A DYNAMIC VISEME MODEL FOR PERSONALIZING A TALKING HEAD

WANG Zhiming, CAI Lianhong, AI Haizhou

Department of Computer Science and Technology, Tsinghua University, State Key Laboratory of Intelligent Technology and Systems Beijing 100084, PR China

ABSTRACT

Personalizing a talking head means not only to personalize a head model but also to personalize his talking manner. In this paper, we propose a dynamic viseme model for visual speech synthesis that can deal with co-articulation problem and various pauses in continuous speech. Facial Animation Parameters (FAPs) defined in MPEG-4 are estimated according to the tracked feature points from two orthogonal views via a mirror setup. Individual talking manner described by model parameters is learnt from FAP data to implement a personalized talking head.

1. INTRODUCTION

To synthesize a talking head is an important issue in human computer interaction to enhance the intelligibility of speech and to make the computer interface more friendly. There are mainly two approaches to make a talking head. One is model-based that is a parameter control approach, which builds a 2D or 3D face model and adjusts its control parameters to mimic various facial expression and mouth movement [1][2][3]. The other is sample-based that is data driven approach, which builds a talking head by selecting a sequence of video triphones or a group of facial parts from a sample database [4][5]. In order to personalize a talking head, the model-based approach usually only changes the head model without considering individual talking manner so as to not so realistic. In the case of sample-based approach it can make the talking head more realistic and more like a particular individual with the expense that it need rebuild the sample database, which is extremely a hard work because a great amount of primitive training data should be acquired. In this paper we follow the model-based approach to build a personalized talking head through a parameter learning procedure based on the tracked feature points from two orthogonal views via a mirror setup. In this way we make the personalizing procedure can be easily implemented with both personalized head model and reasonable individual talking manner.

This paper organizes as follows. Section 2 describes the FAP extraction method, section 3 gives the dynamic viseme model, section 4 illustrates the personalizing talking head procedure, and finally section 5 gives the conclusion.

2. EXTRACT FAPS FROM ORTHOGONAL VIEW

In accordance with MPEG-4 standard, we describe the face model by Facial Animation Parameters (FAPs) in MPEG-4. In order to extract 3D FAPs from video automatically, we use a mirror to acquire two orthogonal views frontal and profile simultaneously as in Fig.1. The 3D FAPs are estimated from the tracking results of feature points in both views.

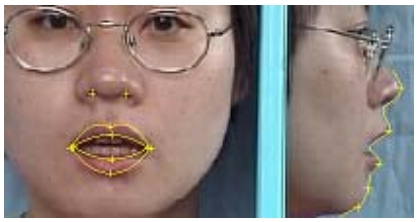

Fig.1 Frontal and profile views

In frontal views, the nostrils and the outer lip contour are to be tracked, and then the inner lip parameters are estimated based on the outer lip parameters. In the profile view, those points represent the nose tip; the protrusion of upper and lower lip, the thrust of jaw, and the openness of jaw are to be located. The correlative FAPs are estimated from the positions of these feature points.

In the literature, there are many methods to locate the feature points on face [6][7][8] but the precision is to be improved. We develop a multi step method to locate the mouth contour robustly and precisely.

2.1 Frontal view feature points tracking

The region of nostril is estimated according to the previous frame, and the corresponding image region is transformed into binary image by a proper threshold. The two nostril points are the barycenter of the left and right black pixels area as shown in Fig.2 that the upper shows the threshold result, and the lower shows the original image and located nostril points.

To track the lip contour precisely, which is the most flexible part in the face area, is very difficult. To locate the

Fig 2 Nostril

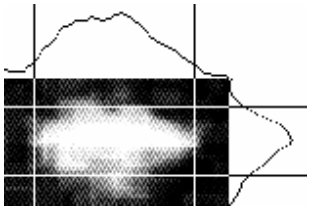

Fig 3(a) Fisher transfer projections of mouth area

Fig 4 left and right corner detector

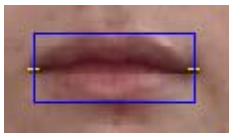

Fig 3(b) Mouth Located

exact positions of the two mouth corners are the key to the tracking precision. We work out the following procedure:

(1). Coarsely position the ROI (Region Of Interest) of mouth according to previous frame, and then compute a Fisher transfer in this region to make the mouth contour more evident;

(2). Project the area vertically and horizontally as shown in Fig. 3(a). By choosing proper thresholds applied to the projection curves the mouth rectangle is determined. The thresholds can be decided based on heuristics about mouth area projections that the more flat the vertical integral curve (mouth is nearly close), the higher the horizontal threshold value to guarantee smaller height, and vice versa.

(3). Find left and right mouth corners along left and right vertical edges in the rectangle area by using two special corner detectors shown in Fig.4. The result is shown in Fig. 3(b).

(4). Having located the mouth rectangle and the mouth corners, the deformable template method [6] is used to track the outer lip contour as shown in Fig.5.

The template consists of two parabolas (inner lip) and three quartics (outer lip), which can be described by following equations:

$$y = h(1 - \frac{x^2}{w^2}) \qquad (1)$$

$$y = h(1 - \frac{x^2}{w^2}) + 4q(\frac{x^4}{w^4} - \frac{x^2}{w^2}) \qquad (2)$$

Inner lip contour is difficult to locate even by our eyes, so we estimate inner lip contour from outer lip contour based on statistical knowledge as mentioned in [8] according to:

$$W_i = a_1 * W_o + b_1 * h_1 + c_1 * h_2 + d_1 * q_1 + f_1 * q_2 + e_1 \qquad (3)$$

$$h_3 = a_2 * W_o + b_2 * h_1 + d_2 * q_1 + e_2 \qquad (4)$$

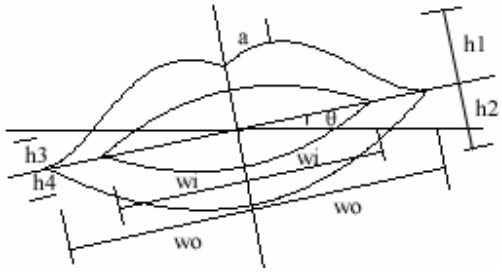

Fig 5 Deformable template

$$h_4 = a_3 * W_o + c_3 * h_2 + f_3 * q_2 + e_3 \qquad (5)$$

Where $W_i, W_o, h_1 \sim h_4$ are shown in Fig. 5. q_1, q_2 are parameters in equation (2) for upper and lower outer lip respectively. After measuring these parameters on more than 6 pictures manually, we can estimate these mapping coefficients $a_i \sim f_i (i = 1,2,3)$ by least square estimation. Certainly, the more pictures we measured, the more precise the coefficients would be.

2.2 Profile view feature points tracking

The face area in the image of profile view is first separated from its background according to color difference since in our experiments the blue background is used that is clearly different to face area. And then the profile contour of face is extracted.

From the nose to the jaw, there are three evident maximum horizontal values that correspond to nose tip and protrusion of upper and lower lip separately. Points for thrust of jaw and openness of jaw may be not obviously, the point for openness of jaw can be found based on the maximum second derivative in corresponding range. After deciding the point for openness of jaw, we can find the point for thrust of jaw through finding the point in the edge contour where its tangent is vertical or most close to vertical. If there is no such a point, we use the middle between the point for openness of jaw and that protrusion of lower lip instead. Fig.1 also shows an over all tracking result, includes all the feature points that are located in frontal view and profile view.

Finally, we should point out, the automatic tracking method can't track all the feature points in all frames precisely. Whenever the automatic tracking result is failed to be accurate, the tracking will be paused and the correct points will be manually input by mouse and then go on.

3. DYNAMIC VISEME MODEL

Static viseme in MPEG-4 is defined as follows: Viseme is the physical (visual) configuration of the mouth, tongue and jaw that is visually correlated with the speech sound

corresponding to a phoneme [9]. A static viseme could be show by a still human face picture. But when we pronounce a phoneme, the movement of our organ is more like a dynamic process rather than a static state. So we introduce a concept of dynamic viseme, which represent the whole process of the physical organs movement during the pronouncing of a given phoneme.

Like Cohen's co-articulation model [10], the dynamic viseme model is composite of dominance and parameter value. Unlike Cohen's model, we define two special silence dominance functions before and after each phoneme. Each model consists of three dominance functions: main viseme dominance, left and right silence dominance function.

Suppose D_{sp} is the dominance of lower FAP *p#* of phoneme *s*, then

$$D_{sp} = \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau|}, \quad if \ \tau \geq 0 \tag{6}$$

$$D_{sp} = \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau|}, \quad if \ \tau < 0 \tag{7}$$

α_{sp} 、$\theta_{\leftarrow sp}$ and $\theta_{\rightarrow sp}$ are positive parameters correlated with *s and p*. α_{sp} is the magnitude of the dominance function of FAP *p#* of phoneme *s*, $\theta_{\leftarrow sp}$ and $\theta_{\rightarrow sp}$ represents the rate parameter forward and backward. $\tau = t_{cs} - t_{osp} - t$, where *t* is the running time, t_{osp} gives the time offset from the center of segment *s* for the peak of dominance for FAP *p#*, t_{cs} is the time center of phoneme *s*.

The left silence dominance (from silence to sound) is given by following formula:

$$D_{lsp} = \alpha_{lsp} e^{\mathrm{sgn}(\tau)\theta_{lsp}|\tau|} \tag{8}$$

$\tau = t_s - t_l - t$, t_s represents the start time of following phoneme, t_l represents the time offset from the center of left silence to the start time of following phoneme. Similarly, the right silence dominance (from sound to silence) is given by following formula:

$$D_{rsp} = \alpha_{rsp} e^{-\mathrm{sgn}(\tau)\theta_{rsp}|\tau|} \tag{9}$$

$\tau = t_r - t_e - t$, t_e represent the end time of last phoneme, t_r represent the time offset from the center of right silence to the end time of following phoneme.

The final FAP *p#* for a whole process phoneme is decided by:

$$F_{sp}(t) = \frac{D_{sp}(t) \times T_{sp}}{D_{sp}(t) + D_{lsp}(t) + D_{rsp}(t)} \tag{10}$$

Where *Tsp* is the value of FAP *p#* of static viseme for phoneme *s*.

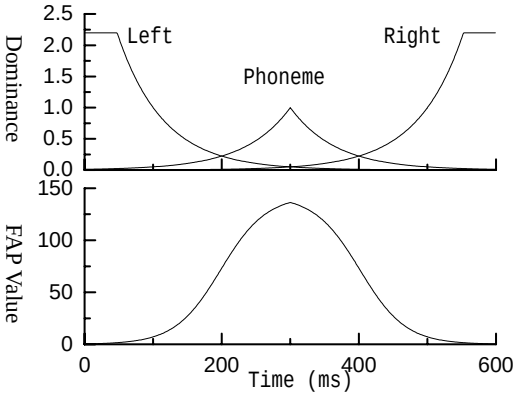

Fig 6 Dynamic viseme model

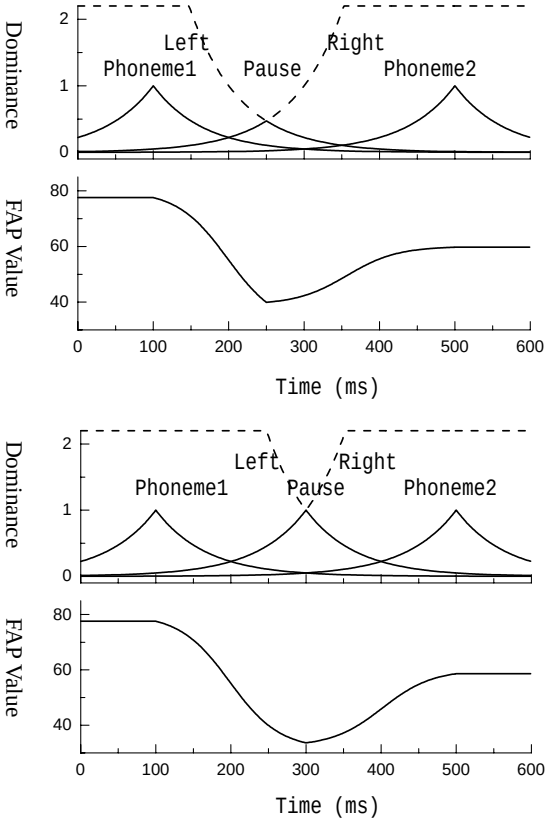

Fig 7 Different pause in continuous speech

The whole dynamic viseme model is shown in Fig.6. In continuous speech, pause between phonemes is formed by the intersection of right silence dominance of last phoneme and left silence dominance of following phoneme. This intersection model solved the problem of different transition parameters for different time interval between phonemes, as show in Fig.7. When there is no pause between two phonemes, this intersection model disappeared automatically.

4. PERSONALIZING A TALKING HEAD

With a new individual, we need to record the talking process of several basic phonemes, extract 3D FAP data, and then learn model parameters or adjust the model parameters quickly from one person to another. With these model parameters, we can build a talking head with the new speaker's talking manner.

We define 19 basic static viseme in Chinese based on both the phonetics knowledge and experiment data, as show in Table 1.

Table 1 Basic static viseme in Chinese

Initial	Final
b, p, m	a, ang
f	ai, an
d, t, n, l	ao
g, k, h	e, eng
j, q, x	ei, en
zh, ch, sh	er
r	i
z, c, s	o
	ou
	u
	ü

The whole procedure to personalize a talking head with his talking manner is as followings:

(1). Record the whole process of pronouncing the basic static viseme.

(2). Extract the basic static viseme picture by audio information, track outer lip contour and measure inner lip parameters on at least 6 pictures manually. Estimate the linear mapping coefficients from outer lip parameters to inner lip parameters by least square estimation.

(3). Track the whole video and extract lower FAP data for every viseme.

(4). Using the sum of square error between FAP data extracted from video and FAP data created by the dynamic viseme model as a criterion, we can learn parameters of dynamic viseme model from new FAP data by gradient descending method, or we can just adapt the general parameters according to the new FAP data.

(5). Using the new dynamic viseme model parameters to synthesis talking head with individual talking manner.

5. CONCLUSION

In this paper, we propose a dynamic viseme model for visual speech synthesis that can deal with co-articulation problem and various pauses in continuous speech. Facial Animation Parameters (FAPs) defined in MPEG-4 are estimated according to the tracked feature points from two orthogonal views via a mirror setup. Individual talking manner described by model parameters is learnt from FAP data to implement a personalized talking head. In this way, we could obtain different parameters for different people that correspond to particular talking manner of each individual. With a personalized 3D face model in real texture, we can realize a true personalized talking head.

6.REFERENCES

[1] Chen, L.S., Huang, T.S. and Ostermann, J., "Animated talking head with personalized 3D head model," IEEE First Workshop on Multimedia Signal Processing, pp. 274 –279,1997.

[2] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," Signal Processing: Image Communication, Special Issue on MPEG-4, vol. 15, pp. 387-421, Jan. 2000.

[3] Shiguang Shan. Wen Gao. Jie Yan. Hongming Zhang and Xilin Chen, "Individual 3D face synthesis based on orthogonal photos and speech driven facial animation", ICIP2000, vol. 3, pp. 238 -241, 2000.

[4] Bregler. C, Covell. M. and Slaney. M., "Video Rewrite: Driving visual speech with audio," Proc DIGGRAPH97, pp. 353-360, ACM SIGGRAPH, July 1997.

[5] Cosatto, E.; Graf, H.P., "Photo-realistic talking-heads from image samples", IEEE Transactions on Multimedia, vol. 2, pp. 152 –163, Sept. 2000.

[6] Rabi, G.; Si Wei Lu, "Energy minimization for extracting mouth curves in a facial image," Intelligent Information Systems, IIS '97. Proceedings, pp. 381 –385, 1997.

[7] Jin-Woo Kim, Munjae Song, Ig-Jae Kim, Yong-Moo Kwon, Hyoung-Gon Kim and Sang Chul Ahn, "Automatic FDP/FAP generation from an image sequence," The 2000 IEEE International Symposium on Circuits and Systems, ISCAS 2000, vol. 1, pp. 40-43, 2000.

[8] Rui Wang, Wen Gao and Jiyong Ma, "An approach to robust and fast locating lip motion," Third International Conference on Multimodal Interfaces, pp. 325-331,2000.

[9] International standard, Information technology -Coding of audio-visual objects-Part 2: Visual; Amendment 1: Visual extensions, ISO/IEC 14496-2: 1999/Amd.1: 2000(E).

[10] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," Models techniques in computer animation, Tokyo Springer-Verlag, pp. 139-156, 1993.