*I.J. Education and Management Engineering* **2011, 4, 59-66**
Published Online October 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2011.04.10

***Available online at http://www.mecs-press.net/ijeme***

# Automatic Estimation of Emotional Parameters for An Intelligent Avatar

Jing Yuan, Baochen Jiang, Menghan Sun

*School of Mechanical, Electrical & Information Engineering, Shandong University at WeiHai, WeiHai, China*

**Abstract**

This paper analyzes automatic estimating of emotional parameters from 2D photos based on the MPEG-4 rules. Each affective face image which obtained from a novel picture questionnaire is divided into three parts. The deformation function which calculates the displacements of vertices meshes within the influence of facial animation parameters are discussed. This paper develops geometric and morphological image processing methodologies to lifelike facial expression synthesis. Finally a 3D audio-visual avatar is created which gives the same expression corresponding to the input image. Experimental results are given to support our point of view. The experimental results show that our system properly applied in describing emotional facial expression.

**Index Terms:** Emotion parameters estimating; speech synthesis; image processing; MPEG-4



## 1.Introduction

A great deal of scientific research proved that the interaction of acoustic speech and visual speech enhances the understanding of the conversational content. According to Mehrabian's statistic for the relative impact of facial expressions and spoken words, total human communication equals to 7% verbal meaning plus 38% voice quality or prosody information, and 55% facial expressions [1]. During 30 years research on text to audio-visual speech synthesis, many research and efforts in both academia and industry have been brought to improve the quality of synthesized speech based on two measurements: intelligibility and naturalness. Major developments in two aspects: acoustic quality and visual quality. D. Jiang proposed an adaptation algorithm to predict the emotional prosody features, which was efficient when there were only small amount of training data [2]. Most research mainly pay attention on how to enhance the prosody of synthesized speech. Parke proposed the earliest geometric model [3, 4]. Geometric facial model composed of polygonal faces and vertices with 3D topological structure. Waters established a series of physical face models based on facial anatomy [5, 6]. With current state-of-the-art 3D modeling technology, both the modeling and animation effects are lifelike. Ekman proposed that the archetypal emotions correspond to distinct facial expressions [7]. He separated facial expression into different AUs (Action Units) from the angle of psychology. Escher parameterized model with MPEG-4

* Corresponding author.
E-mail address: yuanjing1985@gmail.com , jbc@sdu.edu.cn

definition, and extracted FAPs from image and video data [8]. These researches further improve the intelligent of the synthesized speech. After that, considering the shortage of verisimilitude, recent researches focus on define visemes of different languages. Z. Y. Wu et al. presented Facial Animation Parameters (FAPs) for Chinese visemes [9]. Nonetheless, audio-visual speech synthesis by no means considered as sentient as human. To this end, ideally, computer is expected to have the capable of producing emotive expression. The difficulty lies in the fact that how to calculate emotional parameters. However, research into this particular topic is rare.

This paper makes a novel picture questionnaire survey. Respondents are separated into different groups according to personality, and then asked to see some pictures which may stimulate different moods. We take photos with facial expressions that generated in natural condition, and use image processing methods to extract feature points which relate to emotion. Emotional parameter model is built based on statistic theory. Finally, emotional parameters and FAPs are calculated.

## 2.Visual Emotional Parameter Selection

### *A. Facial Model*

A 3D avatar with VRML format is created using the software Autodesk 3ds Max. A VRML format files is stored as binary text which is convenient to read and access. Furthermore, this is the general format for virtual reality, so it is easy to expand its applicable range. Our 3D model has 7 Meshes, including 6435 vertices and 12280 faces in total. The model specifies positional coordinates for rendering, normal coordinates for lighting effects as well as texture coordinates for texture mapping. Each face model is composed of meshes of upper teeth, lower teeth, tongue, throat, left eye, right eye and head. Each mesh which filled with triangles are comprise of connection of positional coordinates.

### *B. Expression Parameter Selection*

We adopt the MPEG-4 Facial Animation standard for selecting parameters, and then use the theory of Ekman to define and render these parameters [10]. The MPEG-4 standard defines 84 Feature Points (FPs) to describe different parts of the face model. The different parts are eyebrows, eyes, nose, mouth, tongue, teeth, etc. Facial Definition Parameters (FDPs) are defined to calibrate the shape of the head, and FAPs are defined to describe the deformations with respect to the neutral face.

We choose 28 relevant FPs for three parts in face. FP 4.2, 4.4 and 4.6 define the camber of right eyebrow, FP 4.1, 4.3 and 4.5 are corresponding points on the left eyebrow. Fig. 1 shows 14 FPs on eyes and 8 FPs on month we used. FP 2.2~2.5 describe the contour of inner lip, and FP 8.1~8.4 give the contour of outer lip.


Figure1 Ps used on eyes and mouth

Some FAPs we concerned are listed in TABLE I. FAP19 to FAP22 define the displacement of eyelids. FAP31 to FAP38 define camber changes of eyebrows. Other FAPs decide inner and outer lip edges. All of above FAPs can be calculate from position of FPs, our work is to track their movements without any advance markers.

TABLE I. FAPS DESCRIPTION TABLE

| Feature Animation Points Number and Describe | | | |
|---|---|---|---|
| *FAP#* | *Text description* | *FAP#* | *Text description* |
| 4 | lower_t_lip_i[a] | 33 | raise_l_m_eyebrow |
| 5 | raise_b_midlip_i | 34 | raise_r_m_eyebrow |
| 6 | stretch_l_cornerlip_i | 35 | raise_l_o_eyebrow |
| 7 | stretch_r_cornerlip_i | 36 | raise_r_o_eyebrow |
| 12 | Raise_l_cornerlip_i | 37 | squeeze_l_eyebrow |
| 13 | Raise_r_cornerlip_i | 38 | squeeze_r_eyebrow |
| 19 | close_t_l_eyelid | 51 | lower_t_midlip_o |
| 20 | close_t_r_eyelid | 52 | raise_b_midlip_o |
| 21 | close_b_l_eyelid | 53 | stretch_l_cornerlip_o |
| 22 | close_b_r_eyelid | 54 | stretch_r_cornerlip_o |
| 31 | raise_l_i_eyebrow | 59 | raise_l_cornerlip_o |
| 32 | raise_r_i_eyebrow | 60 | raise_r_cornerlip_o |

a. t- top; b- bottom; l- left; r- right; i- inner; o- outer.

## 3. Feature Point Tracking

### *C. Facial Expressive Image Acquisition*

Previous studies tend to let respondents act 7 basic emotions: natural, happy, angry, sad, surprise, fear and disgust [9]. Then use interpolation algorithm to calculate the intermediate expressions. These image frames finally become composite facial animation. However, this limits the expression not only exaggerate but also lack of individuality. According to Ekman's facial muscle movement theory, we try to extract micro expressions. First we choose 20 respondents and divide them into different groups on the basis of personality. Group A represents "lively and cheerful" and group B represents "quiet and introverted". After that we show them 6 pictures which present happy, moving, bloody etc. content respectively. Using this psychological hint method, people perform unconscious expressions. In this way, naturalness of emotion synthesis is enhanced and the input of the emotion synthesis system is simplified.

### *D. Geometric Image Registration*

Considering head movement along horizontal and vertical directions. We need to do image geometric registration before extracting FPs. Image registration aims in two or more images alignment in the same scene. Consequently each image has approximate size and in the same position as far as possible. Generally we take benchmark image as reference, and find out a suitable relationship of space transformation through some fiducial points. In order to eliminate the translation errors and rotation errors, three variables should be calculated: $\Delta x$, $\Delta y$, and rotation angle $\theta(n)$. Let $P_i(n)$ signifies FP $i$ of frame $n$, $y_l$, $y_r$ signify the $y$ coordinates of left and right nostrils. As in (1) ~ (3), $\Delta x$, $\Delta y$, and $\theta(n)$ are obtained. An example of geometric image registration result is showed in Fig. 2.

$$\Delta x(n)=\frac{1}{2}\sum_{i=1}^{2}\left(x_{Pi}(0)-x_{Pi}(n)\right), \tag{1}$$

This work is supported by the science and technology progress program of Shandong Province (2005GGA10111)

$$\Delta y(n)=\frac{1}{2}\sum_{i=1}^{2}\left(y_{Pi}(0)-y_{Pi}(n)\right), \tag{2}$$

$$\theta(n)=\arccos\left[\frac{y_r(n)-y_l(n)}{y_r(0)-y_l(0)}\right]. \tag{3}$$

.

Figure2 Image Registration result.

*E. FPs Track*

In order to extract the contour of different organ: brows eyes and mouth, we separated face image into three parts, and use image processing method respectively. Difficult is how to express mouth shape around which color is rich.

- First we change pixels in brows part from RGB to Gray, then use automatic threshold 0.38 to obtain binary image. Let the binary image corroded by $3\times3$ Square structure elements image corrosion. As shows in Fig. 3, we have obtained the contour of brows.

Figure3 Gray brows (left) and Binary Contour (right)

- Previous works on separating lips with the skin around achieved good results. Wang used Fisher transform to do best binary classification [11]. Chen employed GMM to calculate probabilities [12]. In order to simplify working, we adopt edge detection method to extract FPs on eyes and mouth. First use smoothing filter to reduce noise because edge detection easily affected by noise. Then use image binaryzation and Sobel operator to detect vertical and horizontal edge. Fig. 4 and Fig. 5 indicate the edge detect results.

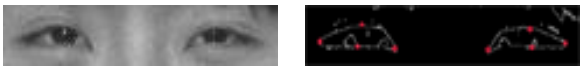


Figure 4 Gray eyes (left) and canny operater edge detection (right)

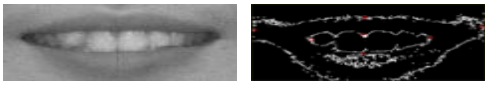


Figure 5 Gray mouth and Sobel operater edge detection

## 4. Parameter Estimation

### *F. FAP Value Calculation*

The FAPs is a group of facial animation parameters. FAP is independent of the facial model while FAPU depends on the facial model. The definitions for 6 FAPUs are as following: *IRISD=IRISD0/1024, ES=ES0/1024, ENS=ENS0/1024, MNS=MNS0/1024, MW=MW0/1024, AU=$10^{-5}$*. As mentioned above, FAP values decide the animation of the 3D avatar we made. When a new face image sequence is inputted into the facial expression synthesis system, the current nature position of each FP can be extracted corresponding to the previous one. We calculate the difference between them. On the basis of mouth wide in nature condition (MW) and mouth-nose distance (MNS) in nature condition, the corresponding FAP can be estimated as in (4) ~ (7) [10]:

- FAP # 4, 5, 12, 13, 51, 52, 59,60:

$$Fap = \left(P_i'.y - P_i.y\right)/MNS \tag{4}$$

- FAP # 6, 7, 53, 54:

$$Fap = \left(P_i'.x - P_i.x\right)/MW \tag{5}$$

- FAP # 19~22:

$$Fap = \left(P_i'.y - P_i.y\right)/RISD \tag{6}$$

- FAP # 31~38:

$$Fap = \left(P_i'.x - P_i.x\right)/ES \tag{7}$$

$P_i'.x$ and $P_i.x$ present current FP's $x$ coordinate and the corresponding $x$ coordinate in neutral expression condition. $P_i'.y$ and $P_i.y$ present current FP's $y$ coordinate and the corresponding $y$ coordinate in neutral expression condition.

### *G. Emotional Parameters Calculating*

What's important in animating a facial model is to use the specified FAP values to generate the desired expression.

MPEG-4 puts forward an algorithm to compute the displacements of the vertices within the influence of FAPs.

Let $P_m$ be the coordinate of vertex $m$ in benchmark expression. $P_m^{'}$ is the coordinate after $m$ changes its position according to FAP. $D_{m,K}$ is the motion factor in the $k$th segment, $P_m^{'}$ can be given by following steps:

Assume the range of FAP is divided into *max+1* segments, $[I_0, I_1]$, $[I_1, I_2]$, $[I_{\max}, I_{\max+1}]$. Assume present FAP lies in $[I_j, I_{j+1}]$, and *0* lies in $[I_k, I_{k+1}]$, $0 \le j,k \le \max$. If FAP or *0* lies on one of the boundaries, we can choose freely from the adjacent segments of the boundary as the segment for this FAP or *0*. And we can compute $P^{'}_{m}$ as follows:

- *if* $j > k$:

$$\begin{aligned} P'_m = FAPU \times \big(&\left(I_{K+1} - 0\right) \times D_{m,k} + \left(I_{k+2} - I_{k+1}\right) \times D_{m,k+1} + \\ &\ldots + \left(I_j - I_{j-1}\right) \times D_{m,j-1} + \left(FAP - I_j\right) \times D_{m,j}\big) + P_m \end{aligned} \tag{8}$$

- *if* $j < k$:

$$P'_m = FAPU \times \left( \left( FAP - I_{j+1} \right) \times D_{m,j} + + \left( I_{j+1} - I_{j+2} \right) \times D_{m,j+1} + \right.$$
$$\left. \ldots + \left( I_{k-1} - I_k \right) \times D_{m,k-1} + \left( I_K - 0 \right) \times D_{m,k} \right) + P_m \quad (9)$$

- if $j = k$:

$$P'_m = FAPU \times FAP \times D_{m,j} + P_m \quad (10)$$

If the range of FAP is divided into only one segment, i.e., the motion of the vertex is strictly linear. We can use (10) to compute $P_m^{'}$.

## 5.Experiment Results

As discussed above, Fig. 6 demonstrates frame structure of our facial emotional parameter estimating process. From a color facial image sequence taken by CCD camera, face area is clipped as the input of the system. Then the input image is divided into three facial organ parts which are brow part, eye part and mouth part. Geometric and morphological image processing methods extract the contour profile of these three parts. We manual annotate 28 FPs mentioned using an interactive interface. Finally, feature animation point values are calculated, and an expressive lifelike digital character is created.

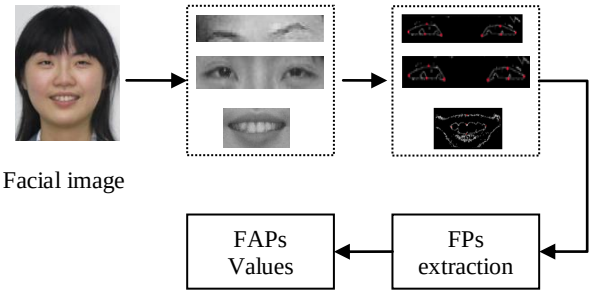
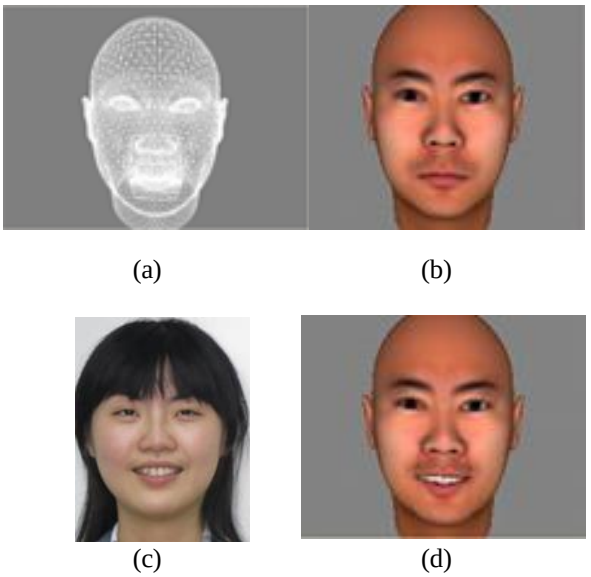



Figure 6 Parameter Estimating Frame Structure

Fig. 7 (a) shows the 3D facial mesh we made, (b) is face with natural expressive, (c) is our input image, and (d) is synthesized result.



(a) (b)

(c) (d)

Figure 7 Experiment result.

## 6.Conclution

This paper developed image processing methods for visual emotional parameter estimate. Our research aim is to analyze a simple and effective algorithm for automatic estimation of facial emotional parameters. A picture survey which adopts psychological hint method is taken to stimulate different moods. 28 Feature Points are tracked form each facial image through image geometric and morphological transform algorithms. According to the displacements of these FPs, three deformation functions are built. Not only 6 basic expressions but also some intermediate expression can be regenerated.

Ultimately, future work focus on more complex deformation functions, in order to create lifelike man-machine interaction interface. At this point, automatic generation of parameters for a 3D face model from 2D expressive picture by mathematics or other theoretical methods need been enhanced.

## References

[1] A. Mehrabian, "Communication without words," Psychol. Today, vol.2, pp. 53–56, 1968.
[2] D.N. Jiang, W. Zhang, L. Shen, L.H., Cai, "Prosody analysis and moeling for enottonal speech synthesis", International Conference on Acoustics, Speech and Signal Processing (ICASSP-2005), pp. 281-284, 2005.
[3] F.Parke., "Computer generated animation of faces", ACM Annual Conf., Boston, Massachusetts, United States, pp. 451-457, 1972.
[4] F. Parke, "A model for human faces that allows speech synchronized animation", Journal of Computers and Graphics , vol.1, pp. 1-4, 1975.
[5] K. Waters , D. Terzopoulos, "A physical model of facial tissue and muscle articulation", 1st Conf. on Visualization in Biomedical Computing , Atlanta , USA , pp, 22-25, May 1990.
[6] D. Terzopoulos , K. Waters. "Analysis and synthesis of facial image sequences using physical and anatomical models ", IEEE Trans on Pattern Analysis and Machine Intelligence, vol.15, pp. 569 – 579, Jun 1993.

[7] P. Ekman , W. V. Friesen, "Facial Action Coding System", Palo Alto , CA : Consulting Psychologist Press , 1978

[8] M. Escher , T. Goto , S. Kshirsagar , et al. , "User interactive MPEG-4 compatible facial animation system", International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging ( IWSNHC3DI'99) , Santorini , Greece , 1999.

[9] Z.Y. Wu, S. Zhang, L.H. Cai, and H.M..Meng, "Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a Three-dimensional Avatar", Proc. Int. Conf. on Spoken Language Processing, pp. 1802-1805, 2006.

[10] Motion Pictures Expert Group, ISO/IEC 14496-2:1999/Amd. 1: 2000(E). International Standard, Information.Technology – Coding of Audio-Visual Objects. Part 2:Visual; Amendment 1: Visual Extensions.

[11] R. Wang, W. Gao, J. Y. Ma, " An approach to robust and fastlocating lip motion", 3rd Conf. on Multimodal Interfaces,Heidelberg, vol.1948, pp. 332-339, 2000.

[12] T. Chen. "Audiovisual speech processing", IEEE Signal Processing Magazine ,vol.18, pp. 9-21, 2001.