INDIVIDUAL 3D FACE SYNTHESIS BASED ON ORTHOGONAL PHOTOS AND SPEECH-DRIVEN FACIAL ANIMATION

Shiguang Shan[1], Wen Gao[1,2], Jie Yan[3] , Hongming Zhang[2], Xilin Chen[2]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080,China
Department of Computer Science, Harbin Institute of Technology, Harbin, China
Microsoft Research(China), Beijing, China
{sgshan, wgao}@ict.ac.cn, i-jieyan@microsoft.com, {hmzhang, xlchen}@cti.com.cn

ABSTRACT

In the paper, a methodology for individual face synthesis using given orthogonal photos is proposed. And an integrated speech-driven facial animation system is presented. Firstly, in order to capture given subject's personal facial configuration, a novel coarse-to-fine strategy based on facial texture and deformable template is proposed to localize some facial feature points in the image of frontal view. And the corresponding feature points in the profile are extracted by using polygonal approximation. Secondly, all these feature points are aligned to fit the generic 3D face model to a specialized one to reflect the given person's facial configuration. Then a multi-direction texture-mapping technique is presented to synthesize a lifelike personal face. Finally, muscle-based expression and lip-motion models are built up. All above technologies are integrated into a speech-driven face animation system. We are aiming at a MPEG-4 compatible video-driven face animation system.

1. INTRODUCTION

Given some facial images of one person, it is a great challenge to synthesize his personal virtual face realistically and animate the face to look, talk and behave like the person himself naturally. Nevertheless It is essentially important to cartoon industry, interactive VR, 3D interactive games, personal virtual agent, multi-modal human-computer interface and model-based coding as emphasized in newly standardized MPEG-4. In this field some effort has been done as in[1,2]. In this paper, we address all the related issues and solved the problems in our own way. An method for automatic individual face synthesis based on given orthogonal photos is proposed and an integrated speech-driven facial animation system is presented. Fig. 1 outlines the overview of our system.

The organization of this paper is as follows: we begin with the automatic facial feature extraction from two orthogonal photos, then proceed to describe the synthesis of 3D virtual face for the given subject. Then expression and lip-motion model is constructed. Finally, the integrated speech-driven face animation system is introduced and a short conclusion is given.

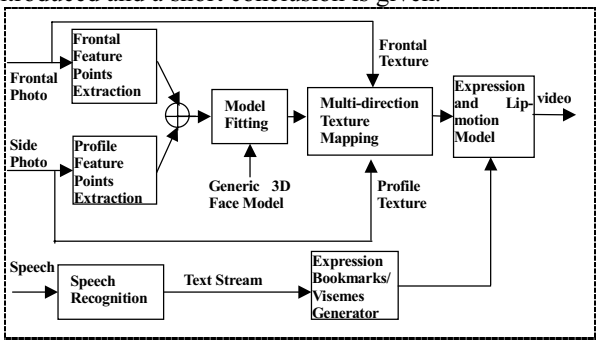


Fig. 1 System Overview

2.AUTOMATIC FACIAL FEATURE EXTRACTION

In order to fit a generic face model to a personal one, we pre-defined totally 46 feature points as illustrated in Fig.2, including 30 in the frontal view and16 in the profile. This section discusses the extraction of these feature points.

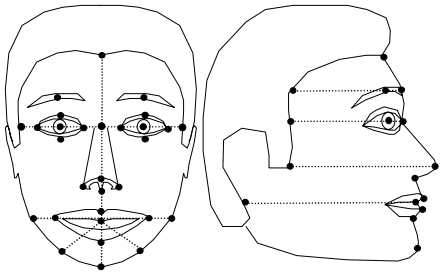


Fig. 2 Pre-defined Facial Feature Points

2.1 Extraction Of Frontal Facial Feature Points

Deformable template [4,5] is an effective method to extract the location and shape of salient facial organs such as eyes, mouth and chin, while it does have some drawbacks as highly dependence on the initial parameters and being subject to trapping into the local minimal. Moreover, it is too time-consuming. To partly solve these problems, a coarse-to-fine strategy is proposed.

2.1.1 Coarse Localization of the Facial Features

Based on the results of our three-level face detection system[8] and the observations that the two irises are the most salient features, the two irises are localized first. Then other organs are localized by integral projection.

1. Localization of the two irises The output of our face-detection system is a region containing all the face organs excluding most part of the hair. The region is then binarized by multi-threshold to a binary image, in which under most circumstances the two irises form two isolated 'black-islands'. It follows a morphological process in order to filter off those noise connected regions. Then in the resulting binary image the following rules are used to search a couple of isolated 'black-islands' denoted as RL and RR: (1) No other connected regions near below RL and RR. (2) The horizontal distance between the center of RL and RR is not too small. (3) The vertical distance between the center of RL and RR is small enough. (4) RL and RR is high-frequency region in the original face image. These rules are enough to find a couple of connected regions corresponding to the two eyes. Then Hough transform is used to search circles in the Canny edge maps of the RL and RR individually. The center of the circles are regarded as the centers of the irises.

2. Feature Localization based on integral projection Based on the observations that the nostrils and the lip form two gray valleys below the midpoint of the two irises, horizontal integral projection is adopted to localize the two organs. Firstly, the approximate region range is estimated by facial configuration a prior, in which the horizontal integral projection curve is calculated by sliding a window vertically. A local minimal is expected to appear at the point corresponding to the location of the mouth. The same strategy as used in the mouth detection is adopted to extract the nose features. Horizontal integral projection is firstly used to obtain the location of the nose. Then the position of the two nostrils is calculated by vertical integral projection. The nose-tip is located by searching for a high luminance point above the nostrils.

3. Key Feature Points detection The key feature points include the eye-corners, mouth corners and some points on the chin edge, etc, according to which template parameters can be well initialized. The corners are located by integral projection in the edge map. The points on the chin are obtained by analyzing the chin edge map.

2.1.2 Feature Shape Extraction Using Improved Deformable Template

Since the feature localization algorithm described above is merely based on local information and low-level processing, it is not accurate enough and therefore subject to noises. As is well known, being an algorithm that makes use of the global information, deformable template can extract features more accurately and is more robust to noises. Detailed algorithm can be found in [4]. The templates for the eye, mouth and chin in our system are illustrated in Fig.3. As to the cost function, we inherit the energy items measuring the goodness of fit between the template and the image properties such as the peak, valley and edge. And at the same time new items are introduced to measure the homogeneity of texture of the feature regions, which is expressed by the intensity and chromatic variance of the region. Additionally, in order to avoid the template deforming toward an unreasonable shape, an internal energy item and a punishment energy item are defined. All these items are combined to one cost function using different weights. Finally an optimal algorithm based on greedy algorithm and multi-epoch cycle is used to search for the minimum cost.

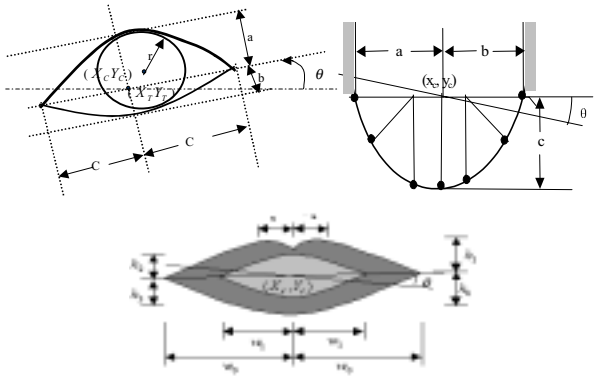


Fig. 3 Template for the eyes, chin and mouth

2.1.3 Experiments Results

Fig.4 illustrates the results of our feature extraction system. The original images are 256x256 with a typical face size of 130x140. On a Pentium III 450MHz computer, 2 seconds are needed for the whole procedure from detecting the face to exacting all the feature points.

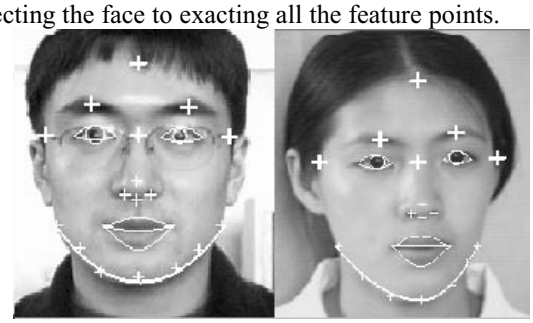


Fig 4. Result of Detecting Facial Features

2.2 Profile Feature Points Extraction

Profile feature points Extraction includes two main steps: the extraction of the profile outline and the localization of feature points. To simplify the problem, we put two constraints on the profile: (1) blue background (2) the subject being with neutral expression.

2.2.1 Extraction of the profile outline

Color information contains three components: intensity, hue and saturation among which hue can be used to represent skin color. Our experiments indicate that facial skin color clusters well in hue components. An appropriate hue threshold is selected by a moment-based threshold approach to compute a binary map in which the profile outline is extracted by Canny algorithm.

2.2.2 Location of profile feature points

Based on the observations that many feature points are turning points along the profile outline, a polygonal approximation method is used to detect them. The nose-tip is located by searching the rightmost point at the outline. And it divides the outline into two parts. For each segment a polynomial function is calculated by a curve-fitting algorithm. Then the extremal points can be extracted by searching points with *0-value* first order derivatives. In this way some feature points are located such as forehead, nose valley, upper lip, mouth point and lower lip. From these points and the geometric relations between the features points, other feature points can be easily estimated such as eye hollow-eyed point, chin point, lower mouth point. Fig.5 shows the results of the process.

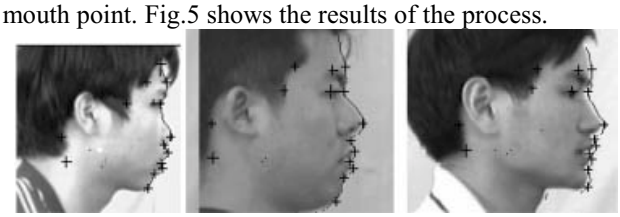


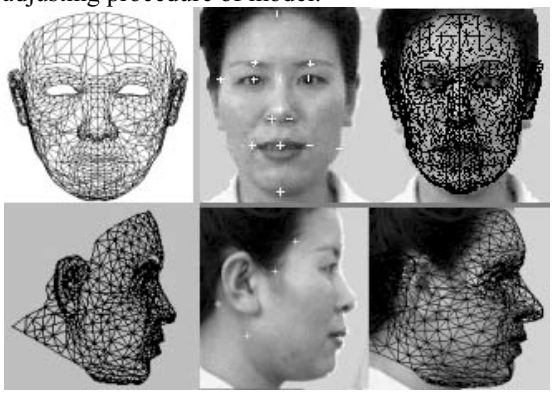
Fig 5 Profile feature points extracted

3. INDIVIDUAL 3D FACE SYNTHESIS BASED ON TWO ORTHOGONAL PHOTOS

Feature points automatically extracted as described in the previous section are sometimes not accurate enough. In order to rectify these possible errors, an interactive correcting mechanism is introduced. Then based on these feature points a generic 3D-face model is fitted to a personal one. And multi-direction texture mapping is proposed to synthesize a personal and lifelike face.

3.1 Fitting a generic 3D-face model to a personal one

Global and local transform are designed to fit the generic 3D-face model to a personal one. Global transform is used to adjust the global facial contour and the positions of the organs in the face. It is accomplished by scaling the coordinate values of each vertex of the model. The scaling factors are calculated according to the relations between the coordinates of the feature points in the generic face model and in the photos. While the local transform aims at adjusting the shape of each organ such as the eyes, eyebrows, mouth, nose and chin to fit the given persons' characteristic. Details can be found in [7]. Fig.4 illustrates the adjusting procedure of model.



(a) (b) (c)

Fig.6 Fitting a Generic 3D-Face Model to a Personal one (a) Generic 3D-Face Model (b) Feature Points Extracted (c) Personal 3D-face model

3.2 Multi-Direction Texture Mapping

Texture mapping provides a valuable technique for further enhancing the factuality of the synthetic face. A multi-direction texture mapping technique is presented to map appropriate texture to the surface of the 3D model. For each Bézier patch in the face surfaces, from which photo its texture is mapped depends on the whole normal direction of the Bézier patch. When the angle of the directional vector is less than 30 degree, frontal face image is used. Otherwise profile image is used. Fig.5 illustrates the synthetic face for a given person.

(A) (B) (C)

Fig 7. Given Person's Synthetic Face (a) Input Given Person's Front Image (b, c) Synthetic lifelike face

4. SPEECH-DRIVEN FACE ANIMATION

A vitural face is expected to be able to look, talk and behave like human beings naturally. In this section, we talk about how to animate the synthetic face to be expressive and 'talk' like a human being. To integrate all the technologies mentioned in this paper, an integrated speech-driven face animation system is presented.

4.1 Expression Synthesis

Expressions are driven by the motion of facial muscles. In our system facial expression is described as action units

(AUs) which reflect the contraction of some bunches of facial muscles. The expansion and contraction of each muscle vector reflect the motions of facial feature points and the motions' propagation to the related areas in the mesh. In our system the AUs for six basic expressions are defined individually according to physiological and anatomical analyses. Details can be found in [7].

4.2 Lip Motion Synthesis

In Chinese, speech unit that can be distinguished naturally is syllable. In general, one Chinese word is just one syllable. A syllable is composed of consonant and rhyme. When a syllable is pronounced, the lip-shape for of the consonant sustains very short and quickly switches to that of the rhyme. In Chinese phonetics, there are 19 consonants and 39 rhymes. Rhymes are classified into single rhymes, compound rhymes and nose rhymes. Since some phoneme's lip-shapes are very similar, we only build up seven typical lip-shape models corresponding to the phonemes: 'a', 'o', 'e', 'i', 'u', 'b', 'ng'. For any syllable in Chinese, its lip-motion is constructed by the combination of the seven basic lip-shapes. And the duration of each phoneme is carefully controlled.

4.3 Speech-driven Face Animation

An integrated speech-driven face animation system as described in Fig.1 is implemented, in which the speech recognition module developed from IBM Viavoice SDK is embedded. And the module named as "expression bookmarks/visemes generator" is used to extract affective meanings and visemes from the Chinese text stream, which is used to drive the expression/lip-motion model to generate lifelike face actions. Figure 8 illustrates some animation results.

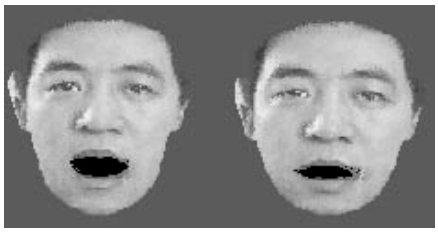


Fig8. Face 'Speaking' 'a' and 'e'

5. CONCLUSION

The presented method for automatic 3D individual face synthesis based on given orthogonal photos and speech-driven facial animation system work very well. The presented coarse-to-fine strategy can accurately localize the facial feature points in the frontal view while the polygonal approximation method can extract corresponding feature points in the profile. A generic 3D-face model is fitted to a personal one reflecting the given person's facial configuration. The presented multi-direction texture mapping technique greatly enhances the realism of the graphic face. And corresponding expression and lip-motion model are built up, by which the face can be animated to 'talk' like human beings.

ACKNOWLEDGMENT

This research is supported partly by Natural Science Foundation of China (No.69789301), National Hi-Tech Program of China(No.863-306-ZD03-01-2), and 100 talent foundation of Chinese Academy of Sciences.

REFERENCE

[1] Li-an Tang, and Thomas S.Huang Automatic Construction of 3D Human Face Models Based on 2D Images, *Proceedings of ICIP'96*, III, pp.467-470, 1996

[2] Y.C.Lee, D.Terzopoulos and Keith Waters. Realistic modeling for facial animation. In SIGGRAGH'95 Conference Proceedings, pp.55-62, Los Angels, 199

[3] P. Ekman and W. V. Friesen. Facial action coding system. *Consulting Psychologists Press Inc*., California, 1978

[4] A.L.Yuille, P.W. Hallinan and D.S. Cohen. Feature extraction from faces using deformable templates. *Int. Journal on Computer Vision* ,vol.8, pp.99-111, 1992

[5] X. Xie, R. Sudhakar and H. Zhuang On improving eye feature-extraction using deformable templates. *Pattern Recognition*, 27(6), pp.791-799, 1994

[6] C.J.Wu and J.S,Huang. Human face profile recognition by computer, *Pattern Recognition*, 23(3), pp.225-259, 1990

[7] Wen Gao, Jie Yan, Baocai Yin, Yibo Song, An individual facial image synthesis system for virtual human, *Proceedings of the Second International Conference on Multimodal Interface*, pp.20-25, 1999

[8] Wen Gao, Mingbao Liu. A hierarchical approach to human face detection in a complex background. *Proceedings of the First International Conference on Multimodal Interface*,pp.289-292, 1996