Computer Animation and Simulation 2000, Springer, pp.29-40
(Proc. of the Eurographics Workshop in Interlaken, Switzerland)

A Low Bit-rate Web-enabled Synthetic Head with Speech-driven Facial Animation

I-Chen Lin, Chien-Feng Huang, Jia-Chi Wu*, Ming Ouhyoung

Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan,
CyberLink Corporation, Taiwan*
{ichen, cardy, ming}@cmlab.csie.ntu.edu.tw, jc_wu@cyberlink.com.tw*

Abstract. In this paper, an approach that animates facial expressions through speech analysis is presented. 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 realistic facial expressions of the 3D head model, key frames of facial expressions are calculated from motion-captured data. A 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 *14 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. Currently, the data streaming are implemented under Microsoft ASF format, Internet Explorer, and Netscape's Navigator.

Keywords Web-based animation, facial animation, face modeling.

1. INTRODUCTION

It is difficult to "stream" high-resolution videos due to the bandwidth constraint. So model-based video coding approach, using synthetic faces and talking heads instead of current frame-based videos, is one of the most popular research topics in this area. In the international standard MPEG-4 [1][2], the head model parameters and the controls of facial expressions are defined as a set of Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs) respectively. However, synthesizing video-realistic facial animation is still difficult, since our eyes can be very sensitive to any tiny imprecision on a synthetic face. How to model one's head, how to animate the facial expression in real-time, and how to synchronize the animation with the speech are the three critical problems to generate realistic facial animation.

In general, modeling one's head can be roughly divided into three kinds of approaches, 3D model, 2D mesh and sample based. Some use physical 3D models such as bones and muscles to synthesize one's face and action [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 with view-dependent 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 using a 3D head model is that it is flexible for synthesizing facial actions and can be viewed from any viewpoints. Nevertheless, efficient modeling and rendering of hairs in 3D is still considered difficult. The approach of image warping based on a 2D mesh is simpler and more computationally effective. The MTV video

clip "black or white" is an impressive demonstration while the Image Talk [9], our previous research 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. Synthetic talking head with this technique can look quite real, but it suffers from large storage space.

The issue of automatic lip synchronization can be tackled from two directions: synchronization to synthetic speech and synchronization to real speech. Most of researches take the former direction to reduce the difficulty [2,13,21]. However, synthetic speech is difficult to sound natural and personal; thus, we adopted real speech and exploited speech analysis techniques to drive facial expressions. *Voice Puppetry* [8] is another voice driven facial animation system, which analyzed the video to yield a probabilistic state machine, mapping vocal features to facial configuration space.

In our work, a set of motion-captured data of human face is utilized to animate the talking head; an algorithm for compensating these data is described in this paper. A low bit-rate web-enabled talking head is our target. We propose to use a hybrid model composed of a 3D half cut head and hair image patches to synthesize one's head. In addition, an automatic lip-synchronization module by speech analysis is also presented.

The remainder of this paper is structured as follows: In Section 2, the proposed two and half dimension head model is first introduced. Section 3 describes how we got facial motion by compensating global motion of captured data. Synchronization between speech and synthetic facial expression is presented in Section 4. A complete web-enabled system is described in Section 5. We conclude this paper in Section 6.

2. HEAD MODELING

The requirement of the proposed system can be stated as photo-realistic but low bit-rate animation data over Internet. 2D image warping technique was employed on a single face image in VR-Talk [9][10], our previous speech driven talking head system. But the above animation is view morphing based and so is not very natural in rotation. When developing a system purely based on 3D model, we can't overcome the problem of hair rendering, which is one of the most difficult issues in real-time

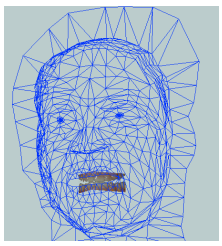


Fig 1. The wire-frame display of 2½D head model.

Fig 2. 2½ D hybrid head model in different scenes.

computer graphics. Thus, we adopt a two and half dimension head model, which consist of a half-cut 3D model and an image plane (Fig. 1, Fig. 2) with a front-side view head image. With this extra image plane, our talking head can exhibit one's hair, neck, and smooth contour. The major advantage of this model is to combine both nice features from 2D mesh and 3D model: simple, vivid, and natural when a small-scale rotation less than 30 degree is applied.

2.1 Combination of 2½ D Hybrid Head Model and Natural Scenes

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 scenic image. To achieve the goal of replacing a background dynamically, the alpha blending technique is employed as following.

First, an image-processing tool is applied to find the contour of the original image, and then build a front alpha mask, which has value zero at non-face area, one at face area, and values obtained by linear interpolation around contour. Then the following equation is used to generate the final image for display.

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

3. GLOBAL MOTION COMPENSATION OF CAPTURED FACIAL MOTION

For 3D lip motion and facial animation, we use a commercial optical motion capture device for "viseme" generation. The same "viseme" can then be modified to fit into different talking heads. 3D facial motion is captured at our industrial collaborator, Digimax Production Center, where a VICON 8 motion capture system is used. Eight cameras are set up, and 23 optical markers are attached on a performer's face (Fig. 3). The VICON 8 system captures the performer's facial expressions at 60 frames/second. After the process of feature extraction and 3D reconstruction, the output file with C3D data format contains the 3D position of 23 features for each frame.

The retrieved 3D coordinates of marker points attached on the actor's face fully recorded the facial actions. We are interested in the mouth movement from speaking and facial expressions. Unfortunately, the global motion, such as head rotation and translation, also moves the positions of feature points. It is not a reasonable requirement to ask the performer to fix his head when he acts. Hence, the first task we have to solve is to compensate for the global motion and then the remaining offsets can be applied to drive the facial animation.

3.1 Algorithm for motion estimation

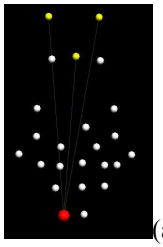
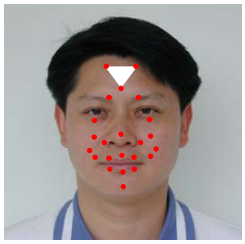

(a)

(b)

Fig 3. (a) The 23 captured optical markers. The red point is the rotation pivot, and the three yellow points are used for global motion estimation. (b) Illustrates the placements of the 23 optical markers on one's face for facial motion capture, where the three points forming a rigid white triangle are identical to the three yellow points in (a).

In this case, the problem belongs to "3D-to-3D feature correspondences" [23]. Suppose the features *pi, pi'* are 3D coordinates of points on the surface of the rigid body, observed at time t_1 and t_2. Given N corresponding pairs (*pi, pi'*), which obey the relationship of

$$p_i' = Rp_i + t, \quad i = 1,..N \tag{1}$$

It is well known [24] that three non-collinear point correspondences are necessary and sufficient to determine *R* and *t* unique. With three point correspondences, we will get nine non-linear equations while there are six unknown motion parameters. Because the 3D points obtained from motion capture system are accurate, linear algorithm is good enough for this application, instead of iterative algorithms or methods based on the least square procedure. The improved method based on translation invariants [25] is adopted to solve the motion estimation problem.

If two points on the rigid body, p_i and $p_{i+1,}$ which undergo the same transformation, move to p_i' and p_{i+1}' respectively, then $p'_i = Rp_i + t$ and $p'_{i+1} = Rp_{i+1} + t$. Subtraction eliminates the translation t, and using the rigidity constraint yields

$$\frac{p_{i+1}' - p_i'}{\left| p_{i+1}' - p_i' \right|} = R \frac{p_{i+1} - p_i}{| p_{i+1} - p_i |} \tag{2}$$

Define the above equation as $\hat{m}'_i = R\hat{m}_i$, $1 \le i \le 3$, which $\hat{m}$ is a unit vector. If the rigid body undergoes a pure translation, these $\hat{m}$ parameters do not change, which means translation invariant.

After rearranging these three equations, we can solve a 3×3 linear system to get R, and afterward obtain t by substitution into equation 1. In order for a unique solution, the 3×3 matrix of unit $\hat{m}$ - vectors must be of full rank, that is the three $\hat{m}$ - vectors must not be coplanar. As a result, four point correspondences are needed instead of three points, the minimal requirement. To overcome the problem of supplying the linear method with an extra point correspondence, a "pseudo-correspondence" can be artificially constructed due to the property of rigidity. In our case, the problem is resolved by finding a third $\hat{m}$ - vectors orthogonal to the other two. Via this improvement, the system is of lower dimension, only three point correspondences are required, and it helps to reduce the singularity problem of a matrix. The third vector can be achieved by setting $\hat{m}_3 = \hat{m}_1 \times \hat{m}_2$ and $\hat{m}'_3 = \hat{m}'_1 \times \hat{m}'_2$. These artificial vectors are generated to span the three dimension spaces.

3.2 Rotation Pivot Estimation

In the above algorithm, we assume the rotation pivot is at the origin of world coordinate. Although the position of pivot point does not influence the result of estimated rotation matrix *R*, it changes the translation vector *t* slightly. Let *O* be the rotation pivot, and p is a point on a rigid body, which undergoes *R* and *t*, then $p'-O = R \cdot (p - O) + t$. After rearranging this equation, we get:

$$t = p' - R \cdot p + (R - I) \cdot O \tag{3}$$

Since the global rotation angle is not large, this implies the rotation matrix is close to identity matrix. Besides, the pivot position is not far from the world coordinate origin due to the preprocessing based on domain knowledge; thus the translation vector can be assumed invariant to the position of the pivot point. Here we propose an algorithm to estimate the real position of rotation pivot to obtain better 3D data compensated global motion. In Figure 4, p_1, p_2 and p_3 are points on a rigid body, where l_i is the edge length between p_i and pivot point and θ_i is the angle between two vectors OP_i and OP_{i+1}. When the rigid body undergoes a transformation about the pivot, l_i and θ_i are invariant due to the property of rigidity.

In order to estimate the position of pivot point, we need to develop an objective function that measures if the points on rigid body obey the above-mentioned properties. The cost function consists of three metrics as follows:

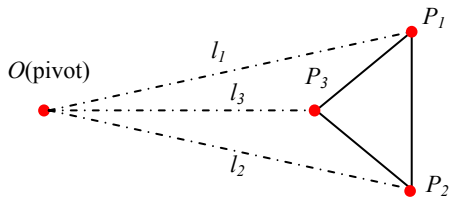


Fig 4. The relation between rigid body and rotation pivot.

$$E_k(p^k) = w_1(\sum_i \sum_j \left\| \theta_j^0 - \theta_j^i \right\| * l) + w_2(\sum_i \sum_j \left\| l_j^0 - l_j^i \right\|) + w_3(\frac{\left\| p^k \right\|}{C}) \quad (4)$$

where p^k is the candidate pivot at iteration k, θ_j^i is the jth angle on frame i and l_j^i is the jth edge length on frame i. C and l are two parameters which must be adjust depending on the 3D data itself and coefficients w_i's are weighting factors, $1 \le i \le 3$. It is obvious that the first two terms measure the variation of the angle and edge length, while these terms should be zero if the pivot point is at correct position because of the constraint of rigid body motion. The third metric represents a pulling force to impose restriction on pivot point not far away from the origin. Otherwise, one of the components of p^k will diverge, because the longer the length of vectors is, the smaller the angle between two vectors is, which implies that the first metric will be almost zero. Fortunately, we have preprocessed the 3D raw data according to domain knowledge, so the assumption that the pivot point is close to origin is reasonable. In other word, we only search the neighbor of the origin to get a more reliable pivot, instead of searching in the whole three-dimension space. To minimize the objective function $E_k(p^k)$ with unknown parameters *p(x, y, z)*, we follow the concept of the gradient descent algorithms.

3.3 Motion Compensation

The above two algorithms depend on the results of each other; in other words, for global motion estimation, it needs the position of rotation pivot, and for rotation pivot estimation, the translation offsets for each frame are required. Hence, we run these two procedures alternately to update the unknown parameters until these unknowns converge. Once the rotation matrix and translation vector for each frame are determined, the inverse of these affine transformations can be applied directly following equation (1) on feature points to get the new position without global motion.

4. SPEECH-DRIVEN FACIAL ANIMATION

4.1 Speech-driven Lip Synchronization

The synchronization of synthetic lip motion and the input speech is an important issue for video-realistic facial animation. In order to generate appropriate mouth shapes corresponding to input speech signal, one has to know what is the current utterance, and when the utterance starts and ends.

In this work, we include a commercial speech analysis package developed by Applied Speech Technologies [15] in our system. At this moment, our system is developed for Mandarin Chinese and English. We have developed the 14 visemes defined in MPEG-4 standard [1] in our system. The details of speech driven facial animation are described in our previous work [16].

Computer Animation and Simulation 2000, Springer, pp.29-40
(Proc. of the Eurographics Workshop in Interlaken, Switzerland)

4.2 Synthesis of Facial Expression

Adjusting captured data for head models

To apply the captured 3D motion data on a head model, firstly, we have to modulate the data to fit the facial features of the model. For feature points on the upper part of the face, the motion data are scaled according to the distances between two lower eyelids, and the distance between the forehead and nose tip. For points on the lower part of face, the data are scaled in proportion to the mouth width and the distance between nose tip and the chin; and the distance between the cheek and the lips determines the scale value in z-axis direction. How to interpolate the unmarked vertices and constraints for generating the facial animation are described in the following subsections.

Scatter Data Interpolation

After adjusting 3D motion data for a specified $2^1/_2$D head model, we can directly deform the feature points on the face mesh according to the modified motion data. However, we still have to construct a smooth interpolation function that gives the 3D displacements between the original points positions and the new position in the following frames for every vertex. Constructing such an interpolation function is standard problem in scattered data modeling. Given a set of known displacements $u_i = p_i - p_i^{(0)}$ away from the original positions $p_i^{(0)}$ at every constrained vertex i, which are the marker point on neutral face after motion compensation, we should construct a function that finds the displacement u_j for every unconstrained vertex j.

In different applications, various considerations should be taken to select a method for modeling scattered 3D data with minimum error. In our case, a method based on *radial basis functi*ons is adopted, that is, functions of the form

$$f(p) = \sum_i c_i \phi(\|p - p_i\|) + Mp + t \tag{6}$$

where $\phi(r)$ are radial symmetric basis functions. p_i is the constrained vertex; low-order polynomial terms M,t are added as affine basis. Many kinds of function for $\phi(r)$ have been proposed [22]. We have chosen to use $\phi(r) = e^{-r/64}$.

To determine the unknown coefficients c_i and the affine components M and t, we must solve a set of linear equations that includes $u_i = f(p_i)$, the constraints $\sum_i c_i = 0$ and $\sum_i c_i p_i^T = 0$. In general, if there are n feature point correspondences, we will have n+4 unknowns and n+4 equations with the following form:

Computer Animation and Simulation 2000, Springer, pp.29-40
(Proc. of the Eurographics Workshop in Interlaken, Switzerland)

Fig 5. The synthetic head rotates about the joint of the neck with exaggerated expressions.

$$\begin{bmatrix} \cdot & \cdot & \cdots & & p_{1x} & p_{1y} & p_{1z} & 1 \\ \cdot & e^{-\|p_i - p_j\|/64} & \cdots & & p_{2x} & p_{2y} & p_{2z} & 1 \\ \vdots & \vdots & \cdot & & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & p_{nx} & p_{ny} & p_{nz} & 1 \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 \\ p_{1x} & p_{2x} & \cdots & p_{nx} & 0 & 0 & 0 & 0 \\ p_{1y} & p_{2y} & \cdots & p_{ny} & 0 & 0 & 0 & 0 \\ p_{1z} & p_{2z} & \cdots & p_{nz} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \\ a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (7)$$

where $1 \le i, j \le 3 \quad P_i = (p_{ix}, p_{iy}, p_{iz})$.

Face Regions and Force Constraints

Since there are only 23 markers with our current captured data, and the facial actions of human beings are so subtle, some constraints must be applied to generate reasonable and smooth animation. In this system, we separate the head model into five regions: the hindbrain, the upper lip, the lower lip, the face, and the neck.

The hindbrain By the proposed method described in Section 3, we can compensate the global head motion, and thus the hindbrain becomes stationary when compared to the feature point motion. To avoid the abnormal motions of vertices intruding into this stationary region, some static points around the hindbrain are considered as feature points in the radial basis function computation. With this approach, those kinds of abnormal motions can be gradually reduced from the hindbrain.

Face We took all the feature points including static points as constrained vertices in the interpolation of facial vertices motion. This is because that the influence of a constrained vertex decreases exponentially according to the distance in our interpolation, and the effects of feature points far away are almost zero. However, taken all the vertex in the same "field of force" can help us to avoid the problem of discontinuousness at the boundary between different regions.

Upper and Lower lips Since the upper lip of human beings are controlled by muscles on the upper mouth and cheeks, the motion of vertices on the upper lip are interpolated from the marked feature points on the upper mouth and cheeks. Similarly, the vertices on the lower lips are interpolated from the marked feature points on the

lower mouth and the chin. In certain drastic motion involving lips, the discontinuity may occur at corners of the mouth. Some curve, such as Bezier curves or B-spline can be applied to smooth the boundary.

Neck The same as the hindbrain, the connected region of the neck and the head are stationary after applied global motion compensation, and some static points are also located at the connected region. After facial expressions are calculated over the face, we take the whole head as an object and rotate it about the joint of neck; then we can simulate one's nodding and head shaking. (as Fig. 5)

Emotion of Synthetic Face

Six expressions "neutral", "joy", "sadness", "anger", "fear", "disgust", and "surprise", specified in MPEG-4, are also defined in the proposed system. A facial expression with emotion is defined as following: (*i=1~9)*

$$Facial\ expression = basic\ viseme_i + \alpha \times Emotion\ vertex\ offset$$

where α is the degree of emotion intensity. As shown in Figure 6, an "emotion index" slider is drawn to change the emotion of the synthetic face.

5. WEB-ENABLED TALKING HEAD

In order to be web-enabled, our system must have characteristics of very-low bit-rate, short responsive 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; speech data can be encoded by CELP (Code Excited Linear Prediction) coding techniques such as G.723.1, the bandwidth requirement of our system "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 transmission 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.

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 can be implemented as a plug-in for web browser. When a user enter a web page with a link to a VRT file, our plug-in downloads the VRT data in streaming and plays back the speech with corresponding facial animation.

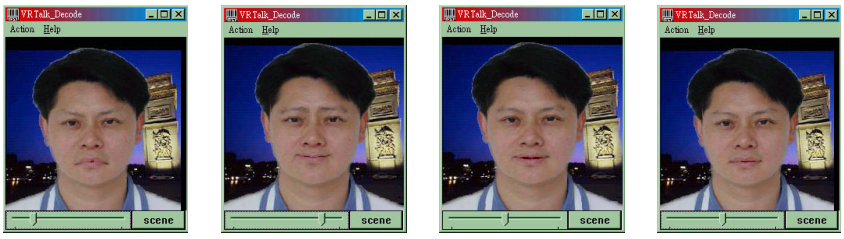


Fig 6. Facial expressions with an "emotion index" slider for real-time manipulation.

Computer Animation and Simulation 2000, Springer, pp.29-40
(Proc. of the Eurographics Workshop in Interlaken, Switzerland)

5.1 Compatibility with Microsoft's Advanced Streaming Format

The advanced streaming format (ASF) of MicroSoft recently becomes a popular streaming technology on PCs. The ASF streaming technology handles the streaming flow issue on Internet, and it also invokes the proper decoder specified in the streaming data to decompress the data after a block of data are accumulated.

The ASF format is a frame-based framework. Once a frame is received, the decoder must decompress the frame immediately and the raw data of color map should be sent to the render filter at the next step. With this issue, not only the key frames but also intermediate frames should be interpreted while data is being encoded. Figure 7 is VRT head and packet format of ASF stream. The required information of constructing the talking head such as triangles, textures, etc, is transmitted in the earliest frames. Frames of Facial Animation Parameters (FAPs) then follow the head model information.

5.2 System Implementation

In our self-defined VRT streaming data, images and speech data are major parts of it. To reduce the VRT streaming size, we adopt the JPEG image coding approach to encode the texture image and background; the speech coding standard G.723.1 with silent detection is applied to reduce the speech stream to less than 5.3 Kbps.

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 ASF streaming file, the facial animation part contains FAPs only. 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. A prototype talking head controlled by ASF bit-streams is also available. At this moment, since high-level FAPs are sufficient to describe vivid and realistic facial actions, and low-bit rate is also our target, not all 68 FAPs, but 9 high-level FAPs (viseme, expression, eye lid

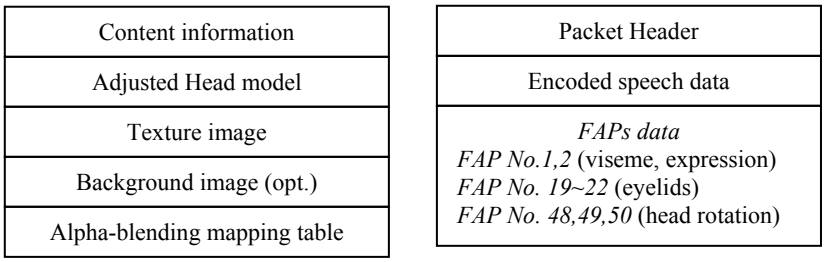


Fig 7. The header and packet format of VRT.

motion, and head rotation) defined in MPEG-4 are included in our streaming packets. These raw high-level FAPs data are transmitted frame-bye-frame without compression (should be done as specified in MPEG-4 with DCT, or arithmetic coding) and the bit-rate of animation control stream is about 8.6Kbps. Figure 8. is a figure of our talking head with a modified $2^1/2$D model of "Ananova" [26], who is a famous synthetic reporter originated from U.K. The original "Ananova" is rendered off line, and the streaming video is supported by Real Player [27]. With our web-enabled talking head techniques, the similar "Ananova" (we did this for comparative purpose only) requires only 14K bits per second including 5.3K sounds and the display window could be scaled up in resolution. The major difference is that our system follows the specifications of MPEG-4, and so, the streaming data contains not just video data, while the current implementation of "Ananova" is video based. Video compression based technology (H.263, e.g.) requires more than 56Kbps if QCIF format is used. However, as the MPEG-4 synthetic/natural hybrid coding specifies, we put 3D wireframes, textures, FAPs and speech into the streaming file. Therefore, our implementation can have a higher resolution (being model based) and yet requires less bandwidth.

For the time being, our system is developed on Windows 98/2000. Two kinds of web browsers, Internet Explorer (IE) and Netscape Navigator are supported. In addition, the Windows Media Player is also supported via the ASF stream. The MPEG-4 streaming format support is under development, and will be released in six months. On a Pentium III 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.

6. CONCLUSION AND FUTURE WORK

The proposed system "Web-enabled VR Talk" is a lifelike synthetic talking head. It now can be a vivid web-site presenter, and may also be used in "chat room like" applications on Internet. The demo web page of the proposed system is at http://www.cmlab.csie.ntu.edu.tw/~ichen/VRTalkDemo.html.

In addition to an ongoing project of "virtual meeting", some features of this system can be extended and improved. Captured facial motion data can be used to formulate the change between visemes, for instance co-articulation effects, as mathematical

Fig 8. Our talking head of the model "Ananovo" in ASF format.

models. Besides, how human's emotion will affect their mouth movement while speaking should also be analyzed. "View morphing" techniques can be applied to extend the range of view direction of 2½D head model. Compression techniques for triangles, bit-streams, etc. may be exploited to further reduce the bandwidth requirement.

ACKNOWLEDGMENTS

We would like to thank Digimax Production Center for providing the 3D facial motion data and technical supports in motion capture. The project is partially supported by the National Science Council of Taiwan, under the grant number NSC88-2622-E-002-002.

REFERENCES

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. Thaddeus Beier, Shawn Neely. Feature-Based Image Metamorphosis", Proc.of SIGGRAPH 92. Computer Graphics, pp. 35- 42, 1992.
12. Steven M.Seitz, Charles R. Dyer. View Morphing, Proc. SIGGRAPH 96, pp. 21-30.
13. 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.

14. Microsoft Speech Technology SAPI 4.0 SDK, http://www.microsoft.com/iit/projects/sapisdk.htm
15. Applied Speech Technologies Corporation. http://www.speech.com.tw
16. 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.
17. 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.
18. P.E Kmon, W.Fresen. Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, CA, 1978.
19. 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.
20. 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.
21. K. Waters and T. Levergood. An Automatic Lip-Synchronization Algorithm for Synthetic Faces. In Proceeding of ACM Multimedia, pp. 149-156, San Francisco, CA, USA, 1994, ACM Press.
22. Gregory M. Nielson. Scattered data modeling, in IEEE Computer Graphics and Applications, 13(1), pp.60-70, Jan. 1993.
23. Thomas S. Huang, and Arun N. Netravali. Motion and Structure from Feature Correspondences: A Review, in Proceedings of the IEEE, 82(2), pp. 252-268, Feb. 1994.
24. H. Goldstein. Classical Mechanics. MA: Addison Wesley, 1980.
25. S. D. Blostein and T. S. Huang. Algorithms for motion estimation based on 3-D correspondences, in Motion Understanding, W. Martin and J. K. Aggrawal, Eds. Norewell, MA: Kluwer, 1988.
26. "Ananova", http://www.ananova.com.
27. Real Player, http://www.realplayer.com.