Text-driven Automatic Frame Generation using MPEG-4 Synthetic/Natural Hybrid Coding for 2-D Head-and-Shoulder Scene

Chun-Ho CHEUNG & Lai-Man PO

CityU Image Processing Laboratory, Department of Electronic Engineering,
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
Tel: (852) 2788 9902 Fax: (852) 2788 7791 Email : chcheung@ee.cityu.edu.hk

***Abstract* —** In this paper, we propose a facial modeling technique based on the MPEG-4 Synthetic/Natural Hybrid Coding for automating frame sequence generation of a talking head. With the definition and animation parameters on a generic face object, the shape, textures and expressions of an adapted frontal face can generally be controlled and synchronized by the phonemes transcribed from plain text. By this developed facial modeling technique, it increases the intelligibility of an non-verbal facial communication for potential audiovisual lip-synch application on news reporting, lip-reading for the hearing-impaired or the deaf, virtual meeting through internet, and Story Teller on Demand (STOD).

1. Introduction

A facial animation system, regardless of the type of character, extremely realistic or cartoon-like, is able to animate the desired facial motion as well as to reuse the animation and expression libraries. The most impressive facial modeling is to extract the realistic canonical facial model by a 3-D scanner, which results in high computational load during facial animation, texture mapping and rendering for the 3-D polygons [3], especially for B-Spline Modeling [9]. This highly time consuming 3-D rendering is not suitable for real-time interactive MPEG-4 system. Therefore, the system is designed in two-dimensional space with triangular mesh model on a multimedia PC platform. Our facial modeling technique can be regarded as 2-D Synthetic/Natural Hybrid Coding (SNHC) [1,2] with frames generated from visemes that are synchronized by phonemes.

The coming MPEG-4 standard involves the combination of natural and synthetic data, both for video and audio. It allows freedom of interactivity with the individual objects, such as combining a synthetic talking head onto a prior known stationary scene of conference room, rather than at the level of the composite video frames. Fig.1 shows the main idea how the MPEG-4 SNHC system works. It supports different languages, segmenting type, segment duration, amplitude contour and pitch of the chosen face object. Users can generate their desired talking face objects to be displayed on other hosts through internet during the virtual meeting (by transmitting Facial Definition Parameters(FDP)/Facial Animation Parameters (FAP)/phonemic parameters at phrase A). Besides, the SNHC system can produce video clips with the talking head as TV news production and STOD for broadcasting by either transmitting the required parameters or the encoded video at phrase A or B, respectively.

2. Lip-synch effect and Phonetics

The segmenting type can be syllable, intonational phrase or phonetics. Since human speech of any language can be decomposed into their shortest representative phonetics set, as shown in Fig.2, and thus lip/facial synchronization can be achieved. By using US English as our custom-made rule-based text-to-phoneme (RB-TTP) transcription engine [10], the plain text will be transcribed into orthographic phonetic symbols, listed in Table 1, according to International Phonetics Alphabets (IPA)/SAMPA computer readable phonetic alphabet. Some phonemic descriptions are associated with more than one viseme called diphthongs - i.e. shorter duration at the beginning of the utterance and longer at the end, e.g. /AY/, /AW/, /OY/.

3. The Face Object

Our generic wireframe facial model with a neutral expression shown in Fig.3(a) is based on the frontal projection of modified CANDIDE 3-D wireframe model description [7,8]. This extended 2-D triangulated mesh consists of 144 triangular polygons and 85 vertices which can sufficient for representing the 2-D frontal human face, as shown in Fig.3(b), as well as cartoon-like character.

3.1 Facial Definition/Animation Parameters

The FDP set is used to describe the different features among various face objects. The complete face models of a selected frontal view speaker can be obtained by masking (facial adaptation) [8]. For the facial animation, expressions, emotions and lip co-articulation for speech pronunciation, there are viseme parameters chosen from a set of visemes. Table2 shows 18 visemes (*fap1 – viseme_select (6bits)*) for classifying the mostly used 40 US IPA/SAMPA phonemes, as shown in Table 2. The visemes can also be regarded as the Associated Facial Action Units (AFAUs) corresponding to the phonemic pronunciation [4-6]. There are some random facial expressions for making the speaker face much more alive and more pleasant other than visemes. For examples, blinking the eyes and raising the eyebrows. Besides viseme parameters (*fap1*), there are facial expression parameters (*fap2 – expression_select(4bits)* + *expression_intensity (4bits)*) for facial expression, like joy, sadness, anger, fear, disgust and surprise, in a range of intensity levels. Other expressions, like excited, tense and sarcastic, can also be added in future to a maximum number of 16. A neutral face is a result of 0 intensity while the corresponding maximum value of 15 gives the most exaggerated expression, as shown in Fig.4.

4. Methodologies

The transcribed phonemes will be used to generate the corresponding speech and to synchronize with the corresponding visemes at the current selected expression. Once the key frames is generated, the frame sequence will be produced by undergoing temporal bilinear interpolation on the visemes for the in-between frames or tweens, digital image wrapping technique on the synthesized tweens' meshes and finally texture mapping from the original frontal face object.

4.1 Temporal Bilinear Interpolation

We assume NTSC frame rate (30fps) for animation and there is no acceleration between the corresponding vertices. Thus, we can apply a linear estimation by Equ.1 for estimating the i-th vertex of the j-th tween (t_{ij}) from the vertex of the starting frame (src_i) stepping forwards with equal temporal step size to the corresponding vertex in the final frame (dst_i).

$$t_{ij} = \frac{src_i \cdot (n+1-j) + dst_i \cdot j}{n+1}, \qquad \text{(Eq.1)}$$

where $t_{ij} \in \{F_1, F_2, ..., F_n\}$, *i*-th vertex is $[x_0, x_1, \ldots, x_K]$ of K-dimensional space and vertex $i \in \{P_0, P_1, \ldots, P_N\}$. (In our system, n = number of frames between 2 key frames, K=2-D system, N =85.)

4.2 Digital Image Wrapping

1. Image Wrapping Formulation

For the pixels within the same triangle on a triangular mesh will be under the same transform of the triangular vertices to that of the next frame. In order to preserve all the textural information in consecutive frames, affine transformation [11] is estimated from the adapted FDP face model (the original frontal face object) together with textural dental information to the tweens. There are two types of estimation - local and global affine transformation. The local transform is estimated by every pair of corresponding triangular vertices while the global transform is estimated by least-square method between 2 frames.

Local: For frame $F^S \xrightarrow{T_{Ai}} F^D$ through affine mapping T_{A_i},where frame $F^S = \{\Delta_0^S, \Delta_1^S, ..., \Delta_G^S\}$.

$$\Delta_i^D = (\Delta x_i^D, \Delta y_i^D, L) = \Delta_i^S T_{A_i} \qquad \text{(Eq.2)}$$

$$\Delta_i^D = (\Delta x_i^S, \Delta_i^S, L)\begin{pmatrix} a & c & 0 \\ b & d & 0 \\ e & f & 1 \end{pmatrix}, \qquad \text{(Eq.3)}$$

where $\Delta x_i^m = (x_{i1}, x_{i2}, x_{i3})^T$, $\Delta y_i^m = (y_{i1}, y_{i2}, y_{i3})^T$ for *m* being source(*S*) or destination(*D*) frame, $L = (\mathbf{1\ 1\ 1})^T$, and *i* = 1 to *G*-th triangle(G=144). Thus, the 6-parametric affine transform can be determined as,

$$T_{A_i} = (\Delta_i^S)^{-1} \cdot \Delta_i^D. \qquad \text{(Eq.4)}$$

Global: An affine transformation can also be found by more than three pairs of triangular vertices with the least-squares method [12]. Thus, for frame $F^S \xrightarrow{T_{GA}} F^D$ through a global affine transform T_{GA} with parameters (*a, b, c, d, e, f*), where frame $F^S = \{P_1^S, P_2^S, ..., P_N^S\}$ and $P_i^S = (x_i^S, y_i^S)$ for i = 1 to *N*. Thus, we have each transformed vertex as $P_i^D = P_i^S T_{GA}$,

$$(x_i^D, y_i^D) = (x_i^S, y_i^S)\begin{pmatrix} a & c \\ b & d \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \qquad \text{(Eq.5)}$$

or a matrix form of vertex from 1 to *N*,

$$\begin{bmatrix} x_1^S & y_1^S & 0 & 0 & 1 & 1 \\ 0 & 0 & x_1^S & y_1^S & 1 & 1 \\ :: & :: & :: & :: & :: & :: \\ x_N^S & y_N^S & 0 & 0 & 1 & 1 \\ 0 & 0 & x_N^S & y_N^S & 1 & 1 \end{bmatrix}\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} x_1^D \\ y_1^D \\ : \\ : \\ x_N^D \\ y_N^D \end{bmatrix}. \qquad \text{(Eq.6)}$$

In order to give a more refined fit to somatic contours of the actual face and also to alleviate the global deformation, we use local affine mapping for every pair of triangles between pair of frames in our developed system.

2. Texture Mapping

The 6-para affine matrix can then be used to transform each pixel together with the color textural information to that of the next frame. A remedy method, called Neighborhood Spatial Averaging Technique (*NSAT*), for the defects of holes when smaller triangle mapped to larger triangle in the discrete Cartesian System is proposed. By applying a 2-D averaging filter (3x3 or/and 5x5) at the null color pixel, the null colored pixel will be replaced by the average/filtered color value (e.g., R,G,B) of its neighborhoods. However, for a suffered region with more than one null pixel within the window size, we can enlarge the window to 5x5 for increasing the candidates. After 5x5 filtering, the 3x3 average filter can be applied to the remained null color pixel. (Other filters such as lowpass or interleaving for smoothing the suffered region can also be applied.)

5. Application and Results

By using a high quality of phoneme-to-speech synthesizer for producing the speech, text-driven lip-synch application can be easily developed. Our first attempted application is for Digital TV news production. Results should be focused on the lip synchronization of the talking head under the hybrid synthetic and natural composition, and hence it is desirably tested under the human perception on the video playback. Fig.5 shows a synthesized digital frames sequence in NTSC standard with input text "Welcome to here". Each frame last for about 33.33ms for each temporal facial expression. The amounts of lips' opening and mouth shape of each frame represent the corresponding facial motion for the

pronunciation of the phonemes/diphthongs. Another potential application can also be implemented as virtual conferencing. If the FAPs and phonetic parameters each occupy 2kbps bandwidth, totally 4kbps, a 56kbps modem can afford 13 persons and the user to have a virtual text-driven conferencing for the 14 users simultaneously. Besides, a phonemic recognition from human speech, as shown in Fig.1, can substitute the text-to-phoneme engine, the real clean speech can be coded by existing very low bit-rate speech coder (<2kbps or 1.6kbps), like CELPC or HELPC. The phonetic parameters and the FAPs are transmitted to the other hosts and used to synchronize decoded speech for the virtual conferencing.

6. Conclusion

An SNHC for facial modeling on the frontal face object using two dimensional triangulation mesh model for automating frame sequence generation has been proposed. This provides an efficient representation and composition of synthetically and naturally generated audiovisual information. It can also be used to develop real-time interactive applications. The rules for RB-TTP engine and that for visemes or AFAUs always need further developed and tuned since they play the main role of lip synchronization as well as visual prosody. The content-based interactivity with combination of synthetic and natural elements is also an additional feature that is needed in SNHC, in conjunction with our current research direction on the MPEG-4 system as well as on the SNHC with 3-D human head and body animation.

References

[1] ISO/IEC JTC1/SC29/WG11, N1454, MPEG-4 Synthetic/Natural Hybrid Coding Verification Model 2.0 (SNHCVM), Coding and Moving Pictures and Audio.

[2] ISO/IEC JTC1/SC29/WG11, N1666pub, MPEG-4 Synthetic/Natural Hybrid Coding Verification Model 4.0 (SNHCVM), Coding of Moving Pictures and Audio.

[3] K.Waters, D.Terzpoulos, *Modeling and Animating Faces using Scanned Data*, The J. of VCA, 2, 1991.

[4] N.Magenat-Thalmann, E.Primeau and D.Thalmann, *Abstract Muscle Action Procedures for Human Face Animation*, The Visual Computer (1988) 3:290-297.

[5] C.S.Choi, H.Harashima, *Analysis and Synthesis of Facial Image Sequence in Model-Based Image Coding*. IEEE Trans. CSVT. 4(3), June 1994

[6] P.Ekman and W.V Friesen, *Facial Action Coding System*, Consulting psychologists press 1977.

[7] M. Rydfalk, *CANDIDE: A parametrised face*, Dept. Elec. Eng. Rep. LiTH-ISY-I-0866, Linköping Univ., Oct 1987.

[8] M.J.T. Reinders., P.J.L. van Beek, B. Sankur, J.C.A. van der Lubbe, *Facial Feature Localization and Adaptation of a generic face model for model-based coding*. Signal Proc.: Image Comm. 7 (1995) 57-74

[9] M.Nahas, H. Huitric and M.Saintourens, *Animation of B-Spline Figure*, The Visual Computer(1988) 3:272-276.

[10] Jonas Beskow, *Rule-based Visual Speech Synthesis*, ESCA.EUROSPEECH'95.4th Sept,1995.

[11] G.Wolberg, *Digital Image Warping*, IEEE Computer Society Press Monograpgh.

[12] M.Xie, *Feature matching and affine transformation for 2D cell animation*, The Visual. Computer (1995) 11: 419-428.

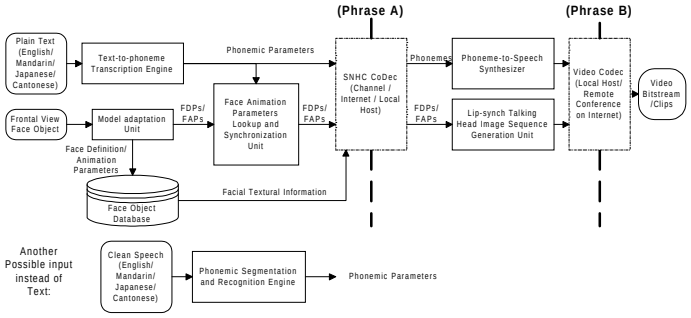


Fig.1 Block diagram of SNHC system

Phoneme	*Example*	*IPA Description*
COP		
p	pin	Voiceless bilabial plosive
b	bin	Voiced bilabial plosive
t	tin meter	Voiceless alveolar plosive & alveolar flap
d	dig	Voiced alveolar plosive
k	kin	Voiceless velar plosive
g	give	Voiced velar plosive
COA		
CH=t+SH	chin	Voiceless alveolar affricate
JH=d+ZH	edge	Voiced alveolar affricate
COF		
f	fork	Voiceless labiodental fricative
v	vat	Voiced labiodental fricative
TH	thin	Voiceless dental fricative
DH	this	Voiced dental fricative
s	sin	Voiceless alveolar fricative
z	zing	Voiced alveolar fricative
SH	shin	Voiceless postalveolar fricative
ZH	azure	Voiced postalveolar fricative
HH	hit	Voiceless glottal fricative
CSN		
m	aim	Bilabial nasal
n	pan	Alveolar nasal
nx	thing	Velar nasal
CSL		
r	red	Retroflex approximant
l	lid elbow	Alveolar lateral approximant Velar lateral approximant
CSG		
w	wasp	Labiovelar sonorant glide
y	yard	Palatal sonorant glide
VC		
IH	pit	Front close unrounded (lax)
EH	pet	Front open-mid unrounded
AE	pat	Front open unrounded (tense)
OW	go	Back close-mid round
AH	cut	Open-mid back unrounded
UH	book	Back close-mid unrounded (lax)
VS		
AX	about	Central close mid (schwa)
VF		
IY	ease	Front close unrounded
EY	ate	Front close-mid unrounded (tense)
AY=AA+IH	tie	Diphthong
OY=AO+IH	toy	Diphthong
UW	lose	Back close round
AW=AA+UH	foul	Diphthong
ER	furs	Central open-mid unrounded rhoticized
AA	car	Back open unrounded
AO	cause	Open-mid back round

Table 1. The mostly used 40 U.S. phonemes

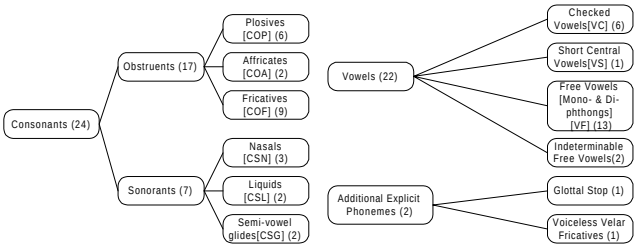
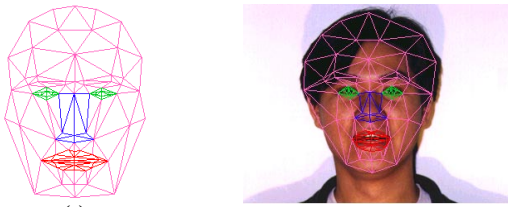


Fig.2 Classification of US Phonetics



(a) (b)

Fig.3 (a) Generic Face Model (b) FDPs on frontal face

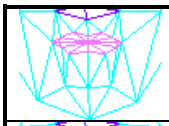
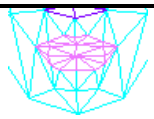
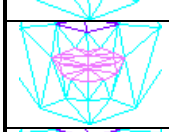
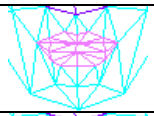
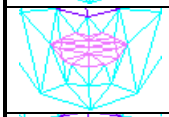
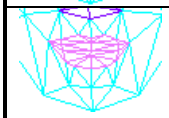
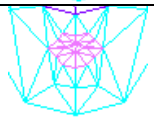
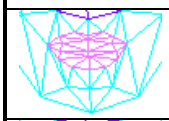
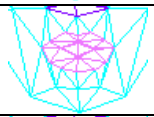
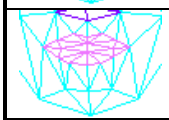
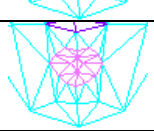
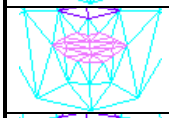
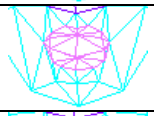
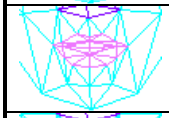
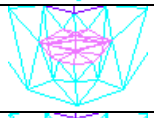
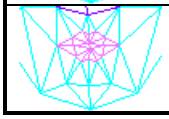
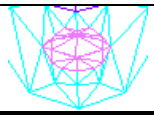
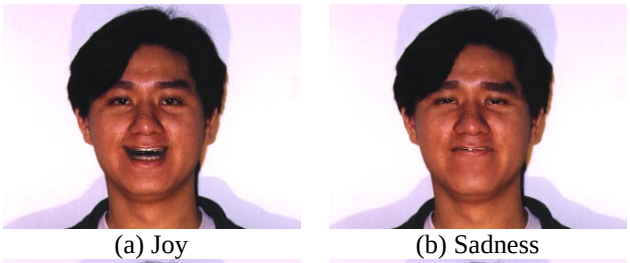
	Viseme=1 /p/, /b/, /m/.		Viseme=10 /AA/, /AH/.
	Viseme=2 /f/, /v/.		Viseme11 /EH/, /AE/, /EY/, /HH/.
	Viseme=3 /TH/, /DH/.		Viseme=12 /IH/, /IY/.
	Viseme=4 /t/, /d/, /y/		Viseme=13 /OW/.
	Viseme=5 /k/, /g/.		Viseme=14 /UH/.
	Viseme=6 /CH/,/JH/, /SH/,/ZH/		Viseme=15 /UW/.
	Viseme=7 /s/, /z/.		Viseme=16 /AX/.
	Viseme=8 /n/, /l/, /NX/.		Viseme=17 /ER/.
	Viseme=9 /r/, /w/.		Viseme=18 /AO/.

Table 2. 18 Visemes correspond to the mostly used 40 phonemes (24 consonants + 13 vowels + 3 diphthongs)



(a) Joy (b) Sadness

(c) Anger (d) Surprise

Fig.4 Facial Expression

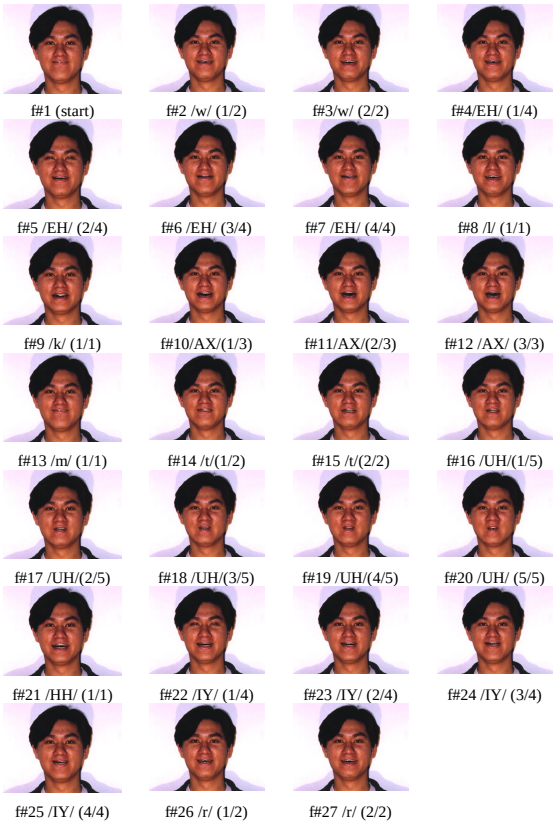


Fig.5 Synthesized lip-synch image sequence saying “Welcome to here” with phonemes /wEHlkAXm/ /tUH/ /HHIYr/