pose estimation", *Image and Vision Computing,* vol. 20, no. 9-10, pp. 665-672.

Krüger, N., Pötzsch, M. & Von der Malsburg, C. 1997, "Determination of face position and pose with a learned representation based on labelled graphs", *Image and Vision Computing,* vol. 15, no. 8, pp. 665-673.

La Cascia, M., Sclaroff, S. & Athitsos, V. 2000, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3 D models", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 4, pp. 322-336.

Lablack, A., Zhongfei, Z. & Djeraba, C. 2008, "Supervised Learning for Head Pose Estimation Using SVD and Gabor Wavelets", *In Proceesings of the tenth IEEE International Symposium on Multimedia,* pp. 592-596.

Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., Malsburg, C.V.D., Wurtz, R.P. & Konen, W. 1993, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Transactions on Computers,* vol. 42, no. 3, pp. 300-311.

Lanitis, A., Taylor, C.J., Cootes, T. & Ahmed, T. 1995, "Automatic interpretation of human faces and hand gestures using flexible models", *Proceedings of the International Workshop on Automatic Face-and Gesture-Recognition,* pp. 98-103.

Lanitis, A., Taylor, C.J. & Cootes, T.F. 1997, "Automatic interpretation and coding of face images using flexiblemodels", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 743-756.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. 1998, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324.

Li, S.Z., Fu, Q., Gu, L., Scholkopf, B., Cheng, Y. & Zhang, H. 2001, "Kernel machine based learning for multi-view face detection and pose estimation", *Proceedings of 8th IEEE International Conference on Computer Vision*, vol. 2 , pp. 674-679.

Li, Y., Gong, S. & Liddell, H. 2000, "Support Vector Regression and Classification Based Multi-View Face Detection and Recognition", *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000,* IEEE Computer Society Washington, DC, USA, pp. 300-305.

Li, Y., Gong, S., Sherrah, J. & Liddell, H. 2004, "Support vector machine based multi-view face detection and recognition", *Image and Vision Computing,* vol. 22, no. 5, pp. 413-427.

Li, Z., Fu, Y., Yuan, J., Huang, T.S. & Wu, Y. 2007, "Query driven localized linear discriminant models for head pose estimation", *Proceedings of the IEEE Conference on Multimedia and Expo,* pp. 1810-1813.

Little, D., Krishna, S., Black, J. & Panchanathan, S. 2005, "A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05),* pp. 89-92.

Ma, B., Zhang, W., Shan, S., Chen, X. & Gao, W. 2006, "Robust head pose estimation using LGBP", *Proceedings of the 18th International Conference on Pattern Recognition, 2006 (ICPR 2006),* pp. 512-515.

Ma, Y., Ding, X., Wang, Z. & Wang, N. 2004, "Robust precise eye location under probabilistic framework", *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition,* pp. 339-344.

Ma, Y., Konishi, Y., Kinoshita, K., Lao, S. & Kawade, M. 2006, "Sparse bayesian regression for head pose estimation", *Proceedings of the 18th International Conference on Pattern Recognition, 2006 (ICPR 2006)* pp. 507-510.

Martinez, J. E., Erol, A., Bebis, G., Boyle, R. & Twombly, X. 2009, "Integrating perceptual Level of Details with Head Pose Estimation and its Uncertainty", *Machine Vision and Applications,* vol. 21, pp. 69-83.

McKenna, S.J. & Gong, S. 1998, "Real-time face pose estimation", *Real-Time Imaging,* vol. 4, no. 5, pp. 333-347.

Moon, H. & Miller, M.L. 2004, "Estimating Facial Pose from a Sparse Representation", *Proceedings of the IEEE International Conference on Image Processing*, pp. 75-80.

Morency, L. P., Whitehill, J. & Movellan, J. 2009, "Monocular Head Pose Estimation using Generalized Adaptive View-based Appearance Model", *Image and Vision Computing* (in press).

Morency, L.P., Rahimi, A. & Darrell, T. 2003, "Adaptive view-based appearance models", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 803-810.

Morimoto, C., Koons, D., Amir, A. & Flickner, M. 2000, "Pupil detection and tracking using multiple light sources", *Image and Vision Computing,* vol. 18, no. 4, pp. 331-335.

Murphy-Chutorian, E., Doshi, A. & Trivedi, M.M. 2007, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation", *Proceedings of the IEEE Intelligent Transportation Systems Conference, 2007,* pp. 709-714.

Murphy-Chutorian, E. & Trivedi, M.M. 2009, "Head Pose Estimation in Computer Vision: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, no. 4 , pp. 607-626.

Murphy-Chutorian, E. & Trivedi, M.M. 2008, "HyHOPE: Hybrid Head Orientation and Position Estimation for vision-based driver head tracking", *Proceedings of the IEEE Intelligent Vehicles Symposium 2008*, pp. 512-517.

Newman, R., Matsumoto, Y., Rougeaux, S. & Zelinsky, A. 2000, "Real-time stereo tracking for head pose and gaze estimation", *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition Gesture Recognition (FG2000),* pp. 122-128.

Ng, J. & Gong, S. 2002, "Composite support vector machines for detection of faces across views and pose estimation", *Image and Vision Computing,* vol. 20, no. 5-6, pp. 359-368.

Ng, J. & Gong, S. 1999, "Multi-view face detection and pose estimation using a compositesupport vector machine across the view sphere", *In Proceedings of International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999*, pp. 14-21.

Niyogi, S. & Freeman, W. 1996, "Example-based head tracking", *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996*, pp. 374-378.

Osadchy, M., Le Cun, Y. & Miller, M.L. 2007, "Synergistic face detection and pose estimation with energy-based models", *The Journal of Machine Learning Research,* vol. 8, pp. 1197-1215.

Perlibakas, V. 2003, "Automatical detection of face features and exact face contour", *Pattern Recognition Letters,* vol. 24, no. 16, pp. 2977-2985.

PICS, *The psychological image collection at Stirling*. Available: http://pics.psych.stir.ac.uk/ [2008, 4/28] .

Raytchev, B., Yoda, I. & Sakaue, K. 2004, "Head pose estimation by nonlinear manifold learning", *Proceedings of the 17th International Conference on Pattern Recognition,* vol. 4, pp. 462-466.

Roweis, S.T. & Saul, L.K. 2000, "Nonlinear dimensionality reduction by locally linear embedding", *Science,* vol. 290, no. 5500, pp. 2323-2326.

Rowley, H., Baluja, S. & Kanade, T. 1998, "Rotation invariant neural network-based face detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 38–44.

Schiele, B. & Waibel, A. 1995, "Gaze tracking based on face-color", *Proceedings of the International Workshop on Automatic Face-and Gesture-Recognition,* pp. 344-349.

Sherrah, J. & Gong, S. 2001, "Fusion of perceptual cues for robust tracking of head pose and position", *Pattern Recognition,* vol. 34, no. 8, pp. 1565-1572.

Sherrah, J., Gong, S. & Ong, E.J. 2001, "Face distributions in similarity space under varying head pose", *Image and Vision Computing,* vol. 19, no. 12, pp. 807-819.

Sherrah, J., Gong, S. & Ong, E.J. 1999, "Understanding pose discrimination in similarity space", *Proceedings of the 10th British Machine Vison Conference,* pp. 523-532.

Shih, F.Y. & Chuang, C.F. 2004, "Automatic extraction of head and face boundaries and facial features", *Information Sciences,* vol. 158, pp. 117-130.

Sim, T., Baker, S. & Bsat, M. 2003, "The CMU Pose, Illumination, and expression database", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1615-1618.

Softopia *, HOIP face database*. Available: http://www.softopia.or.jp/en/rd/facedb.html.

Srinivasan, S. & Boyer, K. 2002, "Head pose estimation using view based eigenspaces", *Proceedings of the International Conference on Pattern Recognition*, pp. 302-305.

Stiefelhagen, R. 2004, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data", *Proceedings of the Pointing 2004 Workshop on Visual Observation of Deictic Gestures*.

Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D. & Soundararajan, P. 2007, "The CLEAR 2006 evaluation", *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006: Revised Selected Papers*Springer-Verlag New York Inc, pp. 1-44.

Stiefelhagen, R., Yang, J. & Waibel, A. 2002, "Modeling focus of attention for meeting indexing based on multiple cues", *IEEE Transactions on Neural Networks,* vol. 13, no. 4, pp. 928-938.

Tang, X., Ou, Z., Su, T., Sun, H. & Zhao, P. 2005, "Robust precise eye location by adaboost and svm techniques", *Proceedings of the International Symposium on Neural Networks,* Springer, pp. 93–98.

Tenenbaum, J.B., Silva, V. & Langford, J.C. 2000, "A global geometric framework for nonlinear dimensionality reduction", *Science,* vol. 290, no. 5500, pp. 2319-2323.

Tian, Y.L., Brown, L., Connell, C., Pankanti, S., Hampapur, A., Senior, A., Bolle, R., Center, I.B.M.T.J.W.R. & Hawthorne, N.Y. 2003, "Absolute head pose estimation from overhead wide-angle cameras", *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures,* pp. 92-99.

Tu, J., Fu, Y., Hu, Y. & Huang, T. 2007, "Evaluation of head pose estimation for studio data", *Lecture Notes in Computer Science,* vol. 4122, pp. 281-290.

Viola, P. & Jones, M.J. 2004, "Robust real-time face detection", *International Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154.

Voit, M. 2007, "CLEAR 2007 evaluation plan: Head pose estimation", http://isl.ira.uka.de/~mvoit/clear07/CLEAR07_HEADPOSE_2007-03-26.doc, 2007.

Voit, M., Nickel, K. & Stiefelhagen, R. 2008, "Head Pose Estimation in Single-and Multi-view Environments--Results on the CLEAR'07 Benchmarks", *Lecture Notes In Computer Science,* vol. 4625, pp. 307-316.

Voit, M., Nickel, K. & Stiefelhagen, R. 2007, "Neural network-based head pose estimation and multi-view fusion", *Lecture Notes in Computer Science,* vol. 4122, pp. 291-298.

Voit, M., Nickel, K. & Stiefelhagen, R. 2006, "A bayesian approach for multi-view head pose estimation", *In Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 31-34.

Wang, X., Huang, X., Gao, J. and Yang, R. 2008, "Illumination and Person-Insensitive Head Pose Estimation Using Distance Metric Learning", *Lecture Notes in Computer Science,* vol. 5303, pp. 624-637.

Wang, J.G. & Sung, E. 2007, "EM enhancement of 3D head pose estimated by point at infinity", *Image and Vision Computing,* vol. 25, no. 12, pp. 1864-1874.

Wang, J.G. & Sung, E. 1999, "Frontal-view face detection and facial feature extraction using color and morphological operations", *Pattern Recognition Letters,* vol. 20, no. 10, pp. 1053-1068.

Wu, J., Pedersen, J.M., Putthividhya, D., Norgaard, D. & Trivedi, M.M. 2004, "A two-level pose estimation framework using majority voting of gabor wavelets and bunch graph analysis", *Proceedings of the Pointing 2004 Workshop: Visual Observation of Deictic Gestures*.

Wu, J. & Trivedi, M.M. 2008, "A two-stage head pose estimation framework and evaluation", *Pattern Recognition,* vol. 41, no. 3, pp. 1138-1158.

Wu, J. & Zhou, Z.H. 2003, "Efficient face candidates selector for face detection", *Pattern Recognition,* vol. 36, no. 5, pp. 1175-1186.

Yilmaz, A. & Shah, M.A. 2002, "Automatic feature detection and pose recovery for faces", *ACCV2002: Proceedings of the 5th Asian Conference on Computer Vision, Melbourne, Australia.*

Yu, S., Kim, J. & Lee, S. 2009, "Iterative Three-dimensional Head Pose Estimation using a Face Normal Vector", *Optical Engineering,* vol. 48, no. 3 (in press).

Zhang, Z., Hu, Y., Liu, M. & Huang, T. 2007, "Head pose estimation in seminar room using multi view face detectors", *Lecture Notes in Computer Science,* vol. 4122, pp. 299-304.

Zhao, L., Pingali, G. & Carlbom, I. 2002, "Real-time head orientation estimation using neural networks", *Proceedings of the International Confererence on Image Processing,* vol.1, pp. 297-300.

Zhou, Z.H. & Geng, X. 2004, "Projection functions for eye detection", *Pattern Recognition,* vol. 37, no. 5, pp. 1049-1056.

Zhu, Y. & Fujimura, K. 2004, "Head pose estimation for driver monitoring", *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 501-506.

APPENDICES

APPENDIX A: A HIGH RESOLUTION COLOUR IMAGES FACE POSE DATABASE

A.1 INTRODUCTION

This appendix describes the acquisition and contents of a face pose database of high resolution colour images. Among the existing standard face pose image databases, only Pointing' 04 and CAS-PEAL are available publically. The images in the pointing' 04 database (Gourier, Hall & Crowley 2004a) are low resolution and badly illuminated. Most face pose estimation techniques when applied to the Pointing' 04 database require the facial points to be located manually. The CAS-PEAL database (Gao et al. 2008) contains only grey level images and algorithms designed for colour images can not be evaluated using this database. Therefore, there is a need to create a new face pose database of high resolution colour images. The advantages of the new database are: (1) the author's proposed face pose estimation techniques can be fully automated as there is no need for manual selection of feature points with high resolution, good illumination images, (2) other researchers may also be able to test their pose estimation schemes fully automatically without any manual selection, and (3) it can be used for evaluating and comparing different facial feature extraction schemes for varying face poses.

The rest of the appendix is organized as follow: section A.2 briefly discusses the capturing techniques for face pose databases; section A.3 discusses the existing standard pose databases; section A.4 describes the acquisition and contents of the new database; section A.5 gives a summary of this appendix.

A.2 POSE DATABASE CAPTURING APPROACHES

The following are the most common approaches that have been used for capturing images for pose databases.

Directional Suggestion: In directional suggestion approach, markers are placed at discrete locations around a room and the subjects are asked to direct their face towards each of these markers in turn while a camera captures images for every location. Pointing' 04

dataset (Gourier, Hall & Crowley 2004b) was captured using directional suggestion approach. There are two main drawbacks of this approach: firstly, it assumes that each subject's head is in the exact same physical 3D location. Secondly, it assumes that a person has the ability to exactly direct his head towards a point. Unfortunately, both of these are subjective and it is almost impossible to achieve 100 percent accuracy in both cases. That is why substantial errors have been reported in Pointing' 04 database by (Tu et al. 2007).

Directional Suggestion with Laser pointer: This approach is similar to directional suggestion, but a laser pointer is affixed to the subject's head. The subject is asked to direct their head towards a marker so that the laser light falls on that specific marker. This approach is more accurate than simple directional suggestion, but it still assumes that a subject's head is located at the same point in the 3D space, which is difficult to ensure.

Manual Annotation: In this approach, a person assigns a pose label to each image based on his own perception after examining it. This approach is purely subjective and its accuracy is very low.

Camera Arrays: Multiple cameras are placed at known discrete positions and images of a subject's face is captured simultaneously by all these cameras. If care is taken to keep the subjects' heads in the same location, high accuracy can be achieved using this approach. This approach is more accurate than directional suggestion, directional suggestion with laser pointer, and manual annotation. The disadvantage of this approach is that it is only applicable to near-field images.

Magnetic Sensors: Magnetic sensors are affixed to a subject's head and used to estimate the position and orientation of the head. This approach has a theoretical inaccuracy of less than 1° but it is highly susceptible to noise as reported by Murphy-Chutorian & Trivedi (2009). Therefore, this approach is limited by the environment in which it can be used.

Inertial Sensors: In this approach, inertial sensors such as accelerometer, gyroscopes, or other motion sensing devices are affixed to subjects' heads to capture their head pose estimated images. The advantage this approach has over magnetic sensors is that it is more robust against metallic interference.

Optical Motion Capture Systems: These are the most accurate and most expansive systems. In these systems, an array of calibrated near-infrared cameras use multi-view stereo and software to follow reflective or active markers attached to a person's head.

A.3 EXISTING POSE DATABASES

The following is a brief description of some of the common pose databases.

Pointing '04: Pointing' 04 (Gourier, Hall & Crowley 2004b) is the most widely used pose database and is available publically online. It was included as part of the Pointing 2004 Workshop on Visual Observation of Deictic Gestures to ensure the uniform evaluation of head pose estimation schemes. It was also used as one of the two dataset in the 2006 International Workshop on Classification of Events Activities and Relationships (CLEAR'06) (Stiefelhagen et al. 2007). The Pointing' 04 comprises 15 sets of near-field images, with each set containing 2 series of 93 images of the same person at 93 discrete poses. The pitch and yaw both vary from -90° to 90° and the subjects range in age from 20 to 40 years old. Five subjects possess facial hair and seven subjects wear glasses. The Pointing' 04 was obtained using directional suggestion and it contains substantial error as noted by (Tu et al. 2007).

CHILL-CLEAR06: CHILL-CLEAR06 was created for the evaluation of head pose estimation systems in CLEAR'06 workshop and it contains multi-view video recordings of seminars (Stiefelhagen et al. 2007). The database contains 12 training sequences and 14 test sequences recorded in a seminar room equipped with four synchronized camera in the corners of the room. The videos are far-field, low resolution in natural lighting conditions and manual annotations are made in every tenth video frame, providing a bounding box around a subject's head and coarse pose orientation label from one of the 8 discrete yaws at 45° intervals.

CHILL-CLEAR07: CHILL-CLEAR07 was created for the evaluation of pose estimation systems in CLEAR'07 workshop and similar to CHILL-CLEAR06, it contains recorded sequence of seminars, captured in a seminar room equipped with four cameras in the corners (Voit 2007). However, in CHILL-CLEAR07, the head pose estimate was provided by a magnetic sensor, providing fine head pose information. Apart from pose information from the magnetic sensor, manual annotation of the bounding box around each presenter's head was provided for every 5^{th} video frame. The database comprises 15 video sequences captured at 15fps.

IDIAP Head Pose: The IDIAP Head pose database consists of 8 video sequences recorded with a single camera (Ba, Odobez 2004). Each of the videos is approximately one minute

in duration. In these videos, two subjects who are always visible were continuously annotated using magnetic sensor to measure head location and orientation. The head pose information is provided as fine measures of pitch and yaw both ranging from -90 to 90 degrees. This database was one of the datasets used for evaluation in CLEAR'07 Workshop.

CMU PIE: The CMU Pose, Illumination and Expression (PIE) database contains 68 images of different people across 13 poses, 43 different illumination conditions and 4 different expressions (SIM, BAKER & BSAT 2003). The database was obtained with a 13 cameras array with 7 of cameras placed at 22.5° interval across yaw, one camera above the centre, one camera below the centre, and one in each corner of the room.

Softopia HOIP: Softopia HOIP dataset contains two sets of near-field face images of 300 anonymous persons. The first set consists of 168 discrete poses (24 yaws and 7 pitches at 15 degrees intervals). The second set contains finer yaw increments (73 yaws at 5 degrees interval and 7 pitches at 15 degree intervals). The images were captured using rotation platforms and a camera array.

CVPR-86: The CVPR-86 dataset consists of 3894 near-field images of 28 persons in 86 discrete poses (Wu, Trivedi 2008). Pitch varies from -45° to 45° with an increment of 15°, while yaw varies from -90° to 90° with increment of 15°. Not every pose has samples for each person in the database. The database was captured using the magnetic sensor aproach.

CVPR-363: CVRR-363 contains images of 10 people against a uniform background. The pitch varies from -30° to 20° and yaw varies from -80° to 80° with interval of 5° (Murphy-Chutorian, Doshi & Trivedi 2007). Optical motion capturing system was used to create this database.

USF HumanID: The USF HumanID dataset contains color images of 10 people against a uniform background. There are 181 images of each person with varying yaw from -90 to 90 degrees with 1 degree interval (Bowyer, Sarkar 2001).

BU Face Tracking: The BU face tracking dataset contains 30fps video sequences (Canton-Ferrer, Casas & Pardas 2007). There are two set of videos: the first set consists of 9 video sequences for each of the five subjects with uniform illumination. The second set is the same as the first but with varying illumination. The head pose information was recorded through magnetic sensor in the BU Face tracking dataset.

CVRR LISAP-14: This database contains 14 video sequences of drivers (Murphy-Chutorian & Trivedi, 2008). Each video is approximately 8 minutes long and the lighting condition is either daytime or night time. Head position and orientation were recorded using an optical motion capturing system.

CAS-PEAL: The CASE-PEAL dataset contains 99594 images of 1040 people with varying poses, expression and lighting (Gao et al., 2008). There are 27 discrete poses for each individual. For each person, 9 yaws were simultaneously captured with camera array at 3 pitches. A subset of the database can be obtained on request.

FacePix: The FacePix dataset contains 181 images for each of the 30 subjects spanning -90 to 90 degrees in yaw at 1 degree interval (Little et al. 2005). The images were captured using a camera on rotating platform and they were cropped manually to ensure that eyes appear in the same vertical position in every view.

A.4 LU-RSI-ID: A NEW POSE DATABASE

A laser light with a stand that could be rotated right and left, and tilted up and down, was used to mark the pose markers on the walls of the room. A camera with a stand was used for capturing the images. The laser light was placed in such a position that it coincided with the camera's position vertically. An angle divider was affixed to the laser light to achieve accuracy in marking the pose markers on the wall as shown in Figure A-4. A total of 93 pose points (13 for each yaw which varies from -90° to 90° with an interval of 15° and 7 for each discrete pitch which varies from -45° to 45° with an interval of 15°) were marked on the walls of the room.

During the photo shoot, a chair was placed and fixed in such a position that its centre was in exactly in the same position where the laser light was placed for marking the pose markers. Each of the subjects was then asked to sit on the chair and direct his head towards each marker in turn and images were captured with a 5 Megapixels still camera which was fixed in such a position that it coincided with each subject's head vertically and with zero degree yaw marker horizontally. All the images were captured in good lighting condition. Figures A-1 to A-5 show the experimental setup. Figures A-6 and A-7 show sample images from the database which contains a total of 910 images for 10 subjects. Three of the subjects had facial hair, one wore glasses and one wore a cap.

Figure A-1 Laser Pointer with a Stand

Figure A-2 Laser Pointer Kept in Level with Camera Lens

Figure A-3 Laser Light with Angle Divider

Figure A-4 Laser Pointer Tilted 30 Degrees Down

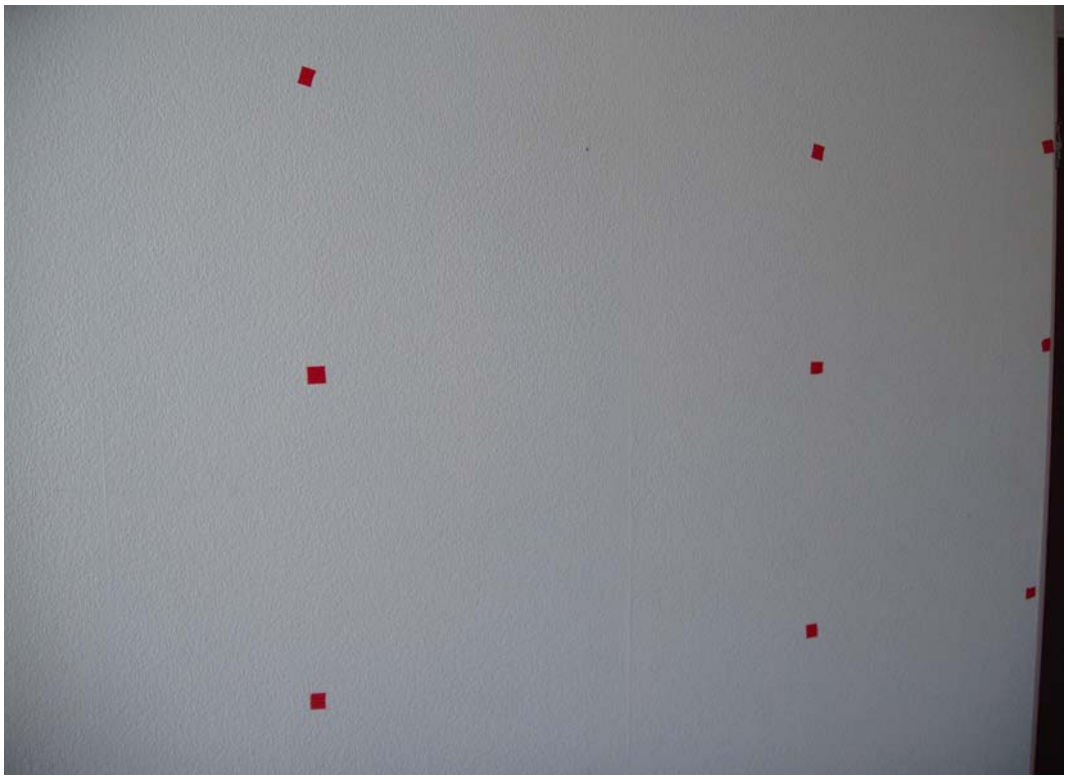


Figure A-5 Pose Markers on a Wall

Figure A-6 Sample Images from the database with 0° yaw and 0° pitch

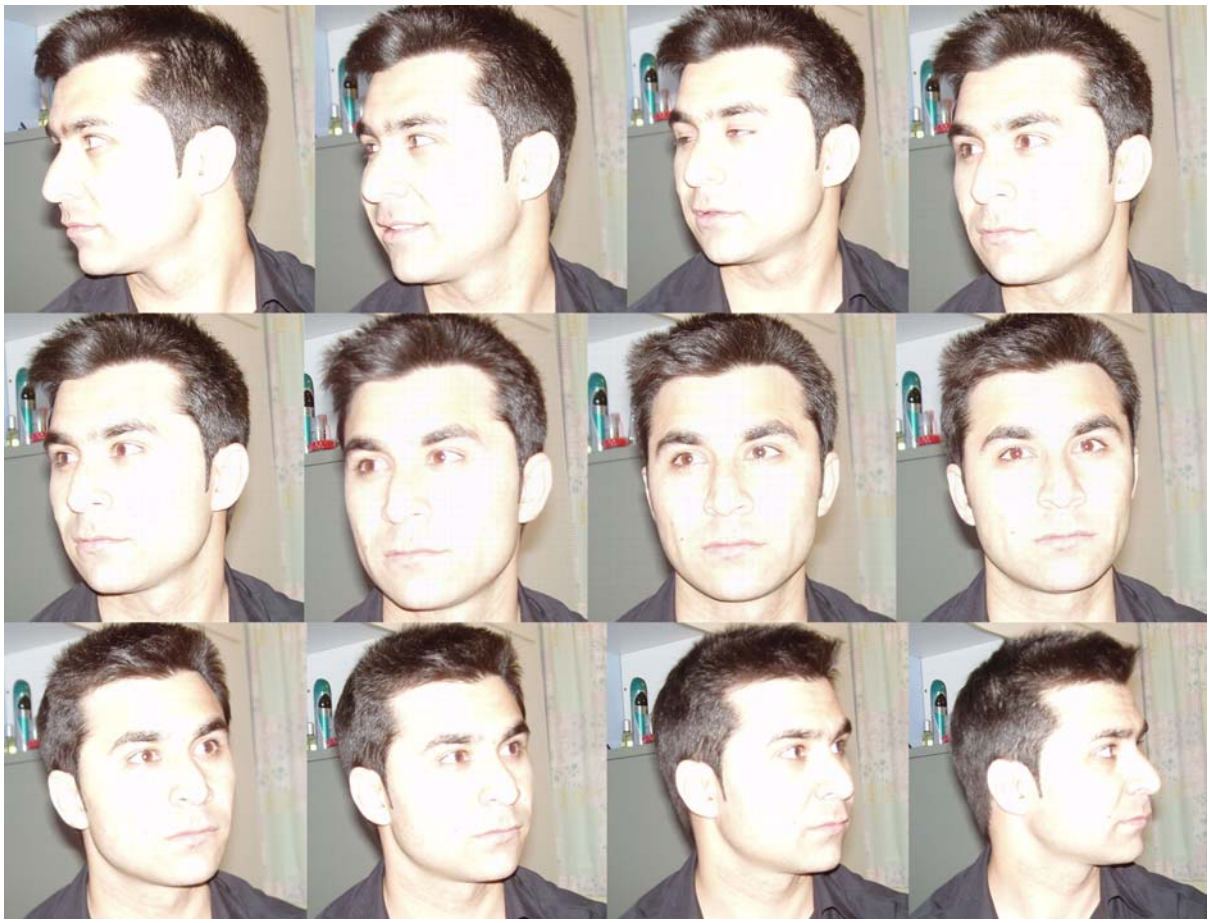


Figure A-7 Sequence of Images with 0° Pitch

Table A-1 Database Summary

No. of Subjects	Condition	Number of Images
1	Wearing Glasses	91
3	Have facial hair	273
6	Others	546

A.5 SUMMARY

The acquisition and contents of a new face pose database with high resolution colour images was discussed in this article. Different pose capturing approaches and existing pose databases were described first, followed by a detailed description of the acquisition process of the new database. As the new database was captured using directional suggestion, so it is prone to errors. However, the new database is of sufficient accuracy and quality for discrete coarse face pose estimation evaluation. Reseachers, including the author, can use the database to test their pose estimation methods fully automatically without any manual point selection because the images are high resolution and taken in good illumination condition. It can also be used for the evaluation of facial feature extraction methods for different face poses. The database can be accessed online at http://www.lboro.ac.uk/schools/informatics/poseDB.html.

APPENDIX B: LIST OF PUBLISHED PAPERS

The work presented in this thesis has resulted in the following contributions.

Shafi, M. and **Chung, P.W.H**., "Eyes Extraction from Facial Images Using Edge Density", *Proceedings of 7th IEEE International Conference on Cybernetic Intelligent Systems*, London, UK, 2008, pp. 317-322, ISBN: 978-1-4244-2914-1.

Shafi, M. and **Chung, P.W.H.,** "A Hybrid Method for Eyes Detection in Facial Images", *International Journal of Electrical, Computer, and Systems Engineering*, 3(4), 2009, pp. 231-236, ISSN: 2070-3813.

Shafi, M. and **Chung, P.W.H.,** "Face Pose Estimation from Eyes and Mouth", *International Journal of Advanced Mechatronic Systems*, 2(1/2), 2010, pp. 132-138, ISSN: 1756-8420.

Shafi, M. and **Chung, P.W.H.,** "Face Pose Estimation using Distance Transform and Normalized Cross-correlation", *IET Computer Vision* (Submitted).