ROBUST FACIAL FEATURE EXTRACTION
FOR VIDEO CODING

PHANG CHUI LIAN

A thesis submitted
in fulfillment of the requirement for the degree of Master of Science.

Faculty Of computer Science and Information Technology
UNIVERSITY MALAYSIA SARAWAK
2006
ACKNOWLEDGEMENT

Firstly, I would like to thank my supervisor, Prof. Khairuddin Ab. Hamid for his invaluable guidance and encouragement throughout this research. I would also like to thank my co-supervisor, Dr. Abdelhamid Abdesselam, for his invaluable technical guidance and encouragement throughout this research.

I would like to thank my friends for useful discussions and moral support, as well as extending my gratitude to all the speakers in my test samples for their kind cooperation.

Besides, I would like to acknowledge the support of National Science Fellowship (NSF) under Ministry of Science, Technology and Environment as well as Faculty of Computer Science and Information Technology and Centre for Graduate Studies, Universiti Malaysia Sarawak.

Finally, I am grateful to my family members for their support.
ABSTRACT

Model-based coding is a good alternative to traditional statistical-based techniques to deal with the bandwidth limitation in video conferencing over low speed networks. In order to apply model-based coding in real life video conferencing, robust extraction of facial features is needed. Since mouth and eyes are usually extracted first and used as reference points for the extraction of other facial features, this research focuses on enhancing robustness of eyes and mouth extraction from video conferencing sequences. The outcome of the research is a technique that is more robust against changes in skin colour, illumination, face orientation, eye colour, shirt colour, and speed of the speaker's movement.
**ABSTRAK**

Pengekodan data berdasarkan model adalah satu alternatif baik untuk menggantikan pengekodan data berdasarkan statistik bagi menangani masalah lebar jalur yang terhad. Demi mengaplikasikan pengekodan data berdasarkan model dalam persidangan yang sebenar, teknik pengekstrakan ciri-ciri penting muka yang tegap atau "robust" sangat diperlukan. Disebabkan mulut dan mata sering diekstrakkan terlebih dahulu dan digunakan sebagai titik-titik rujukan dalam pengekstrakan ciri-ciri penting muka yang lain, maka penyelidikan ini memberi penumpuan kepada peningkatan ketegapan pengekstrakan mata dan mulut daripada rentetan-rentetan persidangan video. Penyelidikan ini telah menghasilkan satu teknik yang lebih tegap terhadap perubahan-perubahan dalam warna kulit, pencahayaan, orientasi muka, warna mata, warna baju, dan kelajuan pergerakan penutur.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgement</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Abstrak</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
</tbody>
</table>

## 1 Introduction

1.1 Background                  | 1
1.2 Problem Statement           | 3
1.3 Scope of the Research       | 3
1.4 Objectives of the Research  | 4
1.5 Significance of the Research| 4
1.6 Thesis Organisation         | 4

## 2 Literature Review

2.1 Introduction                | 6
2.2 Mouth and Eyes Extraction Approaches | 6
   2.2.1 Direct Extraction        | 6
   2.2.2 Extraction through Face Localisation | 7
      2.2.2.1 High-level Approach | 7
      2.2.2.2 Hybrid Approach     | 9
      2.2.2.3 Low-level Approach  | 10
3 System Design

3.1 Introduction 29
3.2 Solution Requirements 29
3.3 Design 29

3.3.1 Silhouette Localisation 30
3.3.2 Face Localisation 31

3.3.2.1 Retaining Hue [6, 31] 31
3.3.2.2 Discarding Quasi-chroma 32

3.3.3 Mouth Extraction 32

3.3.3.1 Including Hue [0, 5] 33
3.3.3.2 Retaining Quadrature Q 33
3.3.3.3 Discarding Intensity 34

3.3.4 Eyes Extraction 34

3.3.4.1 Enhancement of Face Region 34
3.3.4.2 Replacement of Qc, I, Q with Hue [0,31] 34
3.3.4.3 New Geometrical Constraints 35

3.3.5 Repetition of Mouth and Eyes selection 35

3.4 Context Diagram 36
3.5 Data Flow Diagrams 36

3.5.1 Silhouette Localisation (Figures 3.3) 37
3.5.2 Face Localisation (Figure 3.5) 39
3.5.3 Mouth and Eyes Extraction (Figure 3.6) 39
3.6 Summary 40

4 Implementation 41
4.1 Introduction 41
4.2 Algorithms 41
  4.2.1 Silhouette Localisation (Figure 4.1) 41
  4.2.2 Face Localisation (Figure 4.4) 44
  4.2.3 Mouth and eyes extraction (Figure 4.5) 45
4.3 Data Structure 48
4.4 Summary 49

5 Testing and Analysis 50
5.1 Introduction 50
  5.1.1 Source of the Test Samples 50
  5.1.2 Parameter Setting 51
  5.1.3 Threshold Setting 52
  5.1.4 Result Evaluation 52
5.2 Module Testing 53
  5.2.1 Test Samples 54
  5.2.2 Silhouette Localisation 55
  5.2.3 Face Localisation 56
  5.2.4 Mouth Extraction 61
  5.2.5 Eyes Extraction 64
(Skin Colour, Illumination, Face Orientation)

Appendix D: Results From Shiga et al. (2000)'s System 118

(Skin Colour, Illumination, Face Orientation)

Appendix E: Testing under Shiga et al. (2000)'s constraints 126

Appendix F: Determining Suitable Hue Range for Narrowing 127

Appendix G: The related Conference Paper – 128

International Conference on Communication,
Computers and power (ICCCP'05)
LIST OF TABLES

Table 2.1 Robustness of low-level approaches against changes in skin colour, illumination, face orientation, eye colour, shirt colour, and speed of the speaker’s movement

(\(x = \) not robust enough, \(\sim = \) relatively robust, \(\checkmark = \) robust enough)

Table 2.2 Robustness of various features

Table 3.1 Comparison between the features used in previous and the proposed algorithms

Table 5.1 Percentage of correct extraction of mouth and eyes against changes in skin colour, illumination, and face orientation

Table 5.2 Percentage of correct extraction of mouth and eyes against changes in skin colour

Table 5.3 Percentage of correct extraction of mouth and eyes against changes in illumination

Table 5.4 Percentage of correct extraction of mouth and eyes against changes in face orientation
LIST OF FIGURES

Figure 1.1  Structure of a model-based codec based on FAPs  3
Figure 2.1  Points on the face where intensity values are compared  21
Figure 3.1  Context diagram of the proposed system  36
Figure 3.2  First level data flow diagram  37
Figure 3.3  Data flow diagram of silhouette localisation  38
Figure 3.4  Data flow diagram of change detection  38
Figure 3.5  Data flow diagram of face localisation  39
Figure 3.6  Data flow diagram of mouth and eyes extraction  40
Figure 4.1  Algorithm for Silhouette Localisation  42
Figure 4.2  Algorithm for Calculation of Threshold  43
Figure 4.3  Algorithm for Fusing of Changes  44
Figure 4.4  Algorithm for Face Localisation  45
Figure 4.5  Algorithm for Mouth and Eyes Extraction  46
Figure 4.6  Algorithm for Q-thresholding  46
Figure 4.7  Algorithm for Mouth and Eyes Selection  47
Figure 4.8  Data structure of Frame  48
Figure 4.9  Data structure of Silhouette  48
Figure 4.10  Data structure of Eye-pair  49
Figure 4.11  Data structure of OtherColourClusters  49
Figure 5.1  The major modules of the system  53
Figure 5.2  Result of silhouette module under various speeds of speaker's movement  55
Figure 5.3  Result of face module under various skin colours  57
Figure 5.4  Result of face module under various illuminations  
Figure 5.5  Result of face module under various face orientations  
Figure 5.6  Result of face module under various shirt colours  
Figure 5.7  Result of mouth module under various illuminations  
Figure 5.8  Result of mouth module under various shirt colours  
Figure 5.9  Result of eyes module under various skin colours  
Figure 5.10 Result of eyes module under various illuminations  
Figure 5.11 Result of eyes module under various face orientations  
(search range of eyes)  
Figure 5.12 Result of eyes module under various face orientations  
(shadowing effect)  
Figure 5.13 Result of eyes module under various eye colours  
Figure 5.14 Result of system testing under various eye colours  
Figure 5.15 Result of system testing under various shirt colours  
Figure 5.16 Result of system testing under various speeds of the  
speaker's movement
CHAPTER 1 INTRODUCTION

1.1 Background

Video conferencing is a very effective way of communication because it allows face-to-face conversations and meetings to be held between remote stations. With technology advancement, video conferencing has emerged and become increasingly important. Video conferencing has been applied over various types of networks, ranging from local area network (LAN), metropolitan area network (MAN), and wide area network (WAN), regardless of wired or wireless. Unfortunately, low speed networks cannot fully support this application because video conferencing data has large size and cannot tolerate delay. To overcome this problem, a highly efficient coding technique is needed to reduce the video size (Menser & Brünig, 2000) while maintaining visual quality.

Today's coding techniques for video conferencing frames are block-based, such as H.261 standard and H.263 standard. Although these techniques give high coding ratio, they are not effective enough for video conferencing applications. The new MPEG-4 coding technique uses the concept of object-based coding, which has higher compression ratio and better visual quality. Following this, H.264 standard is introduced, somewhat based on MPEG-4. Despite better compression ratio as compared to H.261 and H.263, it is still block-based and not effective enough for video conferencing. Recently, some research works on model-based coding have been carried out (Yau & Duffy, 1989; Choi et al., 1991; Nobori et al., 1992; Reinders et al., 1992; Aizawa & Huang, 1995; Strub & Robinson, 1995; Ngan & Rudianto, 1996; Tang & Huang, 1996; Ahlberg et al., 1997; Davis & Tuceryan, 1997; Eisert & Girod, 1997; Kampmann, 1997; Kampmann & Farhoud, 1997; Eisert & Girod, 1998; Noh & Neumann, 1998; Ström, 1998; Ahlberg,
2001; Kampmann, 2001; Kampmann, 2002; Yin & Basu, 2000). It is a more specific type of object-based coding, whereby a priori knowledge about scene content is used to build models for the objects in the scene, leading to higher coding ratio as compared to general object-based coding. Since there is a priori knowledge about video conferencing scenes, model-based coding approach is recently proposed for low speed networks, in order to reduce data size and thus optimise the performance of video conferencing.

Structure of a model-based codec based on facial animation parameters (FAPs) is as shown in Figure 1.1 (Eisert & Girod, 1998). Firstly, a generic 3-dimensional model is automatically adapted to the speaker's face in video conferencing frames. Then, the encoder estimates the person's 3-dimensional motion and facial expressions to get a set of facial animation parameters (FAPs), which describe (together with the 3-dimensional model) the person's current appearance. Therefore, only a few parameters need to encoded and transmitted, leading to very low data rate. At the decoder, FAPs deform the head model according to the speaker's facial expressions prior to rendering of camera frames (Eisert & Girod, 1998; Phang et al., 2002).
In order to adapt a generic face model to an individual face, the facial features have to be extracted. Generally, mouth and eyes are first extracted and used as reference points for the extraction of other facial features (Kampmann, 2002).

1.2 Problem Statement

Although a number of techniques proposed for the extraction of mouth and eyes are computationally effective and give good results, some assumptions made on speaker and background such as face looking to the front and plain background are not valid in real life. This research aims to produce a more robust technique for extraction of mouth and eyes from video conferencing sequences. It is expected to be robust against skin colour, illumination, face orientation, eye colour, shirt colour, and speed of the speaker's movement.

1.3 Scope of the Research

This research deals with extraction of mouth and eyes from typical head-and-shoulder video conferencing sequences, captured using a digital camera fixed at a short distance in front of the speaker. The scene always consists of a moving speaker in front of a static background, under realistic office environment and illumination. Besides, the head tilt of
the speaker can be in all directions, but only within an extent where mouth and eyes are distinctly visible. The tasks include design and implementation of algorithms to localise silhouette and face, and then extract mouth and eyes.

1.4 Objectives of the Research
Throughout this research, the following objectives are to be met:

i. To study the current development of mouth and eyes extraction from video conferencing sequences.

ii. To investigate and address problems related to robustness of mouth and eyes extraction techniques.

iii. To design more robust algorithms for extraction of mouth and eyes from video conferencing sequences.

iv. To produce a prototype system for the extraction of mouth and eyes from video conferencing sequences.

v. To compare and contrast the performance of the proposed system.

1.5 Significance of the Research
A robust mouth and eyes extraction technique, which is much needed to make model-based coding applicable to real life video conferencing, is proposed and designed.

1.6 Thesis Organisation
Chapter 2 discusses and analyses existing approaches for extraction of mouth and eyes from video conferencing sequences. Chapter 3 describes some design aspects of the proposed mouth and eyes extraction system. Chapter 4 presents the algorithms and major data structure of the proposed system. The analysis of the proposed system is
presented in Chapter 5, supported with results from module testing and system testing. Chapter 6 concludes this thesis with achievements and limitations of this research, and recommendations for future work.
2.1 Introduction

This chapter begins with an overview of the existing mouth and eyes extraction systems, which can be generally categorised into direct extraction and extraction through face localisation. Then, low-level approaches that are more suitable for video conferencing are studied and analysed. Finally, the robustness of low-level approaches are discussed and summarised.

2.2 Mouth and Eyes Extraction Approaches

Various techniques have been proposed for mouth and eyes extraction. A careful study of these techniques shows that mouth and eyes extraction system can be grouped into two major approaches: direct extraction and extraction through face localisation.

2.2.1 Direct Extraction

In this approach, mouth and eyes are directly extracted from the image using their special characteristics or properties in the image.

Reisfeld and Yeshurun (1992) used generalised symmetry operator to extract mouth and eyes based on the assumption that facial features are symmetrical in nature.

Firstly, edges are obtained by performing convolutions with Gaussian derivatives in two directions. The vertical midline of the face is then detected by looking for the best global correlation of directional intensity gradients of left and right half images. On the other hand, each point is assigned a symmetry magnitude and a symmetry orientation using
symmetry operator. Accumulated symmetry of each point is obtained by linking each symmetry value to its neighbours. The centre of each symmetry cluster is singled out as being local maxima in the accumulated symmetry map.

Eyes are regions with high symmetry measures at both sides of the midline while mouth is the region with high symmetry measure that crosses the midline.

Other examples that used this approach are Takaya and Choi (2001), Vincent et al. (1992), and Mu et al. (1996).

Their technique is direct but sensitive to noise because mouth and eyes are relatively small when compared to other components in the image. Besides, background objects may have some properties similar to eyes or mouth and thus can be falsely extracted as eyes or mouth.

2.2.2 Extraction through Face Localisation

Before usage of special characteristics of mouth and eyes as in direct extraction, this approach localises the face area. This approach can be further categorised into high-level approach, hybrid approach, and low-level approach.

2.2.2.1 High-level Approach

In this approach, intelligent and advanced pattern recognition techniques are used to localise face and extract mouth and eyes.

The first level GWN is used for face matching and rough approximation of feature locations. First level GWN, which represents the entire face, is used to search for faces in the database that are similar to the target. For each candidate face, an affine transformation of level-one GWN that registers the candidate with the target image is determined.

The second level wavelet network is used to fine-tune the feature locations. Second level GWNs, representing each feature, are initialised in positions according to the affine transformation from the first level GWN. For each feature, a brute-force search is performed within a limited window for a position that minimizes the score in wavelet subspace between a candidate level-two feature GWN, and the target image. They are only allowed to have slight translation movement to minimise their difference from the new face.

Other examples that used this approach are Lee et al. (2001), Kim et al. (2004), and Zuo and Peter (2005).

This approach is computationally intensive and thus not suitable for real-time video conferencing.
2.2.2.2 Hybrid Approach

This approach is a combination of high-level approach and low-level approach. It involves high-level approach that utilises intelligent and advanced pattern recognition techniques. It also incorporates a low-level approach that makes use of low-level features.

Choi and Takaya (2001) localised face using colour feature (low-level approach) and then extracted mouth and eyes from localised face using Independent Component Analysis (ICA), which is a high-level approach.

Firstly, skin colour filtering in normalized YCbCr colour space is applied to the moving blocks in the difference image. An ellipse, initially set as a circle, is placed such that its upper end is at the upper vertex of the skin-like region and kept fixed. This ellipse is enlarged towards the bottom of the region or until a predefined ratio between the axes of the ellipse was reached. Then, the elliptical face is normalized to a pre-assigned size for the subsequent step of ICA analysis to extract mouth and eyes.

An approximation of the original input image $X$ is reconstructed with ICA bases and their corresponding coefficient. From all the ICA bases, major components are carefully studied to identify those that have highly concentrated values representing eyes or mouth. The selected bases are sufficient to reconstruct localised images of facial parts. To search ICA bases vectors that are regarded as eye-like or mouth-like, a heuristic approach was used. Eye-like features are searched in the upper half and either in the left or right half of an ICA basis matrix. The mouth-like features are searched in the lower half of the ICA matrix.
Other examples that used this approach are Lin and Wu (1999), Lee et al. (2001), Wong et al. (2001a), Wong et al. (2001b), Yen and Nithianandan (2002), and Song et al. (2004).

Although this approach is faster than high-level approach, it is still too computationally intensive for real-time video conferencing application.

2.2.2.3 Low-level Approach

This approach utilises low-level features, such as colour and intensity, together with a priori knowledge of face structure and video conferencing scene content. It is a simplified version of facial feature extraction, and is therefore computationally effective and suitable for real-time video conferencing. Low-level approach can be further categorised into intensity-based approach and colour-based approach.

2.2.2.3.1 Intensity-based Approach

This approach extracts mouth and eyes using intensity images, making use of intensity features only.

Alattar and Rajala (1999) assumed that human head could be thought of as an ellipsoid that sits on the top of the torso and is attached to it by the centre of the neck. Under this ellipsoid model, the front view of head has elliptical shape and tilt angle $\theta$ (side to side rotation). In their algorithm, the ellipse with tilt angle $\theta$ (side to side rotation) that best fits the front view of the speaker is first estimated. However, the fitting steps are not mentioned.
A model describing the relative locations of the top of the head, sides of the head, eyes, nose, mouth, and chin is then used to estimate the positions of mouth and eyes. The estimations made include:

i. The average human head is approximately five eye lengths wide.

ii. Both eyes lie on the line midway between the top of the head and the chin.

iii. The distance between both eyes is one eye length.

iv. The nose lies between face centre and midway between face centre and the chin.

v. The mouth barrel lies between nose base and 2/3 of the distance down from the nose to the chin.

vi. The distance between mouth centre and face centre is approximately 1/3 head length.

vii. Mouth corners align with eye socket centres.

Finally, mouth and eyes locations are refined using vertical and horizontal projections of the pixels in windows around the estimated mouth and eyes locations. Mouth and eyes are assumed to dwell in valleys in their respective windows.

In another work, Aas (1998) first localises silhouette using thresholded difference of two consecutive frames. Bounding box of the head is then estimated. The top of the head is assumed as the top row containing foreground pixels. A dashed line is drawn through the average y-position of the top pixel in each column. The top pixels of all columns are joined as a curve. Left boundary of the head is the leftmost intersection between the curve and the dashed line while the right boundary is defined by the rightmost
intersection point. Based on the assumption that ratio of face height to face width is in the range [1.1, 1.4], the lower edge of the bounding box is determined.

The vertical positions (y-positions) of the eyes, nose, and mouth are estimated by looking for peaks of the horizontal integral projection of the gradients in the face area. Eyes are searched within [40%, 60%] of the distance between the top and the chin. If the interval contains more than one maximum of horizontal integral projection, the principle used to differentiate the eyes from eyebrows and nose is that the eyebrows have approximately the same value of horizontal integral projection compared with the eyes while the nose has lower value. The nose is searched within [30%, 45%] of the distance between the eyes and the chin and then the mouth is searched within [20%, 50%] of the distance between the nose and the chin. If there are more than one or no maximums in the search interval for nose or mouth, the one closest to the middle of the interval is chosen.

Finally, horizontal positions (x-positions) of facial symmetry line and the two eyes are obtained by calculating the vertical integral projection of the differences between the intensity and gradients in each pixel. The x-position of the facial symmetry line is searched by looking for maximum point(s) of vertical integral projection within [30%, 70%] of the distance between the left and right face boundaries. With $\Delta x$ denotes 1/10 of the distance between the left and the right face boundaries and $x_{num}$ denotes the x-position of the facial symmetry line, the left eye is searched by looking for global minimum of vertical integral projection within $[x_{num} - 1.75\Delta x, x_{num} - \Delta x]$ while the right eye is searched by looking for global minimum of vertical integral projection within $[x_{num} + \Delta x, x_{num} + 1.75\Delta x]$. 

12