ANIMATED TALKING HEAD WITH PERSONALIZED 3D HEAD MODEL[1]

L.S. Chen, T.S. Huang
Beckman Institute & CSL
University of Illinois
Urbana, IL 61801, USA
lchen@ifp.uiuc.edu

J. Ostermann
AT&T Labs-Research
Holmdel, NJ 07733, USA
osterman@research.att.com

Abstract - Natural Human-Computer Interface requires integration of realistic audio and visual information for perception and display. An example of such an interface is an animated talking head displayed on the computer screen in the form of a human-like computer agent. This system converts text to acoustic speech with synchronized animation of mouth movements. The talking head is based on a generic 3D human head model, but to improve realism, natural looking personalized models are necessary. In this paper we report results in adapting a generic head model to 3D range data of a human head obtained from a 3D laser range scanner. This personalized model is incorporated into the talking head system. With texture mapping, the personalized model offers a more natural and realistic look than the generic model.

INTRODUCTION

Human Computer Interface is an application area where audio, text, graphics, and video are integrated to convey various types of information. Often conversion is necessary between different media [1]. The objective is to provide more natural interaction between the human user and computer. One approach is to display an animated character or a life-like talking head on the computer screen, with the ability to receive input from the user and respond in a natural and intelligent way. Such an agent may perform information retrieval, reading or replying email messages. Already available are software programs that display a generic animated talking head or a cartoon animal character on the screen to perform various tasks. One program is able to fetch songs at the user's request [2]. Another is able to carry on simple conversations [3]. Yet another is able to convert ASCII text into synthetic speech and synchronized talking head with very realistic lip and jaw movements, with visible teeth and tongue [4].

Most programs use a generic model that is deformable according to a set of parameters. However, to make such interaction more resemblant to that

[1]This work was performed while Lawrence Chen was a summer intern at AT&T Labs-Research.

0-7803-3780-8/97/$10.00 ©1997 IEEE

of a human-to-human interaction, realistic and natural looking displays are required. This calls for models that are based on real measurements of the structures of the human face, as well as facial features such as color, shape, and size. Such information is available through the use of a 3D laser scanner that produces very dense range data of the human head. In addition it also provides the corresponding color image of the head. These information can be utilized to achieve a more natural looking talking head than the generic model.

This paper describes preliminary results of fitting a generic head model to a set of 3D range and color data. The fitted model is incorporated into an existing text-to-speech animated talking head program, and the color image is used for texture mapping for animation. With the personalized model, the interface is more natural looking and more believable.

ANIMATED TALKING HEAD

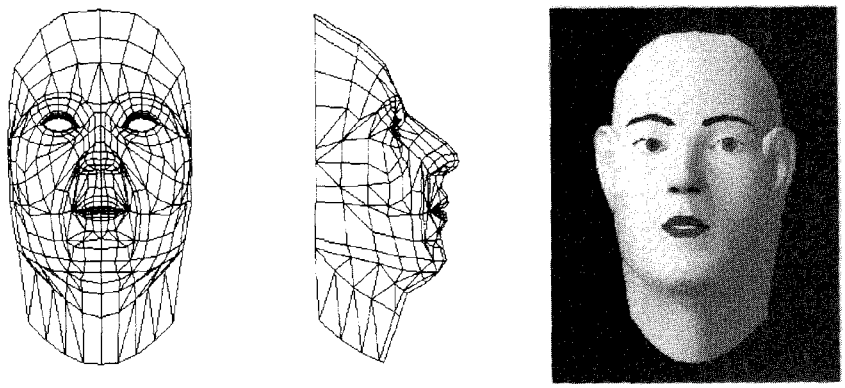
We first describe the animated talking head system used in our work. This system is based on an AT&T text-to-speech(TTS) program. The program converts text to synthetic speech, and displays a generic animated talking head with realistic lip movements synchronized to the speech. The text is parsed and analyzed extensively to produce the speech, and to provide phonetic information for animating mouth movements. The animation subsystem uses a parameterized 3D model, which is a descendant of Parke's model [5], further improved with coarticulation models for synthetic visual speech research at UC Santa Babara [6]. The structure of the 3D model is a wireframe of numbered vertices in 3D coordinate space, with connections specified to form polygons (Figure 1(a)). Prescribed colors are added to each polygon to form smooth-shaded surfaces (Figure 1(b)). This sophisticated model includes the face, eyes, mouth, teeth and tongue, and is capable of producing very realistic mouth movements. It is controlled by a set of deformation parameters. With a set of time-varying parameters, an animation sequence is produced. Occasional head movement and eye blinking are added heuristically.

This system gives satisfactory performance with synchronized speech and lip movements. However, this talking head may seem impersonal because it does not represent any person the human user may know. If one can produce personalized models, the interface would improve, and the user may even have the choice of selecting from models of several different persons.

Our goal is to fit the generic model to 3D range data of a person's head, and this fitted model replaces the generic model in the talking head system to give a personalized display.

3D RANGE DATA

The 3D scanner used is the Cyberware 3D laser range scanner. The subject sits in a chair while the scanner revolves around the person to scan the surface structure of the head. The scanner gives a very dense set (over 260,000 points)



(a) (b)

Figure 1: (a)Wireframe of the generic model;(b)Smooth shaded talking head.

of range data in cylindrical coordinates. In addition to the range data, a color camera captures the surface color the head, so for each range point the corresponding color is also available. This is useful for texture mapping in the final rendering of the talking head.

Carefully scanned dataset ensures that the head is upright and centered. Software manipulation is possible if the head is slightly off-center via coordination transformation. Figure 2 shows two rendered views of the 3D range data.

Figure 2: Two views of the 3D range data.

MODEL FITTING

In this section we describe fitting of the generic model to the 3D range data. Initially, the generic head model is larger in scale than the range data. Vertical scaling factor is first obtained to scale down the model, and the

vertical profile line down the center of the face is fitted. After the profile line is fitted, the rest of the face is fitted through radial projection.

The range scanner gives a data set in the left-handed cylindrical coordinate system (r_L, y_L, ϕ_L), while the wireframe model is in the usual right-handed Cartesian coordinate system (x, y, z). They are transformed into the same right-handed cylindrical coordinate system (r_R, y_R, ϕ_R) by

$$r_R = r_L, y_R = y_L, \phi_R = 2\pi - \phi_L \tag{1}$$

and

$$r_R = \sqrt{z^2 + x^2}, y_R = y, \phi_R = \arctan\left(\frac{x}{z}\right). \tag{2}$$

To find the vertical profile line, we first find the tip of the nose. The tip of the nose is generally the most protruded structure of the face, i.e., its distance is greatest to an imaginary vertical rotation axis in the center of the head. This point is selected manually, but it can also be detected automatically. After the point is obtained, the profile line can be determined. From the profile line, feature points such as the indent above the nose, the upper lip, mouth center, and the lower lip can be detected automatically. The location and separation distance of these features are used to give vertical and scaling factor.

This scaling factor is used to scale down the generic model to the size of the range data in the vertical direction. Then each vertex point on the model is radially projected onto the surface of the range data. For each vertex point, we have the height y_R and the angle ϕ_R in the cylindrical coordinate system. Then we trace a ray radially back towards the vertical axis y, piercing the range data. This ray may not intercept the range data at exactly one range point, so the nearest four range data points are averaged to give the new coordinate for the model vertex. The radial projection is depicted in Figure 3.

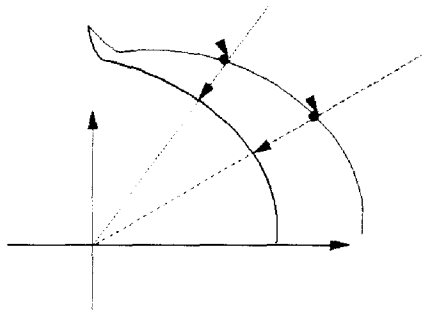


Figure 3: Radial projection: The inner curve is the range data, and the outer curve is the wireframe model. The places where the rays intercept the range data are the new coordinates for the model.

This fits the skin part of the generic model to the range data, but the eye and mouth positions are critical and need to be adjusted manually. Right now only the face is fitted, but the procedure can be easily extended to fit the entire head including the back of the head.

The fitted wireframe (Figure 4) is then ready to be incorporated into the talking head system.

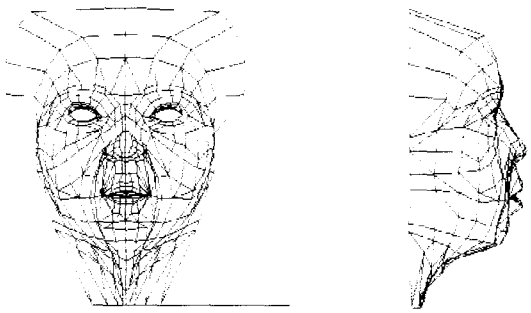


Figure 4: Fitted wireframe model.

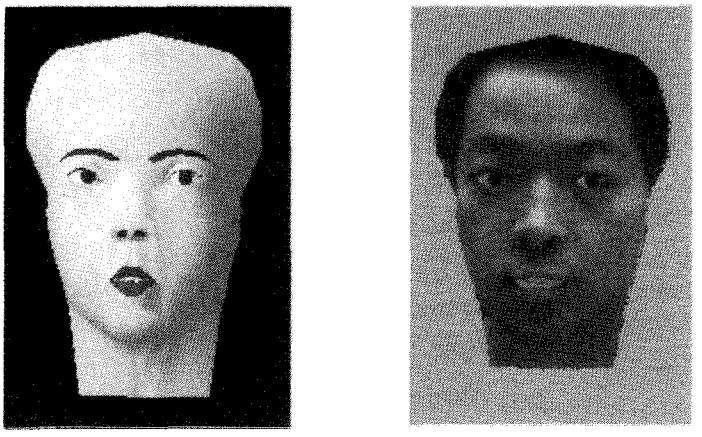
RESULTS

Figure 5(a) shows the talking head with the new model with smooth shading. At first glance, it does not appear to be much better than the original model, because the original model does not contain hair, but the 3D range data contains the structure of hair. This problem is overcome by texture mapping. Instead of smooth shaded polygons with prescribed colors, the color and texture of the polygon surface of the face come from a color image. The image is literally pasted onto the polygonal wireframe. Geometric transformation is done automatically by bilinearly interpolating the texture through the computer graphics library routines. Figure 5(b) is the texture mapped rendering of the head. Clearly the resulting rendering is much more realistic.

CONCLUSIONS

In this paper we described the fitting of a personalized 3D model for a talking head software. The fitted wireframe model with texture mapping looks more realistic and more believable than a simple generic model.

We are now working on adding the back of the head as well as body to the model, so the talking head does not appear to be floating in space. Also in progress is work to further automate the fitting process and texture mapping.



(a) (b)

Figure 5: Fitted model with (a)smooth shading; (b)texture mapping.

Another challenge is to include expressions into the system, so the talking head would smile once in a while.

References

[1] S. Morishima and H. Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface." *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 594–600, May 1991.

[2] D. Kurlander and D. T. Ling, "Planning-Based Control of Interface Animation." *Microsoft Research Technical Report MSR-TR-95-21.* January 1995. Microsoft Research, Redmond, WA, USA.

[3] K. Nagao and A. Takeuchi, "Speech Dialogue With Facial Displays: Multimodal Human-Computer Conversation." *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics,* pp. 102–109, 1994.

[4] R. Sproat and J. Olive, "An Approach To Text-to-speech Synthesis." In W.B. Kleijn & K.K. Paliwal (Eds.) *Speech Coding and Synthesis*, Elsevier Science, 1995.

[5] F. I. Parke, "Parameterized Models for Facial Animation." *IEEE Computer Graphics and Applications*, vol. 2, pp. 61–68, Nov. 1982.

[6] M. M. Cohen and D. W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech." In M. Thalmann & D. Thalmann (Eds.) *Computer Animation '93.* Tokyo: Springer-Verlag.