ICAT '99

Web-enabled Speech Driven Facial Animation

Ming Ouhyoung, I-Chen Lin , David S.D. Lee*,
Communication and Multimedia Laboratory
Dept. of Computer Science and Information Engineering, National Taiwan University
*Cyberlink Inc. Taiwan
http://www.cmlab.csie.ntu.edu.tw/~ming

Abstract

In this paper, we present an approach that animates facial expressions through speech analysis. An individualized 3D head model is first generated by modifying a generic head model, where a set of MPEG-4 Facial Definition Parameters (FDPs) has been pre-defined. To animate facial expressions of the 3D head model, a real-time speech analysis module is employed to obtain mouth shapes that are converted to MPEG-4 Facial Animation Parameters (FAPs) to drive the 3D head model with corresponding facial expressions. The approach has been implemented as a real-time speech-driven facial animation system. When applied to Internet, our talking head system can be a vivid web-site presenter, and only requires 6 Kbps with an additional header image (about 30Kbytes in CIF format, JPEG compressed). The system can synthesize facial animation more than 30 frames/sec on a Pentium III 500 MHz PC.

1. Introduction

Since the growth of Internet usage is exponential, nowadays, web has already been an indispensable part of life. Besides, users are not satisfied with static information such as homepages with only static image and text; hence more and more media files are transferred over Internet, even allowing a user's interaction. It is difficult to use in streaming high resolution videos due to the bandwidth constraint. So model based video coding approach is one of the most popular research topics in this area, using synthetic faces and talking head instead of current frame-based videos. In the international standard MPEG-4 [1][2], synthetic heads are also included. The head model parameters and the control of facial expressions are defined as a set of Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs) respectively.

In general, the problem, modeling one's head, can be roughly divided into three kind of approaches, 3D model, 2D mesh and sample based . Some use physical 3D model such as bones and muscles to synthesize one's face [3]. Most researchers use a generic model with texture mapping from a set of images. Pighin et al. [4] proposed a delicate approach to reconstruct one's 3D head model from fine images; Guenter [5] uses information form six cameras to generate a model plus a changing texture map. Lee and Thalmann proposed [6] a semi-automatic approach, which is based on the front view and side view images of a person. The major advantage of 3D head model is that it's suitable to have arbitrary motion and rotation. In general, it is a trade-off between lifelikeness and efficiency. 2D mesh and image warping approach is simpler and so more computationally effective. The MTV video clip "black or white" is an impressive demonstration while the Image Talk [9], our previous system, is another example of this kind of approach. Sample-based approach means combining individual parts of face features extracted from video clips of a talking person. Bregler [7] recorded the mouth images in the training footage to match the phoneme sequence of the new audio tracks. Brand [8] analyzed the video to yield a probabilistic state machine, mapping to facial configuration space. Synthetic talking head with this technique can look quite real, but it suffers from large storage space.

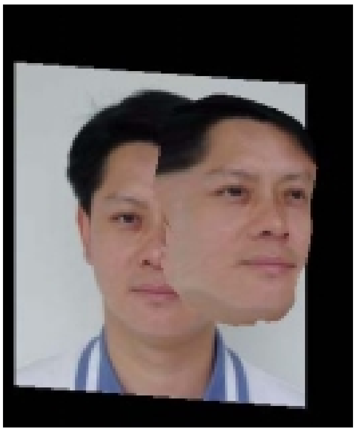

Figure 1. Illustration of two and half dimension head model

In this work, we propose to synthesize one's face using a two and half dimension head model, with the facial expression driven by speech, and make it web-enabled to take advantage of streaming technology.

This paper is organized as the following. In Section 2, the proposed two and half dimension head model is first introduced, and explain how to generate facial expressions based on speech in Section 3. Some streaming issues about data transferring over Internet is discussed in Section 4. A complete web-enable talking head system is presented in Section 5 and we conclude the paper in Section 6.

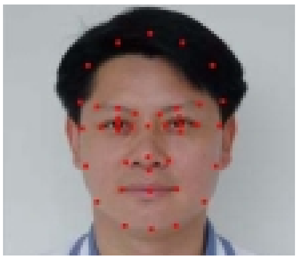

Figure 2. Specify feature points on an image for adjusting from a generic 3D head model

2. Head model

Our basic requirements are simple, i.e. photo-realistic but low bit-rate animation data over Internet. 2D mesh and warping technique is employed on a single face image in VR-Talk [9][10], our previous speech driven talking head system. But the above animation is not natural in nodding direction. When developing a system based on 3D model, to fast construct one's head model is not very easy; besides, we can't overcome the problem for hair rendering. Thus, we adopt a two and half dimension head model, which consist of a front side view head image and a half-cut 3D model (see Fig 1, 3).

The major advantage of this model is to combine both nice features from 2D mesh and 3D model: simple, vivid, and naturally when small-scale rotation is applied. Also we can easily combine the head model with a natural scene image using alpha blending technique. Thus, the background can be easily replaced according to a user's intention.

2.1 Head model fitting

First, a frontal and neutral face image is needed. In order to fit the generic 3D head model, a user must specify about 30 feature points, such as eyes, nose and mouse boundary. Some predefined control points on the 3D head model are adjusted to proper position, and the texture data is obtained by orthogonal projection. As shown in Fig.2, an editor tool is also developed to help a user edit his model in 3 minutes.

Because the 3D head mask only contains frontal half head, the other part is covered by a rectangle patch with texture of the neutral face image, including human hair. It looks real if the rotation angel is constrained to less than 30 degrees.

2.2 Hybrid combination of synthetic object and natural scene

Recently, the concept of object based coding [1] has been getting more and more emphasis. It is an important feature to let a user combine a synthetic talking head with any real background image. Alpha blending technique is employed to achieve the goal to allow replacing a background dynamically.

First, an image processing tool is used to find the contour of the original image, and then build an front alpha mask, which has value zero at non-face area, one at face area, and values obtained by linear interpolation around contour (see Fig.4). The following equation is used to generate final image for displaying.

pixel of display plane *= front alpha * 3D rectangle projected value ＋ (1-front alpha) * Background image pixel value.*

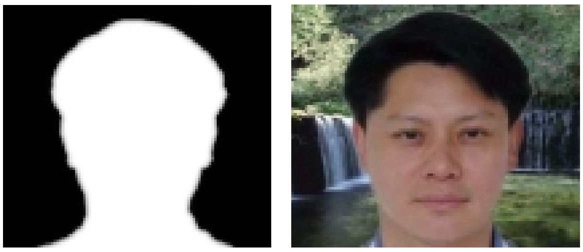

Figure 4. Frontal mask for alpha blending (left) and the result after combining head model with natural background image (right)

2.3 Other head model features

Eye blinking and head motion are common action when a person is speaking. In order to make the talking head more realistic, the above functions should be implemented. To keep head moving, a sequence of transformation parameters based on domain knowledge are applied to the head model. In this way, we can simulate one's action, such as nodding, and head shaking naturally.

In general, a person's eyes blink once every four seconds. The method used is similar to our previous system the Image Talk [9], which is done by dragging the control points of upper eyelid downward.

Teeth usually are ignored in many talking head system except sample-based method because of the lack of teeth information. We propose a generic teeth model to simulate one's teeth inside the mouth. The

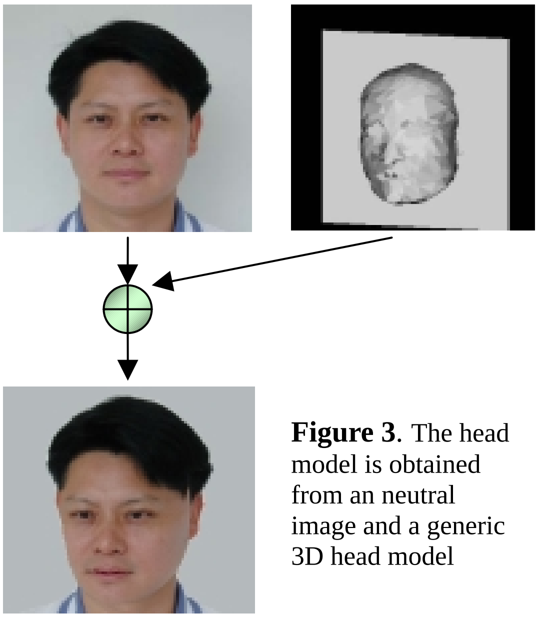

Figure 3. The head model is obtained from an neutral image and a generic 3D head model

teeth model is separated into two part: the upper part and the lower one, moved according to the control vertex at the philtrum and the one at the chin, respectively. In addition, there is a basic assumption the larger one's mouth is opened, the more light his teeth can be illuminated. As a result, the smaller the distance between upper lip and the lower one is, the darker the teeth are.

3. Speech driven facial expressions

A set of MPEG-4 Facial Animation Parameters (FAPs) is used for facial animation, both for visemes and expressions. In MPEG-4, there are 14 visemes and 6 expressions; in our system, there are currently 9 visemes implemented. All the visemes and expressions are edited and saved by our expression editor, and can be applied to any individualized head model without modification.

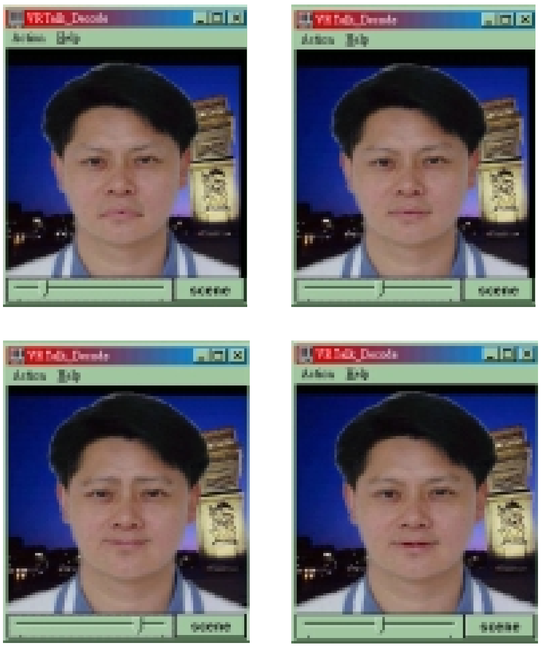

Figure 5: Facial expressions with an "emotion index" slider for real-time manipulation.

3.1 Expression editor

All the animation parameters are defined on the generic head model. Users can edit the expression on the generic head model by dragging the predefined FDP feature points on the face. After fine-tuning the expression, it is saved in FAPs format (see table 1). Because all individualized head models are deformed from the same generic head model, they all exhibit the same FDP feature points. Therefore, all the FAPs tables are applicable to these deformed head model without any modification.

In current implementation, two typical expressions are edited: joy and anger (Fig.5). All the other expressions defined in MPEG-4 can also be generated by our expression editor (Fig. 10).

3.2 Viseme

In Mandarin Chinese, there are in total 408 utterances without tonal variation [16], and 1333 Chinese utterances with tonal variation. All Mandarin Chinese utterances are combinations of 37 syllables. The 37 syllables are shown in Table 2. These 37 syllables are classified into 9 clusters, and each cluster corresponds to a basic viseme. The 9 basic visemes, together with their corresponding syllables are listed in Table 2. In our system, only 9 visemes are used to pronounce all the Chinese syllables.

Viseme	FAP number	Displacement
A	4	-273
	5	-273
	6	14
	7	43
	8	-205
	9	-205
	10	-171
	11	-171
	12	34
	13	137
	51	-34
	52	34
	53	14
	54	43
	55	0
	56	0
	57	34
	58	34
	59	34
	60	137

Table 1. *FAPs* table for viseme 'A'. (The units of FAPs with number 6,7,53,54 are MW (mouth width); the units of other FAPs are MNS (mouth-nose separation).

3.3 facial animation driven by preprocessed speech

First, we will preprocess the input speech file using either Microsoft DirectSpeechTecognition API [17] or the speech recognition engine from Applied Speech Technologies [18], and then save the recognition results as an index file. The detail is shown in our previous work [9,10].

Visemes and other facial expressions are animated independently, so the talking head can change its expression while it is speaking. Changing from one expression to another is by linear interpolation in a user-defined time interval.

4. Web-enabled talking head system

In order to be web-enabled, our system must have characteristics of very-low bit-rate, short responsive

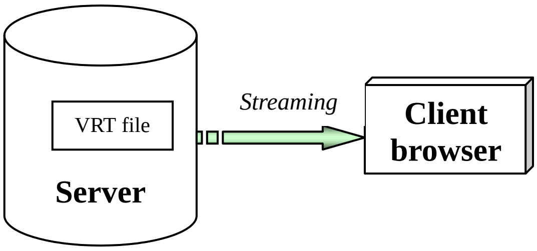

Figure 6. The concept diagram of web-enabled talking head

time, and natural animation.

Since facial expressions of the proposed system are controlled by phonetic and emotional information which are sets of key frame numbers and time-slice data; the speech data can be encoded by CELP (Code Excited Linear Prediction) coding techniques such as G.723.1, the bandwidth requirement of VR Talk is very low. To minimize the responsive time and make the animation play smoothly, we adopt streaming architecture with ring buffers to manage the data transformation on Internet. A VRT (VR Talk streaming data) format is also proposed, which includes information of head model, facial animation control, and encoded speech. This format can be transformed to other streaming data format such as ASF (advanced streaming format) of Microsoft.

Fig.6 is a conceptual diagram of our web-enabled talking head system. The system can be separated into two parts: the server side and the client side. In the server side, a VRT file is prepared in advance. Our web-enabled VR Talk player is implemented as a plug-in for web browser. When a user clicks a link to VRT file, our plug-in downloads the VRT data in streaming and plays back the speech with relative facial animation.(Fig. 5)

4.1 VR Talk streaming data format

A VRT file consists of a VRT header and a chain of packet data. In the VRT header, essential information to reconstruct a talking head such as the head model, and its texture image is included (see Fig. 7). Besides, a background scene and an alpha blending mapping table can be also packed into the header and then our plug-in can apply the approach mentioned in section 2 to make the talking head blend into the background scene.

Viseme	Corresponding Chinese Syllables
A	a, ai, au, ang
B	b, p, m
D	d, t, n, l, g, k, h
E	e
EN	eh, ei, an, en, eng, er
F	f
J	ji, chi, shi, j, ch, sh, r, tz, ts, s, i
O	iu, o, ou
U	u

Table 2. The syllable-to-viseme table. 9 basic visemes used in our system, and their corresponding syllables.

4.2 Streaming and ring buffer

The architecture diagram of VR Talk client plug-in is shown in Fig.8. The plug-in can be viewed as three components: de-multiplexer, visual manager, and audio manager. The de-multiplexer is in charge of storing streaming data into the ring buffer and dispatching the de-multiplexed data in the ring buffer to the other components.

After the plug-in is activated, a specified VRT file is downloaded from the web server. The de-multiplexer stores streaming data into the ring buffer. When the VRT header is transmitted completely, the de-multiplexer sends the data to the visual manager to reconstruct the talking head.

To play back speech and animation smoothly, the de-multiplexer can’t dispatch speech and animation data until at least 12 seconds of equivalent streaming data are stored in the ring buffer, depending on the effective network bandwidth.

In the speech manager, once it receives encoded speech data, it decodes the data and then plays the speech segment. The visual manager synchronizes facial animation with speech by synchronizing the current speech position to facial animation.

Unlike viseme controller and expression controller that receive control information, there is an autonomous controller in the visual manager, it is in charge of

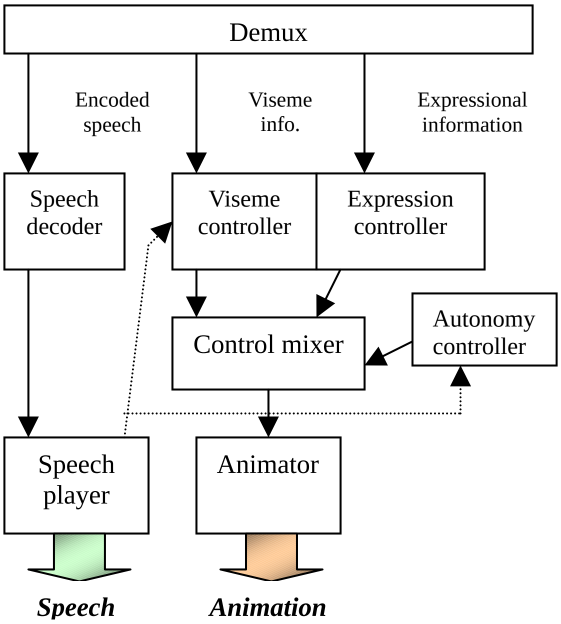

Figure 8. The architecture of a client player. (dotted lines mean synchronization between modules)

autonomic action such as head motion and eye blinking.

4.3 System implementation

In a VRT file, images and speech data are the major part of it. To reduce the VRT file size, we adopt the JPEG image coding approach to encode the texture image and background; the speech coding approach G.723.1 with silent detection is applied to reduce the speech stream to less than 5.3 K bits per second.

At this moment, the display window is of size 256 x 256 pixel. The size of texture image or background image is about 15K to 20K bytes, and the size of alpha blending mapping table is about 12K bytes. There are about 900 triangles in the generic head mask. Currently, we just store the triangle information without further encoding, and triangle data size is about 70K bytes. To sum up, the VRT header size is about 120K bytes.

In streaming packet data, the animation control stream is about 600bps, which is much less than that of speech. Comparing with current encoding techniques such as H.261 and H.263, whose bit-rate is about 40K to 4M bits per second in QCIF format, our proposed system can provide a low bit-rate and high-quality tool for video applications on Internet.

For the time being, our system is developed on Windows 95/98. Two kinds of web browsers, Internet Explorer (IE) and Netscape Navigator are supported. The OpenGL is adopted as the graphic-rendering library.

On a PentiumIII 500Mhz PC without OpenGL hardware acceleration, the frame rate is about 20 frames per second. However, once the OpenGL hardware acceleration is turned on, the frame rate can reach more than 300 frames per second.

5. Results and applications

A complete web-enable talking head system is proposed. A user can dynamically change the foreground speaker, the background scene, and the mood of the speaker. Featuring low bit-rate streaming and real-time user controlled emotion index, our web-enabled system is now available at the web site http://www.cmlab.csie.ntu.edu.tw/~ming.

One immediate application of the web-enable talking head system is a merchandise presenter. The talking head animator is packed in both ActiveX control module and Netscape Plugin. In a typical application, speech data about a target product and the corresponding index file, which is generated by off-line processing, are put on the web-server. After downloading the package, a user can see vivid introduction presented by the artificial talking head.

6. Future work

In current implementation, head motion is driven by a set of motion parameters generated from domain knowledge plus random number generator. Occasionally, the synthetic head looks a bit strange because the motion is not natural. A simple approach for enhancement is to apply transformation parameters that are generated by 3D head motion estimation algorithm [11] based on an user's head motion. After recording the real motion, the talking head will behave more naturally. Besides, the current streaming format is a temporal solution, later it should be compatible to a standard format, such as Advanced Streaming Format proposed by Microsoft or MPEG-4 streaming format. It is easily modified since the difference is not much.

Reference:

[1] MPEG4 Systems Group, *"Text for ISO/IEC FCD 14498-1 Systems,"* ISO/IEC JTC1/SC29/WG11 N2201, 15 May 1998.

[2] J. Ostermann, *"Animation of Synthetic Faces in MPEG-4"*, Proc. of Computer Animation, pp.49-51, Philadelphia, Pennsylvania, USA, June 8-10, 1998.

[3] Demetri Terzopoulos, Keith Waters, *"Analysis and synthesis of Facial Image Sequences using Physical and Anatomical Models,"* IEEE Tran. On Pattern and Machine Intelligence,15(6), Jun.1993, pp.569-579.

[4] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Pichard Szeliski, David H. Salesin, *"Synthesizeng Realistic Facial Expressions from Photographs,"* Proceedings of ACM Computer Graphics (SIGGRAPH 98), pp. 75-84 Aug-1998.

[5] B. Guenter, c. grimm, D. Wood, H. Malvar, F. Pighin, *"Making Face"*, Proc. of Computer Graphics (SIGGRAPH '98), pp. 55-66, Aug. 1998.

[6] Won-Sook Lee, Nadia Magnenat Thalmann, "*Head Modeling from Picutes and Morphing in 3D with Image Metamorphosis Based on Triangulation*," Proc. CAPTECH'98, Geneva, pp.354-267, 1998.

[7] C. Bregler, M.Covell, M.Slaney, *"Video Rewrite: Driving Visual Speech with Audio"*, Proc. SIGGRAPH'97, pp.353-360, 1997.

[8] Matthew Brand, "*Voice Puppetry*", Proc. SIGGRAPH'99, pp.21-28, 1999.

[9] Woei-Luen Perng, Yungkang Wu, Ming Ouhyoung, *"Image Talk: A Real Time Synthetic Talking Head Using One Single Image with Chinese Text-To-Speech Capability"* Proc. of PacificGraphics 98, pp. 140-148, Singapore, Oct 1998.

[10] I-Chen Lin, Cheng-Sheng Hung, Tzong-Jer Yang, Ming Ouhyoung, "*A speech Driven Talking Head Based on a Single Face Image*", pp.43-49, Proc. of PacificGraphics'99, Seoul, Oct. 1999.

[11] Tzong-Jer Yang, fu-Che Wu, Ming Ouhyoung, "*Real-time 3D Head Motion Estimation inFacial Image Coding*", Porc. Of Multimedia Modeling, Lausanne, Switzerland, Oct. 12-15, 1998, pp.50-51.

[12] Thaddeus Beier, Shawn Neely *"Feature-Based Image Metamorphosis"*, Proc.of SIGGRAPH 92. In Computer Graphics, pp. 35- 42

[13] Steven M.Seitz, Charles R. Dyer, *"View Morphing"*, Proc. SIGGRAPH 96, pp. 21-30.

[14] Nur Arad, Nira Dyn, Daniel Resfeld, Yehezkel Yeshurun, *"Image Warping by Radial Basis Functions: Application to Facial Expressions"*, CVGIP: Graphical Models and Image Processing", Vol. 56, No.2, pp.161-172, 1994.

[15] Eric Cosatto, Hans Peter Graf, *"Sample-Based Synthesis of Photo-Realistic Talking Heads"*, Proc. of

Computer Animation 98, pp. 103-110, Philadelphia, Pennsylvania, June 8-10, 1998.

[16] Lin-Shan Lee, Chiu-Yu Tseng, Ming Ouhyoung, *"The Synthesis Rules in a Chinese Text-to-Speech System"*, IEEE Trans. On Acoustics, Speech and Signal Processing. Pp.1309-1320. Vol.37, No.9, 1989.

[17] Microsoft Speech Technology SAPI 4.0 SDK, http://www.microsoft.com/iit/projects/sapisdk.htm

[18] Applied Speech Technologies Corporation. http://www.speech.com.tw

[19] Tzong-Jer Yang, I-Chen Lin, Cheng-Sheng Hung, Chien-Feng Huang and Ming Ouhyoung, *"Speech Driven Facial Animation"*, pp. 99-108, Proceedings of Computer Animation and Simulation Workshop'99, Milan, Italy, Sept. 1999.

[20] M.Esoher and N.M. Thalmann, *"Automatic 3D Cloning and Real-Time Animation of a Human Face"*, Proc. Computer Animation 97, pp.58-66, 1997.

[21] D. Decaolo, D. Metaxas, M. Stone, *"An Antropometric Face Model Using Variational Techniques"*, Proc. Computer Graphics (SIGGRAPH '98), pp. 67-74, Aug. 1998.

[22] T. DeRose, M. Kass, T. Truong, *"Subdivision Surfaces in Character Animation"*, Proc. of Computer Graphics (SIGGRAPH'98), pp85-94, Aug 1998.

[23] P.E Kmon, W.Fresen, *"Facial Action Coding System: A Technique for the Measurement of Facial Movement"*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[24] S. Morishima, H.Harashima, *"A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface"*, IEEE J. Selected Areas in communications, 9, pp. 594-600, 1991.

[23] M.M. Cohen and D.W. Massaro. *"Modeling co-articulation in synthetic visual speech"*. In N.M. Thalmann and D. Thalmann, editors, Models and Techniques in Computer Animation. Springer-Verlag, 1993.

Content information	Header
Adjusted Head model	
Texture image	
Background image (opt.)	
Alpha-blending mapping table (opt.)	
Packet Header	Packet
Encoded speech data	
Utterance data	
Emotional data (opt.)	

Figure 7. The header and packet format of VRT.

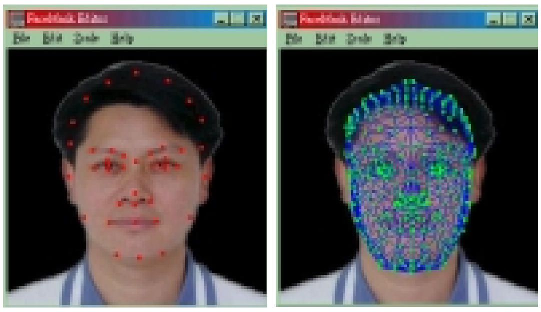

Figure 9. The wireframe model resulted from feature point selection.

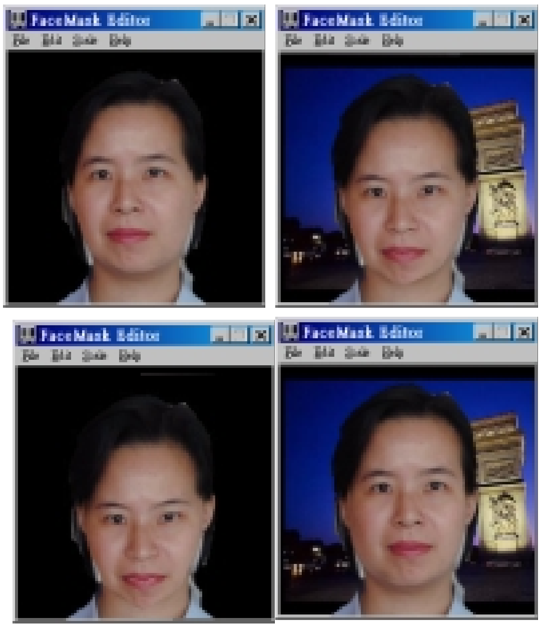

Figure 10. Different viewing angles of the same model