International Journal of Image and Graphics, Vol. 1, No. 1 (2001)

IFACE: A 3D SYNTHETIC TALKING FACE

PENGYU HONG[*], ZHEN WEN[†], THOMAS S. HUANG[‡]
Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

We present the iFACE system, a visual speech synthesizer that provides a form of virtual face-to-face communication. The system provides an interactive tool for the user to customize a graphic head model for the virtual agent of a person based on his/her range data. The texture is mapped onto the customized model to achieve a realistic appearance. Face animations are produced by using text stream or speech stream to drive the model. A set of basic facial shapes and head action is manually built and used to synthesize expressive visual speech based on rules.

Keywords: iFACE; Face Modeling and Animation; Visual-Speech.

1. Introduction

Face modeling and animation have attracted researchers in several disciplines, including computer vision, computer graphics and cognitive science.[1] A graphics-based human model provides an effective solution for information display, especially in collaborative environments. Examples include 3D model-based very low bit-rate video coding for visual telecommunicaiton,[2,3] audio/visual speech recognition,[4] and talking head representation of computer agent.[5] In noisy environments, the synthetic talking face can help users to understand the associated speech, and it helps people react more positively in interactive services.[6] Researchers showed that a virtual sales agent inspires confidence in customers in the case of e-commerce, and a synthetic talking face is also found to assist students learn better in computer-aided education.

In the past few years, our research has successfully led to a system called iFACE, which provides functionalities for face modeling and animation. The 3D geometry of a face is modeled by a triangular mesh. A few control points are defined on the face mesh. By dragging the control points, the user can construct different facial shapes. Two kinds of media, text stream and speech stream, can be used to drive the face animation. The phoneme information is extracted form text streams and speech streams and is used to determine viseme transitions. The time information is extracted from either the synthetic speech or the natural speech to decide the rate at which the face animation process should occur. In this way, the system is able to synchronize the generated visual stream and audio stream.

[*] E-mail: hong@ifp.uiuc.edu
[†] E-mail: zhenwen@ifp.uiuc.edu
[‡] E-mail: huang@ifp.uiuc.edu

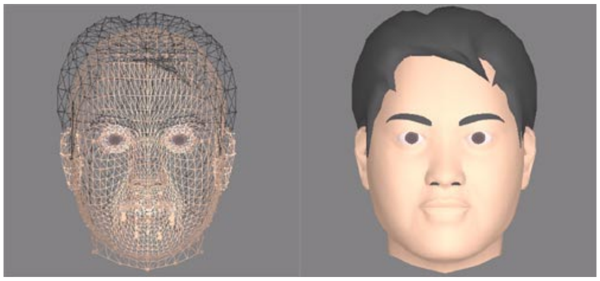
International Journal of Image and Graphics, Vol. 1, No. 1 (2001)

2. Head Modeling

A realistic 3D head model is one of the key factors of natural human computer interaction. Parke was probably the first to use a parameterized polygonal facial model for computer face animation.[7] In recent years, researchers have been trying several ways to build realistic head models. Lee used Cyberware^{TM} scanner to obtain 3D face range information for facial geometry modeling.[8] The texture map extracted by the Cyberware^{TM} scanner is mapped onto the facial geometrical model to increase visual realism. There are also some photogrammetric methods for modeling facial geometries from a number of photographs taken of a human subject. [9,10,11,12]

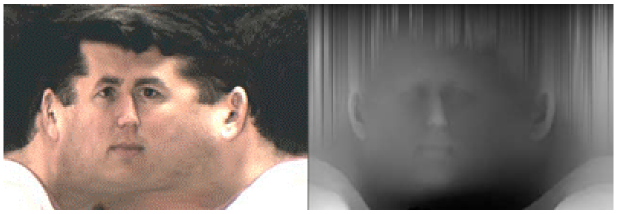
2.1. Customizing head model

In iFACE, the process of making a head model is nearly automatic with only a few manual adjustments necessary. We start with a generic model, which is shown in Fig. 1. The head model consists of all the head accessories such as face, eyes, teeth, tongue, ears, etc. Currently, it consists of 2240 vertices, 2946 triangles, and a non-uniform rational b-splines (NURBS) sub-model with 63 control points. The facial surface is approximated by triangular meshes. To customize the head model for a particular person, we first obtain both the texture data and range data of that person by scanning his/her head using Cyberware^{TM} cyberscanner. An example of the cyberscanner data is shown in Fig. 2.



(a) (b)

Fig. 1. The generic head model. (a) Shown as wire-frame and (b) shown as shaded.



(a) (b)

Fig. 2. Cyberscanner Data. (a) Texture map and (b) range Map.

Thirty-five feature points are defined on the face component of the generic head model. If we unfold the face component of the head model onto 2D, those feature points triangulate the face mesh into several local patches. The 2D locations of the feature points in the range map are manually selected on the scanned texture data (see Fig 3(a)). The system calculates the 2D positions of the remaining face mesh vertices on the range map by deforming the local patches based on the range data. By collecting the range information according to the positions of the vertices on the range map, the 3D facial geometry is decided. The remaining head components are automatically adjust by shifting, rotating, and scaling. Manual adjustments on the fitted model are required where the scanned data are missing. Texture map is mapped onto the customized model to achieve photo-realistic appearance. Fig 3(b) shows an example of a customized head model.

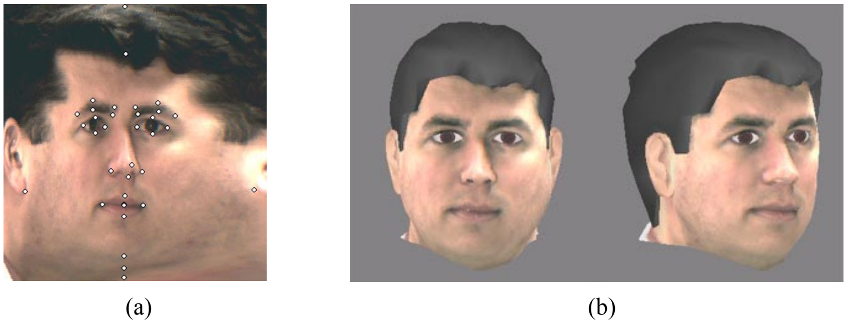


Fig. 3. Customize head model. (a) Feature points and (b) the customized face model.

2.2. *Synthesize facial expressions and head motion*

A control model is defined on the face component (see Fig. 4). The control model consists of 101 vertices and 164 triangles. It covers the facial region and divides it into local patches. For each triangle we define a local affine deformation that is applied to a patch of face region. Changing the 3D positions of the feature points, the user can manually deform the shape of the face component. A set of basic facial shapes is built by adjusting the control points of the face model. Those basic facial shapes are in spirit similar to the Action Units of Ekman.[13] They are built so that all kinds of facial expressions can be approximated by linear combinations of them. Some examples of the basic shapes used by iFACE are shown in Fig. 5. Some head actions, such as nodding, shaking, etc., are also predefined by specified the values of six action parameters and their temporal patterns. Given a script of expression sequence, we can use key frame technique to synthesize expressive animation sequences, such as nodding head, eye blinking, raising eyebrow, etc. By combining the script of expression and text, we can generate an expressive talking head.

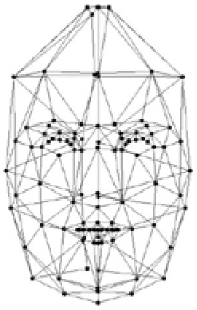


Fig. 4. Control model.

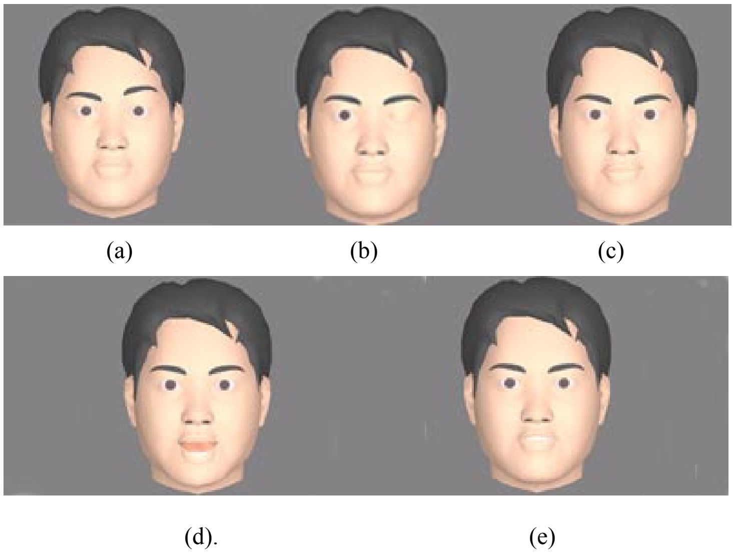


Fig. 5. Examples of basic facial shapes. (a) Raise left eyebrow, (b) close left eye, (c) raise left mouth corner, (d) open lower lip, and (e) raise upper lip.

3. Text and Speech Driven Face Animation

3.1. Text Driven Face Animation

When text is used in communication, for example, in the context of text-based electronic chatting over Internet and visual email, visual speech synthesized from text will greatly help deliver information. Recent work on text driven face animation includes the work of Waters and Levergood, [14] Cohen and Massaro,[15] LeGoff and Benoit,[16] and Ezzat and Poggio.[17] Our work is close in spirit to that of Waters and Levergood.[14] iFACE uses Microsoft Text-to-Speech (TTS) engine for text analysis and speech synthesis. The structure of text driven face animation is illustrated in Fig. 6.

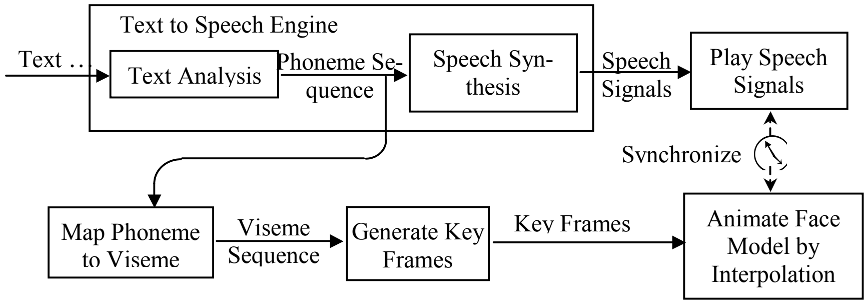


Fig. 6. The architecture of text driven talking face.

First, the text is fed into TTS. TTS parses the text and generates a phoneme sequence, timing information and synthesized speech stream. Phoneme sequence is mapped to the viseme sequence based on a lookup table. To synthesize the animation sequence, we adopt a key frame based scheme similar to Waters and Levergood.[14] Visemes as key frames are located on the starting utterance frames for each phoneme at intervals indicated by phoneme durations. The face shapes between key frames are decided by an interpolation scheme.

3.2. Speech Driven Face Animation

When human speech is used in one-way communication, for example, news broadcasting over the networks, offline speech driven talking face is required. The process of offline speech driven face animation is illustrated in Figure 7. An advantage of the offline process is that the phoneme transcription and timing information can be precisely extracted for accurate animation. Recognizing phonemes using only speech signals requires a complicated continuous speech recognizer, and the phoneme recognition rate and the timing information may not be accurate enough. The text script associated with speech, however, provides the accurate word-level transcription so that it not only greatly simplifies the complexity of phoneme recognition but also helps improve the recognition rate. iFACE uses a phoneme speech alignment tool, which comes with HTK 2.0 for phoneme recognition and alignment. Speech stream is decoded into phoneme sequence with duration information. Once we have the phoneme sequence and the timing information, the remaining part of the procedure of the visual speech synthesis is similar to text driven face animation. Figure 8 shows an example of offline speech driven face animation.

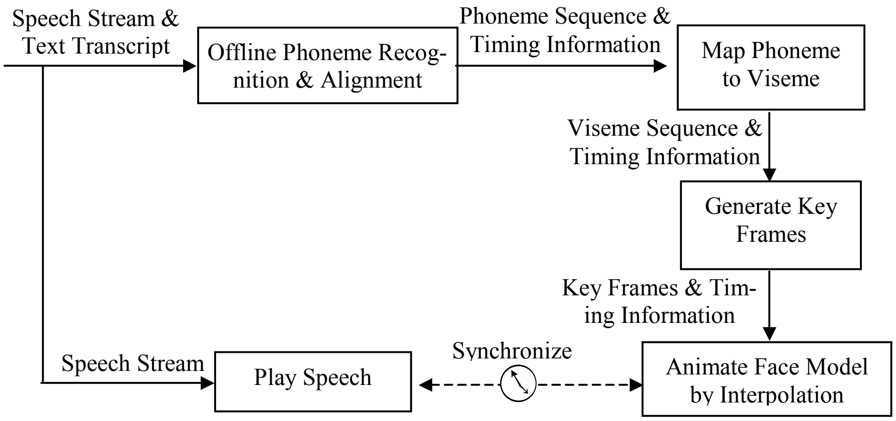
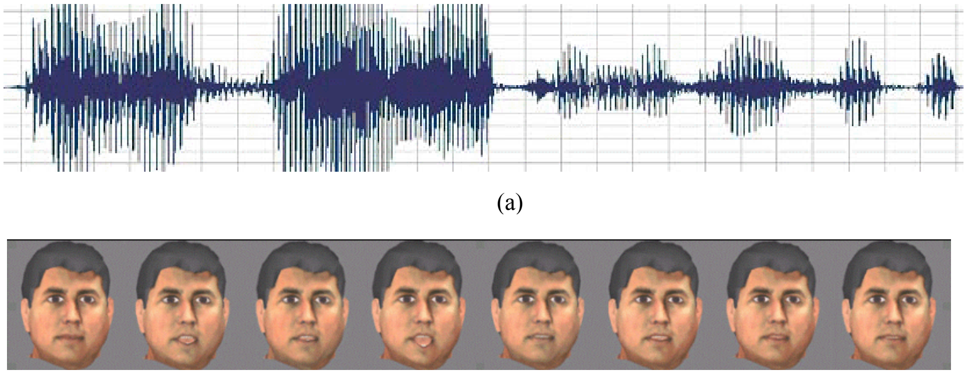


Fig. 7. The architecture of offline speech driven talking face.

4. Conclusions

This paper gives an overview of an expressive talking head system with text-to-animation conversion scheme and voice-to-animation scheme implemented. The system can be used to construct a realistic 3D model and synthesize natural facial animation from text, voice and emotion states. The system is useful for applications such as human computer intelligent interfaces, collaborative applications, computer language education, and automatic animation productions. Further research will be conducted to achieve real-time voice driven facial animation suitable for two-way communication, to evaluate the effectiveness of the system by user study, and to extend the 3D model to include the upper body.



(a)

(b)

Fig 8. An example of speech driven face animation. (a) A sample sound track: “I want high-res images.” and (b) The typical frames in the corresponding face animation sequence.

Acknowledgements

This work was supported by Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003.

References

1. P. Ekman, T. S. Huang, T. J. Sejnowski and J. C. Hager, eds., *Final report to NSF of the planning workshop on facial expression understanding*, Human Interaction Laboratory, University of California, San Francisco, March, 1993.

2. K. Aizawa and T. S. Huang, “Model-based image coding”, *Proc. IEEE*, 83 (1995) 259-271.

3. H. Li, P. Roivainen and R. Forchheimer, “3-D motion estimation in model-based facial image coding”, IEEE Trans. On Pattern Analysis and Machine Intelligence 15, 6 (1993) 545-555.

4. D. G. Stork and M. E. Hennecke, eds., *Speechreading By Humans and Machines*, *NATO ASI Series*, (Springer, 1996).

5. K. Nagao and A. Takeuchi, “Speech dialogue with facial displays,” in *Proc. 32nd Annual Meeting of the Asso. for Computational Linguistics* (ACL-94), (1994) 102-109.

6. I. Pandzic, J. Ostermann, D. Millen, “User evaluation: Synthetic talking faces for interactive services,” The Visual Computer, vol. 15, Issue 7/8, 4 November 1999, pp. 330-340.

7. F. I. Parke, A parametric model of human faces. Ph.D. Thesis, University of Utah, 1974.

8. Y. Lee, D. Terzopoulos, K. Waters, “Realistic modeling for facial animation”, in *Proc. SIGGRAPH ’95*, (Los Angeles 1995), 55-62.

9. F. Pighin, J. ecker, *et al.*, “Synthesizing realistic facial expressions from photographs”, *in Proc. SIGGRAPH ’98*, 1998.

10. B. Guenter, C. Grimm, D. Wood, et al. “Making faces”, in *Proc. SIGGRAPH '98*, 1998.

11. V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” in *Proc. SIGGRAPH’99*, 1999.

12. T.S. Huang and L. Tang, “3-D face modeling and its applications”, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 10, No. 5, (1996) 491-520.

13. P. Ekman and W.V. Friesen, *Facial Action Coding System: Investigator’s Guide,* Palo Alto, CA: Consulting Psychologist Press, 1978.

14. K. Waters and T. M. Levergood, DECface, “An Automatic Lip-Synchronization Algorithm for Synthetic Faces,” Digital Equipment Corporation, Cambridge Research Lab, Technical Report CRL 93-4.

15. M.M. Cohen and D.W. Massaro, “Modeling coarticulation in synthetic visual speech”, in *Models and Techniques in Computer Animation*, ed. N.M. Thalmann and D. Thalmann, (Springer-Verlag: Tokyo), 1993, p. 139-156.

16. B. LeGoff and C. Benoit, “A text-to-audiovisual-speech synthesizer for french”, in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

17. T. Ezzat and T. Poggio, “Visual Speech Synthesis by Morphing Visemes”, *International Journal of Computer Vision* 38(1), (2000) 45-57.

Photo and Bibliography

Pengyu Hong received the B. Engr and M. Engr degrees, both in Computer Science, from Tsinghua University, Beijing, China, in 1995 and 1997, respectively. Since August 1997, he has been a Ph.D. student in the Department of Computer Science at the University of Illinois at Urabana-Champaign, Urbana, Illinois, USA. His research current interests are face modeling, facial motion analysis and synthesis, temporal and spatial pattern searching and understanding in pattern recogniton and machine learning.

Zhen Wen received the B. Engr degree from Tsinghua University, Beijing, China, and MS degree from the University of Illinois at Urabana-Champaign, Urbana, Illinois, USA, both in Computer Science. Currently he is a Ph.D. student in the Department of Computer Science at the University of Illinois at Urabana-Champaign. His research interests are face modeling, facial motion analysis and synthesis, image based modeling and rendering.

Thomas S. Huang received his B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, China; and his M.S. and Sc.D. degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973, and on the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering,

International Journal of Image and Graphics, Vol. 1, No. 1 (2001)

Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology.

During his sabbatical leaves, Dr. Huang has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, and the Rheinishes Landes Museum in Bonn, West Germany, and held visiting professor positions at the Swiss Institutes of Technology in Zurich and Lausanne, University of Hannover in West Germany, INRS-Telecommunications of the University of Quebec in Montreal, Canada, and University of Tokyo, Japan. He has served as a consultant to numerous industrial firms and government agencies both in the U.S. and abroad.

Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 12 books and over 300 papers in network theory, digital filtering, image processing, and computer vision. He is a Fellow of the International Association of Pattern Recognition, IEEE, and the Optical Society of American; and has received a Guggenheim Fellowship, an A.V. Humboldt Foundation Senior U.S. Scientist Award, and a fellowship from the Japan Association for the Promotion of Science. He received the IEEE Acoustics, Speech, and Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He is a Founding Editor of the International Journal Computer Vision, Graphics, and Image Processing, and Editor of the Springer Series in Information Sciences, published by Springer-Verlag.