Real Face Communication in a Virtual World

Won-Sook Lee, Elwin Lee, Nadia Magnenat Thalmann

MIRALab, CUI, University of Geneva
24, rue General-Dufour, CH-1211, Geneva, Switzerland
Tel: +41-22-705-7763 Fax: +41-22-705-7780
E-mail: {wslee, lee, thalmann}@cui.unige.ch

Abstract. This paper describes an efficient method to make an individual face for animation from several possible inputs and how to use this result for a realistic talking head communication in a virtual world. We present a method to reconstruct 3D facial model from two orthogonal pictures taken from front and side views. The method is based on extracting features from a face in a semiautomatic way and deforming a generic model. Texture mapping based on cylindrical projection is employed using a composed image from the two images. A reconstructed head is animated immediately and is able to talk with given text, which is transformed to corresponding phonemes and visemes. We also propose a system for individualized face-to-face communication through network using MPEG4.

Keywords: Cloning, orthogonal pictures, DFFD, texture mapping, talking head, visemes, animation

1. Introduction

Individualized facial communication is becoming more important in modern computer-user interfaces. To visualize ones own face in a virtual world and let people talk with given input, such as text or video, is now a very attractive research area. With the fast pace in computing, graphics and networking technologies, real-time face-to-face communication in a virtual world is now realizable. It is necessary to reconstruct an individual head in an efficient way to decrease data transmission size, and send few parameters for real-time performance.

Cloning a real person's face has practical limitations in the sense of time, simple equipment and realistic shape. We present our approach to clone a real face from two orthogonal views, emphasizing accessibility for anybody with low price equipment. Our method to give animation structure on a range data from a laser scanner or stereoscopic camera is also described, but not in detail. This is because of the high price of this equipments. Therefore, it may not be a very practical idea to make use of them. The main idea is to detect feature points and modify a generic model with

animation structure. After creating virtual clones, we can use them to animate and talk in a virtual world.

The organization of this paper is as follows. We give a review in Section 2 with classification for existing methods to get a realistic face reconstruction and talking head. In Section 3, we describe the idea of a system for individualized face-to-face communication through a network and our system for creating/animating talking head with given text. Section 4 is dedicated to the reconstruction process from feature detection to texture mapping and then the detailed process for talking head is explained in Section 5. Finally conclusion is given.

2. Related Work

2.1. Face Cloning for Animation

There have been many approaches to reconstruct a realistic face in a virtual world. There are many possible ways such as using a plaster model [12][1], or interactive deformation and texture mapping [2][15], which are time-consuming jobs. More efficient methods are classified into four categories.

Laser Scanning In range image vision system some sensors, such as laser scanners, yield range images. For each pixel of the image, the range to the visible surface of the objects in the scene is known. Therefore, spatial location is determined for a large number of points on this surface. An example of commercial 3D digitizer based on laser-light scanning is Cyberware Color Digitizer^{TM} [14].

Stripe Generator As an example of structured light camera range digitizer, a light striper with a camera and stripe pattern generator can be used for face reconstruction with relatively cheap equipment compared to laser scanners. With information of positions of projector and camera and stripe pattern, a 3D shape can be calculated. Proesmans et al. [13] shows a good dynamic 3D shape using a slide projector, by a frame-by-frame reconstruction of a video. However, it is a passive animation and new expressions cannot be generated.

Stereoscopy A distance measurement method such as stereoscopy can establish the correspondence at certain characteristic points. The method uses the geometric relation over stereo images to recover the surface depth. C3D 2020 capture system [3] by the Turing Institute produces many VRML models using stereoscopy method.

Most of the above methods concentrate on recovering a good shape, but the biggest drawback is that they provide only the shape without structured information. To get a structured shape for animation, the most typical way is to modify an available generic model with structural information such that eyes, lips, nose, hair and so on. Starting with a structured facial mesh, Lee et al. [13] developed algorithms that automatically construct functional models of the heads of human

subjects from laser-scanned range and reflection data [14]. However, the approach based on 3D digitization to get range data often requires special purpose high-cost hardware. Therefore, a common way of creating 3D objects is the reconstruction from 2D photo information, which is accessible at a low price.

Modification with Feature Points on Pictures There are faster approaches to reconstruct a face shape from only a few pictures of a face. In this method, a generic model with an animation structure in 3D is provided in advance and a limited number of feature points. These feature points are the most characteristic points to recognize people, detected either automatically or interactively on two or more orthogonal pictures, and the other points on the generic model are modified by a special function. Then 3D points are calculated by just combining several 2D coordinates. Kurihara and Arai [9], Akimoto et al. [4], Ip and Yin [7], and Lee et al. [16] use an interactive, semiautomatic or automatic methods to detect feature points and modify a generic model. Some have drawbacks such as too few points to guarantee appropriate shape from a very different generic head or accurate texture fitting, or automatic methods, which are not robust enough for satisfactory result, like simple filtering and texture image generation using simple linear interpolation blending.

2.2. Talking Head

There are two approaches for the synthesis of talking heads. Pearce et al. [17] have used an approach to create an animation sequence with the input being a string of phonemes corresponding to the speech. In this case, the face is represented by a 3D model and animated by altering the position of various points in the 3D model. A the more recent work of Cohen and Massaro [18], english text is used as the input to generate the animation. The alternative is the image-based morphing approach. Ezzat and Poggio [19] have proposed a method of concatenating a collection of images, using a set of optical flow vectors to define the morphing transition paths, to create an animated talking head. In another recent work of Cosatto and Graf [20], they have proposed an automatic method of extracting samples from a video sequence of a talking person using image recognition techniques. In this case, the face image is being partitioned into several facial parts and later combined to generate a talking head. They focused to reduce the number of samples that are needed and photo-realistic movements of lips.

2.3. Shared Virtual World

There have been numerous works being done on the topic of shared virtual environments. In most of these works [23][24][25], each user is represented by a fairly simple embodiment, ranging from cube-like appearances, non-articulated human-like or cartoon-like avatars to articulated body representations using rigid body segments. In the work of Pandzic et al. [26], a fully articulated body with skin

deformations and facial animation is used. However, none of these works discussed about the process of creating individual face model in an efficient manner. In other words, each individual face model usually takes hours of effort to complete. The MPEG-4 framework [30] under standardization can be used to achieve a low bandwidth face-to-face communication between two or more users, using individualized faces.

3. System Overview

In networked collaborative virtual environments, each user can be represented by a virtual human with the ability to feel like "being together" by watching each other's face. Our idea for face-to-face communication through a network is to clone a face with texture as starting process and send the parameters for the reconstruction of the head model to other users in the virtual environment. After this, only animation and audio parameters are sent to others in real time to communicate. The actual facial animation can be done on a local host and all the users can see every virtual human's animation. This provides a low bandwidth solution for having a teleconference between distanced users, compared to traditional video conferencing, which sends video stream in real time through network. At the same time, it is able to retain a high level of realism with individualized textured 3D head.

The general idea of a system for individualized face-to-face communication through network is shown in Fig. 1. There are two hosts, host1 and host2. Each host has a generic model and a program to clone a person from a given input such as two orthogonal pictures or range data. A person in host1 is reconstructed with pictures or range data obtained in any range data equipment. The parameters for reconstruction will be sent to host2 through network. They are texture image data, texture coordinates and either modification parameters or geometric position data for points on a 3D head surface. The same procedure happens in host2. Finally, both host1 and host2 have two 3D heads which are textured. Although we send a texture image, which takes large bandwidth, it is only one time process for reconstruction. After this process, only animation parameters and audio data are sent to another host through network. With animation parameters given, each host will animate two heads in their own platforms.

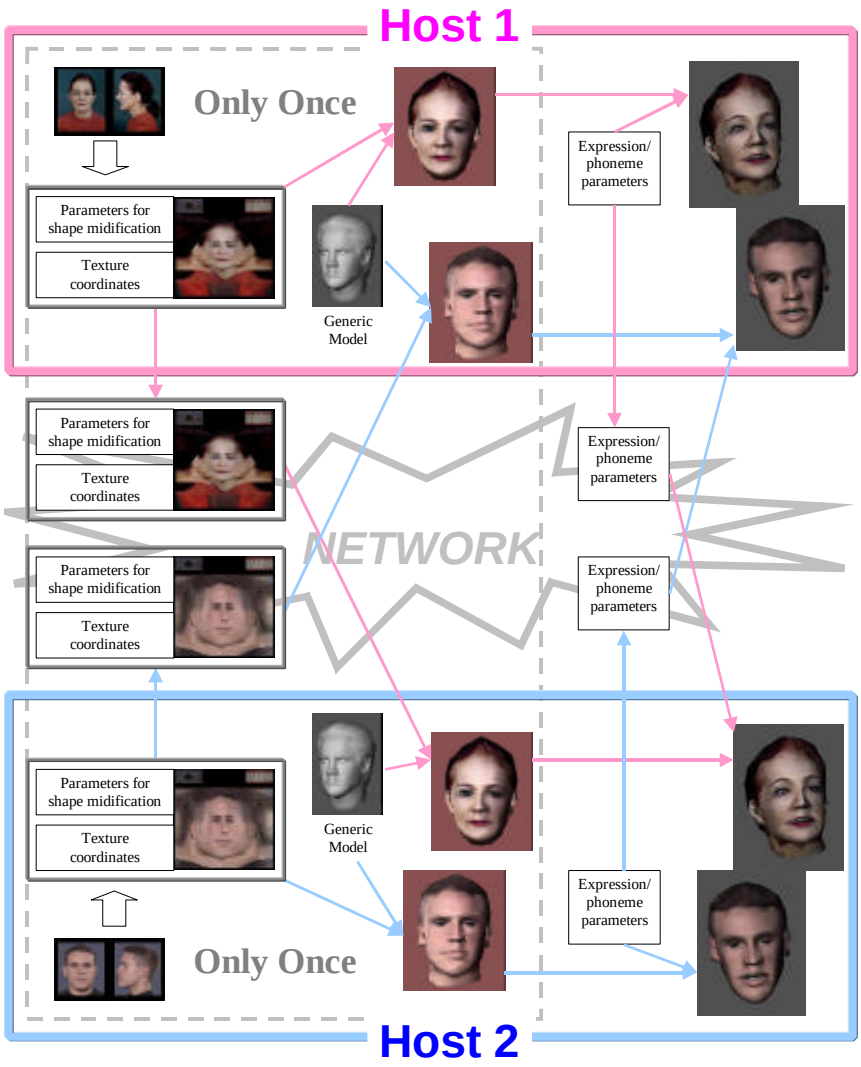

Fig. 1. A system overview for individualized face-to-face communication through network.

The part of this system showing the proceedure of producing an individualized talking head is shown in Fig. 2. We reconstruct a real face from a given input. For reconstruction, only some points (so called feature points) are extracted from front and side views or only from front view if range data is available and then a generic model is modified. After reconstructing a head, it is ready to animate with given animation parameters. We use this model and apply it to talking head, which has animation abilities given text input. The input speech text is transformed to corresponding phonemes and animation parameters, so called visemes. In addition, expression parameters are added to produce final face with audio output.

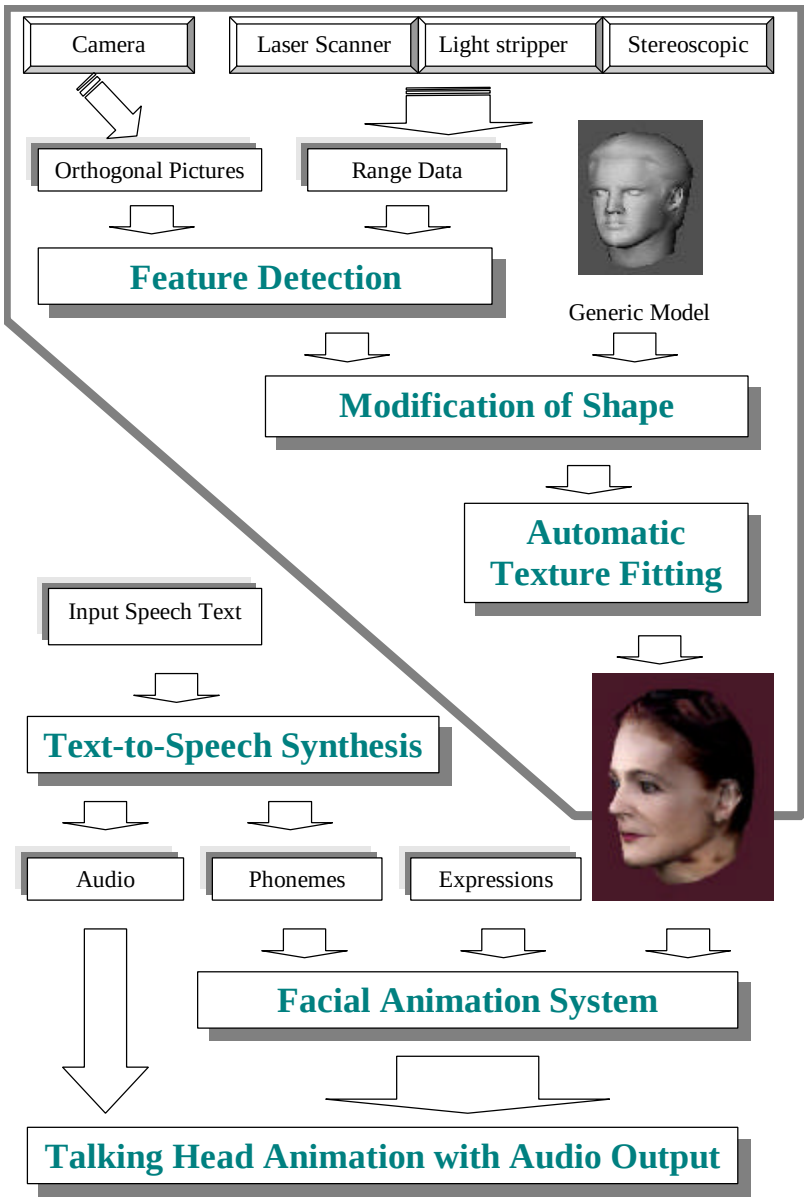

Fig. 2. Overall structure for individualized talking head.

4. Face Cloning

We reconstruct a face from two kinds of input, such as an orthogonal picture or range data in VRML format. The main idea is to modify a generic animation model with detected feature points and apply automatic texture mapping. In this paper, we focus on orthogonal picture input rather than range data input since the approach based on 3D digitization to get a range data often requires special purpose (and high-cost) hardware and the process is quite similar to orthogonal picture case.

We consider hair outline, face outline and some interior points such as eyes, nose, lips, eyebrows and ears as feature points.

4.1. Preparation and Normalization

First, we prepare two 2D wire frames composed of feature points with predefined relations for front and side views. The frames are designed to be used as an initial position for the snake method later. Then we take pictures from front and side views of the head. The picture is taken with maximum resolution and the face is in the neutral expression and pose.

To make the head heights of side and front views the same, we measure them, and choose one point from each view to matching them with corresponding points in prepared frame. Then we use transformation (scaling and translation) to bring the pictures to the wire frame coordinate, overlaying frames on pictures.

4.2. Feature Detection

We provide an automatic feature point extraction method with an interface for interactive correction when needed. There are methods to detect them just using special background information and predefined threshold [4][7] and then use an edge detection method and apply threshold again. In addition, image segmentation by clustering method is used [4]. However, it is not very reliable since the boundary between hair and face and chin lines are not easy to detect in many cases. Moreover color thresholding is too sensitive and depends on each individual's facial image and therefore requires many trials and experiments. We therefore use a structured snake, which has functionality to keep the structure of contours. It does not depend much on the background color and is more robust than simple thresholding method.

Structured Snake First developed by Kass et al. [8] the active contour method, also called snakes, is widely used to fit a contour on a given image. Above the conventional snake, we add three more functions. First, we interactively move a few points to the corresponding position, and anchor them to keep the structure of points when snakes are involved, which is also useful to get more reliable result when the edge we would like to detect is not very strong. We then use color blending for a special area, so that it can be attracted by a special color [5]. When the color is not very helpful and Sobel operator is not enough to get good edge detection, we use a

multiresolution technique [6] to obtain strong edges. It has two main operators, REDUCE with Gaussian operator and EXPAND. The subtraction produces an image resembling the result after Laplacian operators commonly used in the image processing. More times the REDUCE operator is applied stronger are the edges.

4.3. Modifying a Generic Model

3D Points from Two 2D Points We produce 3D points from two 2D points on frames with predefined relation between points from the front view and from the side view. Some points have x, y_f, y_s, z, so we take y_s , y_f or average of y_s and y_f for y coordinate (subscripts s and f mean side and front view). Some others have only x, y_f and others x, y_s. Using predefined relation from a typical face, we get 3D position (x, y, z).

Dirichlet Free-Form Deformations (DFFD) Distance-related functions have been employed by many researchers [4][7][9] to calculate the displacement of non-feature points related to feature points detected. We propose to use DFFD [11] since it has capacity for non-linear deformations as opposed to generally applied linear interpolation, which gives smooth result for the surface. We apply the DFFD on the points of the generic head. The displacement of non-feature points depends on the distance between control points. Since DFFD applies Voronoi and Delaunay triangulation, some points outside triangles of control points are not modified, the out-box of 27 points can be adjusted locally. Then the orginal shape of eyes and teeth are recovered since modifications may create unexpected deformation for them. Our system also provides a feedback modification of a head between feature detection and a resulted head.

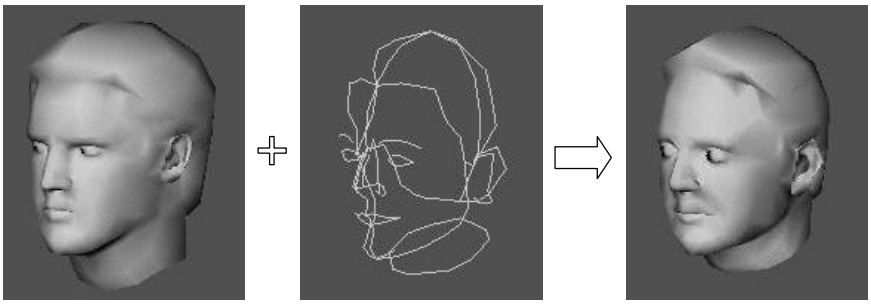

Fig. 3. A result of DFFD modification comparing with the original head.

4.4. Automatic Texture Mapping

To increase realism, we utilize texture mapping. Texture mapping needs a 2D texture image and coordinate for each point on a 3D head. Since our input is two pictures, texture image generation to combine them to one picture is needed.

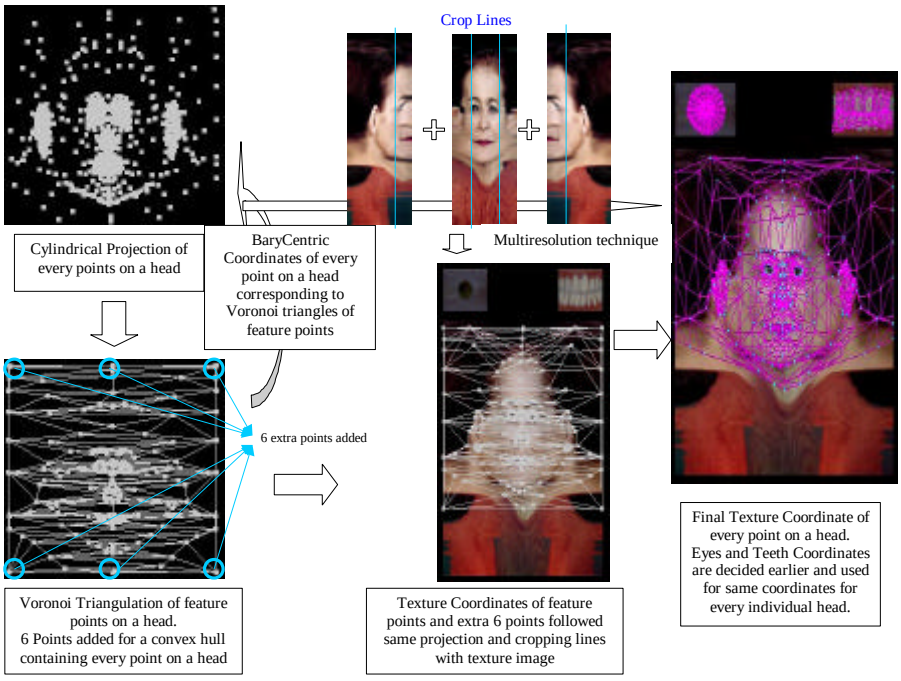

Fig. 4. Texture mapping process.

Texture Image Generation For smooth texture mapping, we assemble two images from the front and side views to be one. Boundaries of two pictures are detected using boundary color information or using detected feature points for face and hair boundaries. Since the hair shape is simple in a generic model, the boundary of a side view is modified automatically using information of back head profile feature points detected to have nice texture for back part of a neck. Then cylindrical projection of each image is processed. Two projected images are cropped at a certain position (we use eye extremes because eyes are important to keep at a high resolution), so that the range of combined image to make the final assembled image is 360^0. Finally, a multiresolution spline assembling method is used to produce one image for texture mapping preventing visible boundary for image mosaic.

Texture Fitting The main idea for the texture fitting is to map a 2D image on a 3D shape. Texture coordinates of feature points are calculated using detected position data and function applied for texture image generation. The problem for texture fitting is how to decide texture coordinates of all points on a surface of a head. We first apply a cylindrical projection of every point on a 3D head surface. Extra points

are added to make a convex hull containing all points, so that a coordinate of every point is located on an image. Then the Voronoi triangulation on control (feature) points and extra points are processed and the local Barycentric coordinates of every point with a surrounding Voronoi triangle are calculated. Finally the texture coordinates of each point on a 2D-texture image are obtained using texture coordinates of control points and extra points and correspond Barycentric coordinate.

4.5. Result

A final textured head is shown in Fig. 5 with input images, whose process from normalization to texture mapping takes a few minutes.

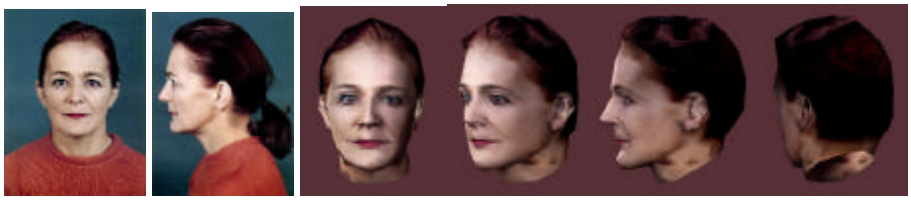

Fig. 5. A final reconstructed head with two input images in left side. The back of head has proper texture too.

5. Talking Head and Animation

In this section, we will describe the steps for animating a 3D face model according to an input speech text, that has been reconstructed as described in the previous section.

Firstly, the input text is being provided to a text-to-speech synthesis system. In this case, we are using the Festival Speech Synthesis System [21] that is being developed at the University of Edinburgh. It produces the audio stream that is subsequently played back in synchronization with the facial animation. The other output that is needed from this system is the temporized phoneme information, which is used for generating the facial animation.

In the facial animation system [22] that is used in our work, the basic motion parameter for animating the face is called Minimum Perceptible Action (MPA). Each MPA describes a corresponding set of visible features such as movement of eyebrows, jaw, or mouth occurring as a result of muscle contractions and pulls.

In order to produce facial animation based on the input speech text, we have defined a visual correlation to each phoneme, which are commonly known as visemes. In other words, each viseme is simply corresponding to a set of MPAs. In Fig. 6, it shows a cloned head model with a few visemes corresponding to the pronunciation of the words indicated.

With the temporized phoneme information, facial animation is then generated by concatenating the visemes. We have limited the set of phonemes to those used in the

Oxford English Dictionary. However, it is easy to extend the set of phonemes in order to generate facial animation for languages other than English, as this can be easily done by adding the corresponding visemes.

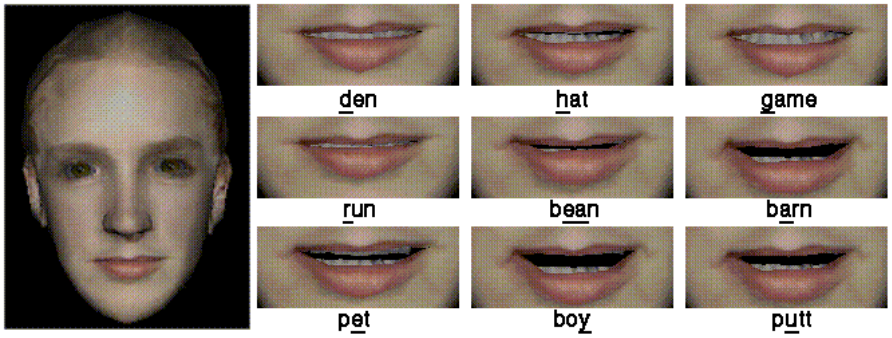

Fig. 6. A cloned head model with some examples of visemes for the indicated words. Our talking head system utilizes 44 visemes and some facial expression parameters.

The facial animation described so far is mainly concerned with the lip movement corresponding to the input speech text. However, this animation will appear artificial without the eyes blinking, and movement of the eyes and head. Therefore, we have also included random eyes blinking and movement of the eyes and head to add more realism into the generated animation. An analysis of the input speech text can also be done to infer the emotional state of the sentence. Then, the corresponding emotional facial expression can also be applied to the animation. For example, the eyebrows can be raised when the emotional state is surprise.

An example of cloned persons interacting in a shared virtual environment is shown in Fig. 7. Two kinds of generic bodies, female and male, are provided in advance. The generic bodies can be adjusted according to several ratios using Bodylib [28]. We connect individualized heads onto bodies by specifying a transformation matrix. Basically we need four types of data, namely the geometrical shape, animation structure, texture image and texture coordinates. In our case, every individualized head shares the same animation structure. Final rendering is then produced in real time. Our whole process from individualization to final talking head with body in a virtual world takes only few minutes.

Fig. 7. Cloned persons interacting in shared virtual environments.

6. Conclusion and Future Research

In this paper, we proposed an efficient way of creating the individual face model, in terms of time needed for the creation and transmitting the information to other users, and sending parameters for facial animation. The input of two orthogonal pictures can be easily obtained from any kind of conventional camera. The process of individualization consists of several steps including feature detection on 2D pictures, Dirichlet Free-Form Deformations for modification of a generic model and automatic texture mapping. The reconstructed heads connected to given bodies are used in a virtual environment. The result is then used immediately to produce an individualized talking head, which is created from given text and viseme database. This whole process can be achieved together with the MPEG-4 framework [27] to create a low bandwidth face-to-face communication between two or more users.

In the television broadcasting industry, the reconstruction method that we have described in this paper will allow actors to be cloned rapidly. Then, the television producer will be able to direct the virtual actors to say the lines and have the appropriate facial expressions in a virtual studio. This will obviously reduce the cost of production.

7. Acknowledgment

We are grateful to Chris Joslin for proof reading this document.

8. References

[1] Meet Geri: The New Face of Animation, Computer Graphics World, Volume 21, Number 2, February 1998.

[2] Sannier G., Magnenat Thalmann N., "A User-Friendly Texture-Fitting Methodology for Virtual Humans", Computer Graphics International'97, 1997.

[3] Exhibition On the 10th and 11th September 1996 at the Industrial Exhibition of the British Machine Vision Conference.

[4] Takaaki Akimoto, Yasuhito Suenaga, and Richard S. Wallace, Automatic Creation of 3D Facial Models, IEEE Computer Graphics & Applications, Sep., 1993.

[5] P. Beylot, P. Gingins, P. Kalra, N. Magnenat Thalmann, W. Maurel, D. Thalmann, and F. Fasel, 3D Interactive Topological Modeling using Visible Human Dataset. Computer Graphics Forum, 15(3):33-44, 1996.

[6] Peter J. Burt and Edward H. Andelson, A Multiresolution Spline With Application to Image Mosaics, ACM Transactions on Graphics, 2(4):217-236, Oct., 1983.

[7] Horace H.S. Ip, Lijin Yin, Constructing a 3D individual head model from two orthogonal views. The Visual Computer, Springer-Verlag, 12:254-266, 1996.
[8] M. Kass, A. Witkin, and D. Terzopoulos, Snakes: Active Contour Models, International Journal of Computer Vision, pp. 321-331, 1988.
[9] Tsuneya Kurihara and Kiyoshi Arai, A Transformation Method for Modeling and Animation of the Human Face from Photographs, Computer Animation, Springer-Verlag Tokyo, pp. 45-58, 1991.
[10] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters, Realistic Modeling for Facial Animation, In Computer Graphics (Proc. SIGGRAPH), pp. 55-62, 1996.
[11] L. Moccozet, N. Magnenat-Thalmann, Dirichlet Free-Form Deformations and their Application to Hand Simulation, Proc. Computer Animation, IEEE Computer Society, pp. 93-102, 1997.
[12] N. Magnenat-Thalmann, D. Thalmann, The direction of Synthetic Actors in the film Rendez-vous à Montrèal, IEEE Computer Graphics and Applications, 7(12):9-19, 1987.
[13] Marc Proesmans, Luc Van Gool. Reading between the lines - a method for extracting dynamic 3D with texture. In Proceedings of VRST, pp. 95-102, 1997.
[14] http://www.viewpoint/com/freestuff/cyberscan
[15] LeBlanc. A., Kalra, P., Magnenat-Thalmann, N. and Thalmann, D. Sculpting with the 'Ball & Mouse' Metaphor, Proc. Graphics Interface '91. Calgary, Canada, pp. 152-9, 1991.
[16] Lee W. S., Kalra P., Magenat Thalmann N, Model Based Face Reconstruction for Animation, Proc. Multimedia Modeling (MMM) '97, Singapore, pp. 323-338, 1997.
[17] A. Pearce, B. Wyvill, G. Wyvill, D. Hill (1986) Speech and Expression: A Computer Solution to Face Animation, Proc Graphics Interface '86, Vision Interface '86, pp. 136-140.
[18] M. M. Cohen, D. W. Massaro (1993) Modeling Coarticulation in Synthetic Visual Speech, eds N. M. Thalmann, D. Thalmann, Models and Techniques in Computer Animation, Springer-Verlag, Tokyo, 1993, pp. 139-156.
[19] T. Ezzat, T. Poggio (1998) MikeTalk: A Talking Facial Display Based on Morphing Visemes, Submitted to IEEE Computer Animation '98.
[20] E. Cossato, H.P. Graf (1998) Sample-Based Synthesis of Photo-Realistic Talking Heads, Submitted to IEEE Computer Animation '98.
[21] A. Black, P. Taylor (1997) Festival Speech Synthesis System: system documentation (1.1.1), Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh.
[22] P. Kalra (1993) An Interactive Multimodal Facial Animation System, PH.D. Thesis, Ecole Polytechnique Federale de Lausanne.
[23] Barrus J. W., Waters R. C., Anderson D. B., "Locales and Beacons: Efficient and Precise Support For Large Multi-User Virtual Environments", Proceedings of IEEE VRAIS, 1996.
[24] Carlsson C., Hagsand O., "DIVE - a Multi-User Virtual Reality System", Proceedings of IEEE VRAIS '93, Seattle, Washington, 1993.
[25] Macedonia M.R., Zyda M.J., Pratt D.R., Barham P.T., Zestwitz, "NPSNET: A Network Software Architecture for Large-Scale Virtual Environments", Presence: Teleoperators and Virtual Environments, Vol. 3, No. 4, 1994.
[26] I. Pandzic, T. Capin, E. Lee, N. Magnenat-Thalmann, D. Thalmann (1997) A Flexible Architecture for Virtual Humans in Networked Collaborative Virtual Environments, Proc EuroGraphics '97, Computer Graphics Forum, Vol 16, No 3, pp C177-C188.
[27] I. Pandzic, T. Capin, N. Magnenat-Thalmann, D. Thalmann (1997) MPEG-4 for Networked Collaborative Virtual Environments, Proc VSMM '97, pp.19-25.

[28] J. Shen, D. Thalmann, Interactive Shape Design Using Metaballs and Splines, Proc. Implicit Surfaces `95 , Grenoble, pp.187-196.