Face Analysis and Synthesis

For Duplication Expression and Impression

Shigeo Morishima

Recently, research in creating friendly human interfaces has flourished. Such interfaces provide smooth communication between a computer and a human. One style is to have a virtual human, or avatar [3], [4] appear on the computer terminal. Such an entity should be able to understand and express not only linguistic information but also nonverbal information. This is similar to human-to-human communication with a face-to-face style [5].

A very important factor in making an avatar look realistic or alive depends on how precisely an avatar can duplicate a real human facial expression and impression on a face precisely. Especially in the case of communication applications using avatars, real-time processing with low delay is essential.

Our final goal is to generate a virtual space close to the real communication environment between network users or between humans and machines. There should be an avatar in cyberspace that projects the features of each user with a realistic texture-mapped face to generate facial expression and action controlled by a multimodal input signal. Users can also get a view in cyberspace through the avatar's eyes, so they can communicate with each other by gaze crossing.

In the first section, the face fitting tool from multiview camera images is introduced to make a realistic three-dimensional (3-D) face model with texture and geometry very close to the original. This fitting tool is a GUI-based system using easy mouse operation to pick up each feature point on a face contour and the face parts, which can enable easy construction of a 3-D personal face model.

When an avatar is speaking, the voice signal is essential in determining the mouth shape feature. Therefore, a real-time mouth shape control mechanism is proposed by using a neural network to convert speech parameters to lip shape parameters. This neural network can realize an interpolation between specific mouth shapes given as learning data [1], [2]. The emotional factor can sometimes be captured by speech parameters. This media conversion mechanism is described in the second section.

For dynamic modeling of facial expression, a muscle structure constraint is introduced for making a facial expression naturally with few parameters. We also tried to obtain muscle parameters automatically from a local motion vector on the face calculated by the optical flow in a video sequence. Accordingly, 3-D facial expression transition is duplicated from actual people by analysis of video images captured by a two-dimensional (2-D) camera

1053-5888/01/$10.00©2001IEEE

without markers on the face. These efforts are described in the third section. Furthermore, we tried to control this artificial muscle directly by EMG signal processing.

To get greater reality for the head, a modeling method of hair is introduced, and the dynamics of hair in a wind stream can be achieved at low calculation cost. This is explained in the last section. By using these various kinds of multimodal signal sources, a very natural face image and impression can be duplicated on an avatar's face.

Face Modeling

To generate a realistic avatar's face, a generic face model is manually adjusted to the user's face image. To produce a personal 3-D face model, both the user's frontal face image and profile image are necessary at the least. The generic face model has all of the control rules for facial expressions defined by the facial action coding system (FACS) parameter as a 3-D movement of grid points to modify geometry.

Figure 1 shows a personal model fitted to a front image and profile image using our original GUI-based face fitting tool. In this system, corresponding control points are manually moved to a reasonable position by mouse operation.

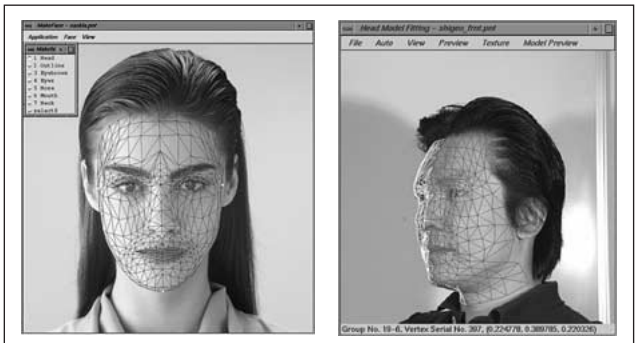
The synthesized face emerges by mapping of the blended texture generated by the user's frontal image and profile image onto the modified personal face model. However, sometimes self-occlusion occurs and then we cannot capture texture from only the front and profile face images in the occluded part of the face model. To construct a 3-D model more accurately, we also introduce a multiview face image fitting tool.

Figure 2(b) shows the fitting result to the captured image from the bottom angle to compensate for the texture behind the jaw. The rotation angle of the face model can be controlled in the GUI preview window to achieve the best fitting to the face image captured from any arbitrary angle. Figure 3 shows the full face texture projected onto a cylindrical coordinate. This texture is projected onto a 3-D personal model adjusted to fit to multiview images to synthesize a face image. Figure 4 shows examples of reconstructed faces. Figure 4(a) uses nine view images and (b) uses only frontal and profile views. Much better image quality can be achieved by the multiview fitting process.

Voice Driven Talking Avatar

The relation between speech and mouth shape is studied widely [1], [6], [7], [9]. To realize lip synchronization, the spoken voice is analyzed and converted into mouth shape parameters by a neural network on a frame-by-frame basis.

Multiuser communication systems in cyberspace are constructed based on a server-client system. In our system, only a few parameters and the voice signal are transmitted through the network. The avatar's face is synthesized by these parameters locally in the client system.



(a) (b)

▲ *1. Fitting front and profile model. (a) Front fitting, (b) profile fitting.*

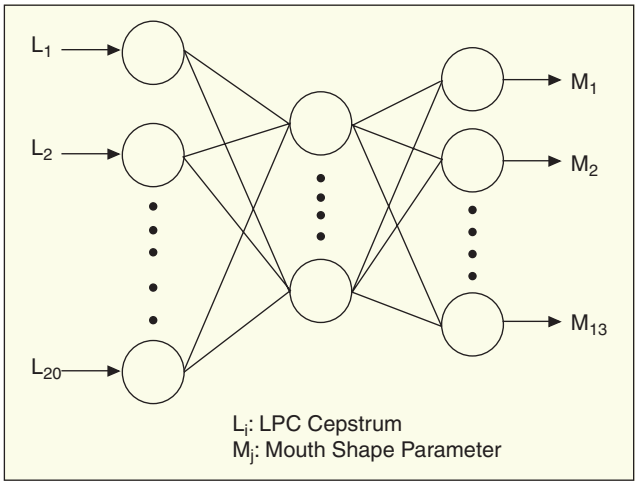
(a) (b)

▲ *2. Fitting front and profile model. (a) From bottom, (b) from diagonal.*

▲ *3. Cylindrical face texture.*

(a) (b)

▲ *4. Reconsructed face. (a) Multiview, (b) two view.*



▲ *5. Network for parameter conversion.*

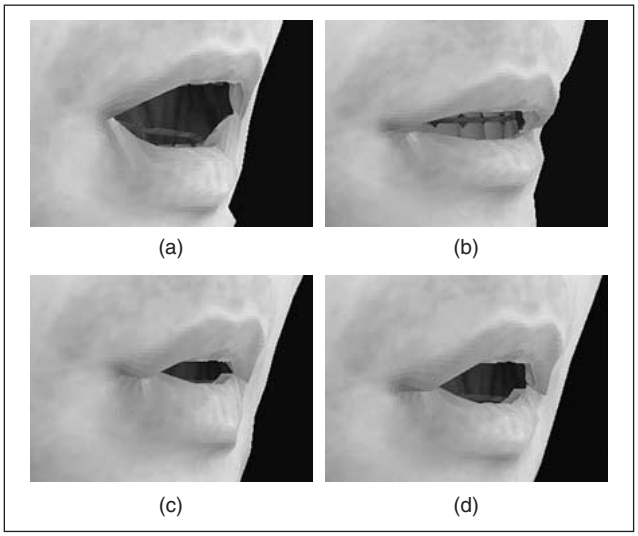
▲ *6. Control panel for mouth shape editor.*

Parameter Conversion

In the server system, the voice from each client is phonetically analyzed and converted to mouth shape and expression parameters.

LPC Cepstrum parameters are converted into mouth shape parameters by a neural network trained by vowel features. Figure 5 shows the neural network structure used for parameter conversion. Twenty dimensional Cepstrum parameters are calculated every 32 ms with 32 ms frame length and given into input units of neural network. The output layer has 17 units corresponding to the mouth shape parameters.

In the client system, the on-line captured voice of each user is digitized to 16 kHz and 16 bits and transmitted to the server system frame-by-frame through a network in the communication system. Then the mouth shape of each avatar in cyberspace is synthesized by this mouth shape parameter received at each client.



▲ *7. Typical mouth shapes. (a) Shape for /a/, (b) shape for /i/, (c) shape for /u/, (d) shape for /o/.*

The training data for back-propagation are composed of data pairs with voice and mouth shape for vowels, nasals, and transient phoneme captured from video sequence.

Hidden Markov model–based mouth shape estimation from voice is proposed [9]. However, their system cannot realize no-delay real-time processing.

The emotion condition is also determined by LPC Cepstrum, power, pitch frequency, and utterance speed by using discriminant analysis into anger, happiness, or sadness. Each basic emotion has a specific facial expression parameter described by FACS [8].

Mouth Shape Editor

Mouth shape can be easily edited by our mouth shape editor (see Fig. 6). We can change each mouth parameter to set a specific mouth shape on the preview window. Typical vowel mouth shapes are shown in Fig. 7. Our special mouth model has polygons for inside the mouth and the teeth. A tongue model is now under construction. For parameter conversion from LPC Cepstrum to mouth shape, only the mouth shapes for five vowels and nasals are defined as the training set. We have defined all of the mouth shapes for Japanese phonemes and English phonemes by using this mouth shape editor. Figure 8 shows a synthesized avatar face speaking phoneme /a/.

Multiple User Communication System

Each user can walk through and fly through cyberspace by mouse control, and the current locations of all users are always monitored by the server system. The avatar image is generated in a local client machine by the location information from the server system. The emotion condition can always be determined by voice, but sometimes a user

gives his or her avatar a specific emotion condition by pushing a function key. This process works as first priority. For example, push anger and the angry face of your avatar emerges.

The location information of each avatar, mouth shape parameters, and emotion parameters are transmitted every 1/30 seconds to the client systems. The distance between any two users is calculated by the avatar location information, and the voice from every user except himself or herself is mixed and amplified with gain according to the distance. Therefore, the voice from the nearest avatar is very loud and that from far away is silent.

Based on facial expression parameters and mouth shape parameters, an avatar's face is synthesized frame-by-frame. Also, the avatar's body is located in cyberspace according to the location information. There are two modes for display: the view from avatar's own eyes for eye contact and the view from the sky to search for other users in cyberspace. These views can be chosen by a menu in the window. Figure 9 shows a communication system in cyberspace with an avatar.

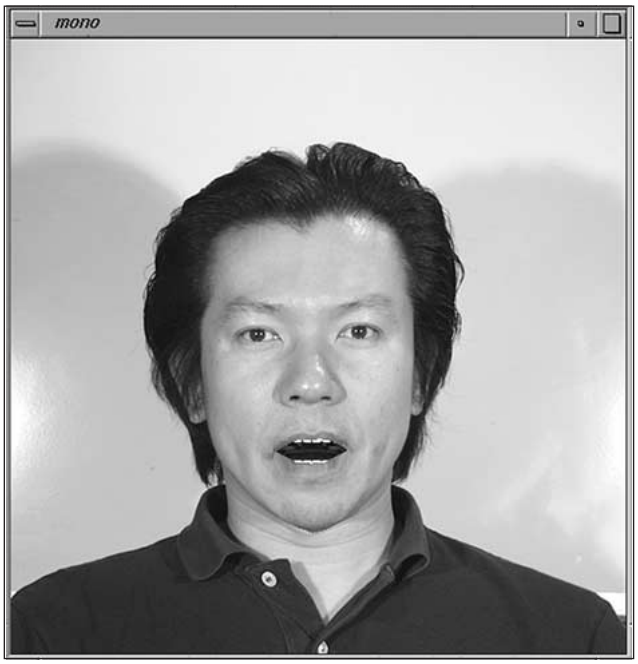
User Adaptation

When a new user enters the system, his or her face model and voice model have to be registered before operation. In the case of voice, ideally new learning for the neural network should be performed. However, it takes a very long time to get convergence of back propagation. To simplify the face model construction and voice learning, a GUI tool for speaker adaptation has been prepared. To register the face of a new user, a generic 3-D face model is modified to fit on the input face image. Expression control rules are defined in the generic model, so every user's face can be equally modified to generate basic expressions using the FACS-based expression control mechanism.

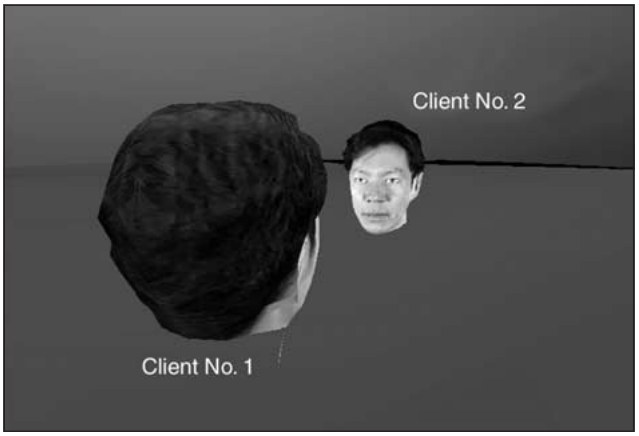
For voice adaptation, 75 persons voice data including five vowels are precaptured, and a database for weights of the neural network and voice parameters is constructed. Accordingly, speaker adaptation is performed by choosing the optimum weight from the database. Training of the neural network for the data of each of the every 75 persons is already finished before operation. When a new nonregistered speaker comes in, he or she has to speak five vowels into a microphone. LPC Cepstrum is calculated for each of the five vowels, and this is entered into the neural network. Then the mouth shape is calculated by the selected weight, and the error between the true mouth shape and the generated mouth shape is evaluated. This process is applied to the entire database one-by-one and the optimum weight is selected when the minimum error is detected.

System Features

A natural communication environment between multiple users in cyberspace by transmission of natural voice and real-time synthesis of an avatar's facial expression is realized. The current system works with three users in an intranetwork environment at the speed of 16 frames/s on an SGI Max IMPACT. An emotion model [12] is intro-

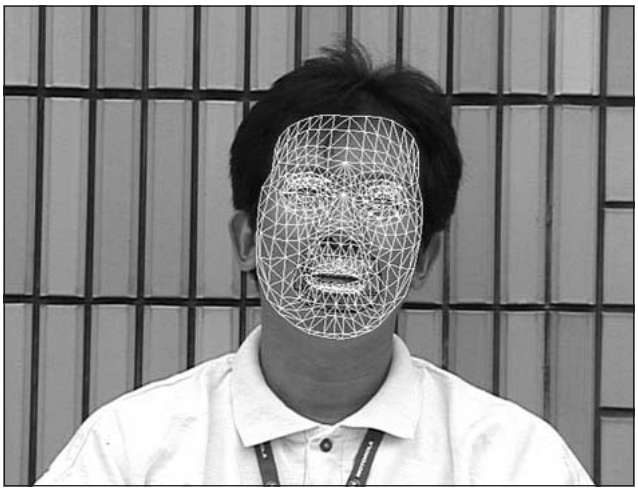


▲ *8. Synthesized face speaking /a/.*



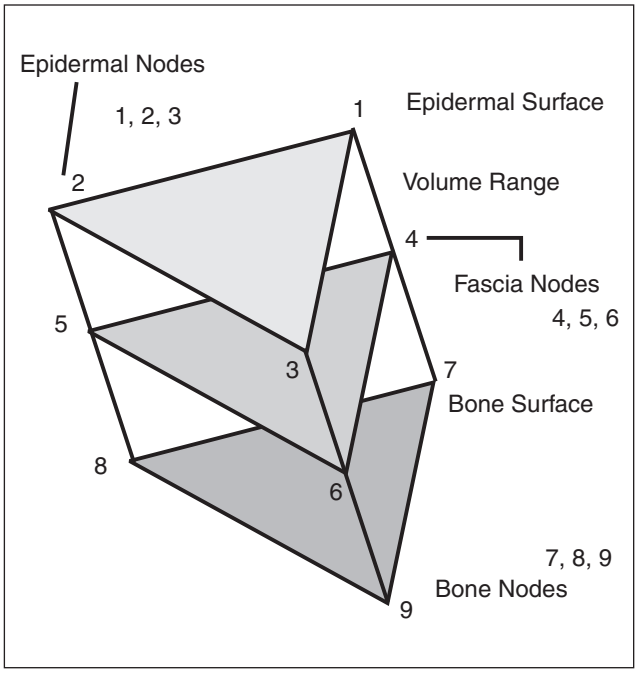
▲ *9. Communication system in cyberspace.*

▲ *10. Original video clip.*



▲ *11. Fitting result of face model.*

▲ *12. Reconstructed face by another face.*



▲ *13. Layered tissue element.*

duced to improve the communication environment. Gaze tracking and mouth closing point detection can also be realized by a pan-tilt-zoom controlled camera.

Entertainment Application

When people watch movies, they sometimes overlap their own figure with the actor's image. An interactive movie system we constructed is an image creating system in which the user can control the facial expression and lip motion of his or her face image inserted into a movie scene. The user submits a voice sample by microphone and pushes keys that determine expression and special effect. His or her own video program can be generated in real time.

At first, once a frontal face image of a visitor is captured by camera, a 3-D generic wireframe model is fitted onto the user's face image to generate a personal 3-D surface model. A facial expression is synthesized by controlling the grid point of the face model and texture mapping. For speaker adaptation, the visitor has to speak five vowels to choose an optimum weight from the database.

In the interactive process, a famous movie scene is going on and the facial region of an actor or actress is replaced with the visitor's face. Facial expression and lip shape are also controlled synchronously by the captured voice. An active camera tracks the visitor's face, and the facial expression is controlled by a CV-based face image analysis. Figure 10 shows the original video clip, and Fig. 11 shows the result of fitting a face model into this scene. Figure 12 shows a user's face inserted into actor's face after color correction. Any expression can be appended, and any scenario can be given by the user independent of the original story in this interactive movie system.

Muscle Constraint Face Model

Muscle-based face image synthesis is one of the most realistic approaches to the realization of a lifelike agent in computers. A facial muscle model is composed of facial tissue elements and simulated muscles. In this model, forces are calculated to effect a facial tissue element by contraction of each muscle string, so the combination of each muscle's contracting force decides a specific facial expression. This muscle parameter is determined on a trial-and-error basis by comparing a sample photograph and a generated image using our muscle editor to generate a specific face image. In this section, we propose the strategy of automatic estimation of facial muscle parameters from 2-D optical flow by using a neural network. This corresponds to the 3-D facial motion capturing from 2-D camera images under a physics-based condition without markers.

We introduce a multilayered back-propagation network for the estimation of the muscle parameter. A neural network is used to classify 3-D objects from a 2-D view [15], recognize expressions [17], and model facial emo-

tion condition [12], and a self-organizing mechanism is well studied [16].

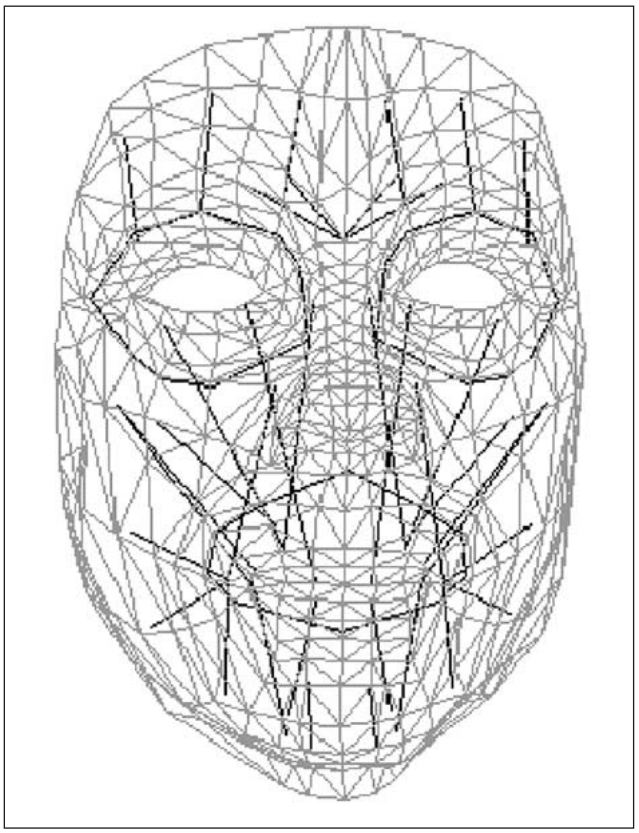
The facial expression is then resynthesized from the estimated muscle parameter to evaluate how well the impression of an original expression can be recovered. We also tried to generate animation by using the captured data from an image sequence. As a result, we can obtain and synthesize an image sequence that gives an impression very close to the original video.

Layered Dynamic Tissue Model

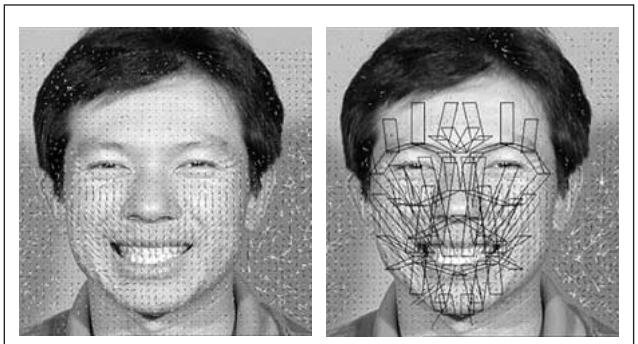
The human skull is covered by a deformable tissue that has five distinct layers. In accordance with the structure of real skin, we employ a synthetic tissue model constructed from the elements illustrated in Fig. 13, consisting of nodes interconnected by deformable springs (the lines in the figure). The facial tissue model is implemented as a collection of node and spring data structures. The node data structure includes variables to represent the nodal mass, position, velocity, acceleration, and net force. Newton's laws of motion govern the response of the tissue model to force [13].

Facial Muscle Model

Figure 14 shows our simulated muscles. The black line indicates the location of each facial muscle in a layered tissue face model. Normally, muscles are located between a bone node and a fascia node. But the orbicularis oculi has an irregular style, whereby it is attached between fascia nodes in a ring configuration; it has eight linear muscles that approximate a ring muscle. Contraction of the ring muscle makes the eyes thin. The muscles around the mouth are very important for the production of speaking scenes. Most Japanese speaking scenes are composed of vowels, so we mainly focused on the production of vowel mouth shapes as a first step and relocated the muscles around the mouth [14]. As a result, the final facial muscle model has 14 muscle springs in the forehead area and 27 muscle springs in the mouth area.



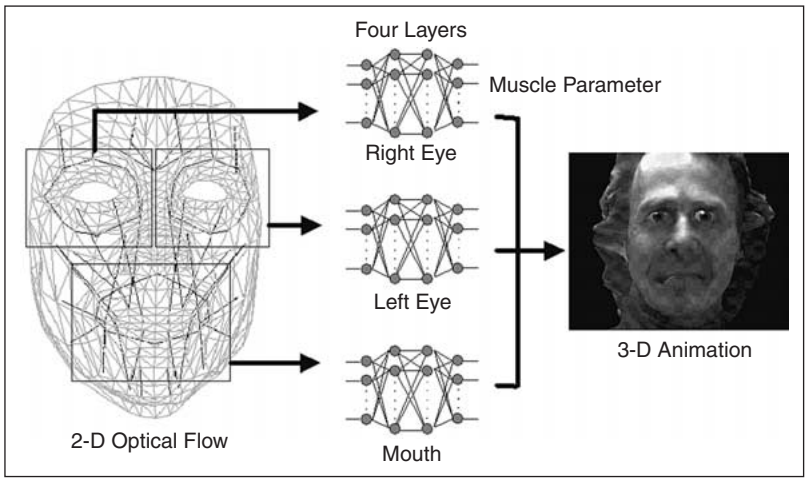
▲ *14. Facial muscle model.*



▲ *15. Motion feature vector on face. (a) Optical flow, (b) feature window.*

Motion Capture on Face

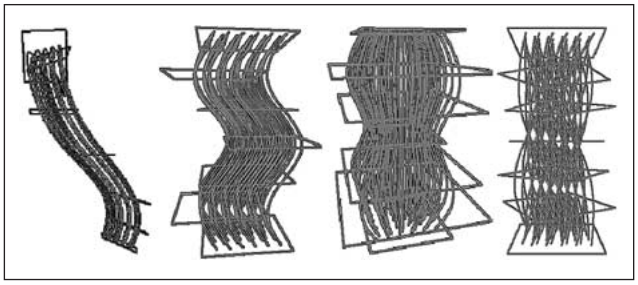
Face tracking [10] and face expression recognition [11] based on optical flow have already been proposed. Their purpose is to find a similar cartoon face by some matching processes; however optical flow is very sensitive to noise. Our goal is to estimate precisely the muscle parameters and resynthesize a texture-mapped face expression using muscle model with the original impression. A personal face model is constructed by fitting the generic control model to personal range data. An optical flow vector is calculated in a video sequence, and we accumulate the motion in the sequence from neutral to each expression. Then motion



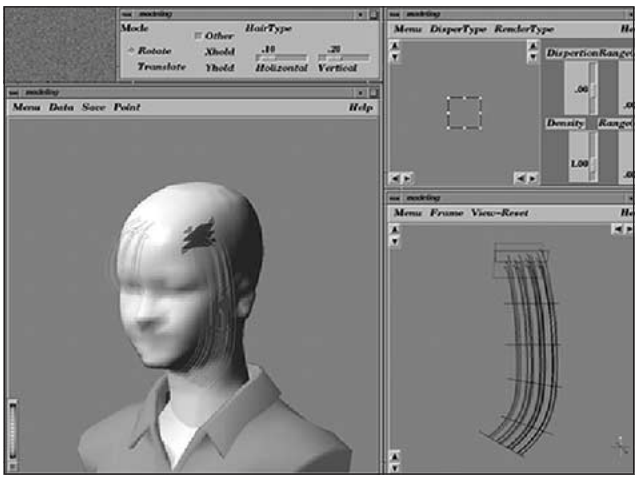
▲ *16. Conversion from motion to animation.*

▲ *17. Face expression regeneration. (a) Original, (b) regenerated.*

▲ *18. Mouth shape control by EMG.*



▲ *19. Style variation with tuft model.*



▲ *20. GUI tool for hair designing.*

vectors are averaged in each window as shown in Fig. 15(b). This window location is determined automatically in each video frame. Our system is not limited to recognize only six basic expressions [17], [18].

A layered neural network finds a mapping from the motion vector to the muscle parameter. A four-layer structure is chosen to model effectively the nonlinear performance. The first layer is the input layer, which corresponds to a 2-D motion vector. The second and third layers are hidden layers. Units of the second layer have a linear function, and those of the third layer have a sigmoid function. The fourth layer is the output layer, corresponding to the muscle parameter, and it has linear units. Linear functions in the input and output layers are introduced to maintain the range of input and output values.

A simpler neural network structure can help to speed up the convergence process in learning and reduce the calculation cost in parameter mapping, so the face area is divided into three sub-areas as shown in Fig. 16. They are the mouth area, left-eye area, and right-eye area, each giving independent skin motion. Three independent neural networks are prepared for these three areas.

Learning patterns are composed of basic expressions and the combination of their muscle contraction forces.

A motion vector is given to a neural network from each video frame, and then a facial animation is generated from the output muscle parameter sequence. This is a test of the effect of interpolation on our parameter conversion method based on the generalization of the neural network.

Figure 17 shows the result of expression regeneration for surprise from an original video sequence.

Muscle Control by EMG

EMG data is the voltage wave captured by a needle inserted directly into each muscle so that the feature of wave expresses the contraction situation of a muscle. In particular, seven di-ball wires were inserted into the muscles around the mouth to make a model of mouth shape control.

The conversion from EMG data to muscle parameters is as follows. First, the average power of an EMG wave is calculated every 1/30 seconds. Sampling frequency of EMG is 2.5 kHz. The maximum power is equal to the maximum muscle contraction strength by normalization. Then the converted contraction strength of each muscle is given every 1/30 seconds and the facial animation scene is generated. However, the jaw is controlled directly by the marker position located on the subject's jaw independent of EMG data.

EMG data is captured from a subject speaking five vowels. Seven kinds of waves can be captured and seven contractions of muscles are determined. Then the face model is modified and the mouth shape for each vowel is synthesized. The result is shown in Fig. 18. In Fig. 18, the camera captured image and synthesized image can be compared, and the impression is very close to each other. Next, EMG time sequence is converted into muscle pa-

rameters every 1/30 seconds, and facial animation is generated. Each sample is composed of a three-second sentence. Good animation quality for speaking the sentence is achieved by picking up impulses from the EMG signal and activating each appropriate muscle. This data offer the precise transition feature of each muscle contraction.

Dynamic Hair Modeling

The naturalness of hair is a very important factor in a visual impression, but it is treated with a very simple model or as a texture. Because real hair has huge pieces and complex features, it's one of the hardest targets to model by computer graphics [19], [20]. In this section, a new hair modeling system is presented [21]. This system helps the designer to create any arbitrary hair style by computer graphics with easy operation by using an editor of tufts. Each piece of hair has an independent model, and dynamic motion can be simulated easily by solving a motion equation. However, each piece has only a few segments and the feature is modeled by a 3-D B-spline curve, so the calculation cost is not so huge.

Modeling Hair

It is necessary to create a model of each piece of hair to generate a natural dynamic motion when wind is flowing. However, the hair feature is very complex, and a real head has more than 100,000 pieces. Therefore, a polygon model for hair needs a huge amount of memory for the entire head to be stored. In this article, each piece has only a few segments to control and is expressed with a 3-D B-spline curve.

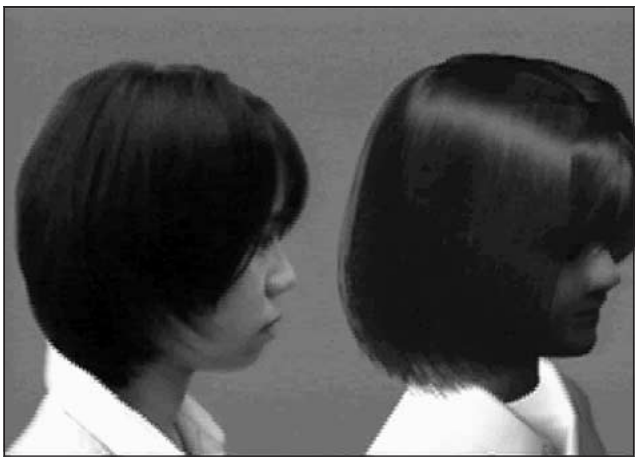
Tuft Model

In our designing system, a head is composed of about 3,400 polygons, and a 3-D B-spline curve comes out from each polygon to create hair style. The hair generation area is composed of about 1,300 polygons, and one specific hair feature is generated and copied for each polygon. To manipulate hair to get a specific hairstyle, some of the hair features are simultaneously treated as one tuft. Each tuft is composed of more than seven squares in which hair segments pass through. This square is the gathering control point of the hair segments, so manipulation of this square can realize any hairstyle by rotation, shift, and modification. This tuft model is illustrated in Fig. 19.

Hair Style Designing System

There are three processes to generate hair style by GUI, and each process is easily operated by mouse control. The first process is to decide the initial region on the head from which the tuft comes out. The second process is to manipulate a hair tuft by modifying the squares by rotation, shift, and expansion. The third process is matching a tuft onto the surface of the head. After these processes, each polygon gets one hair feature, and in total 1,300 features are available. Finally, 20,000 to 30,000 pieces of hair for static images, or 100,000 to 200,000 pieces for animation, are generated by copying the feature in each polygon. The GUI tool for hair designing is shown in Fig. 20.

▲ *21. Example of hair style.*



▲ *22. Copying real hair style.*

Rendering

Each hairpiece is modeled with a 3-D B-Spline curve and also modeled as a very thin circular pipe. Therefore, the surface normal can be determined in each pixel to introduce the Lambert model and Phong model. The typical hairstyle Dole Bob is shown in Fig. 21. Figure 22 shows an example modeling of a real hairstyle. The impression of the generated hair is very natural.

This hair piece is modeled by stick segment and can be controlled dynamically by solving motion equation. Collision detection between hair and head is also considered to generate natural dynamic hair animation.

Conclusion

To generate a realistic avatar's face for face-to-face communication, a multimodal signal is introduced to duplicate original facial expressions. Voice is used to realize lip synchronization and expression control. Video captured images are used to regenerate an original facial expression under a facial muscle constraint. EMG data is used to control directly an artificial muscle. Finally, a hair modeling

method is proposed to make an avatar appear more natural and believable.

Shigeo Morishima received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Tokyo, Tokyo, Japan, in 1982, 1984, and 1987, respectively. Currently, he is a Professor at Seikei University, Tokyo, Japan. His research interests include physics-based modeling of face and body, facial expression recognition and synthesis, human computer interaction, and future interactive entertainment using speech and image processing. He was a Visiting Researcher at the University of Toronto from 1994 to 1995. He has been engaged in the multimedia ambiance communication TAO research project as a Subleader since 1997. He has been also a temporary lecturer at Meiji University, Japan, since 2000 and a visiting researcher at the ATR Spoken Language Translation Research Laboratories since 2001. He is an editor of the *Transaction of the Institute of Electronics, Information and Communication Engineers* (IEICE), Japan. He received the 1992 Achievement Award from IEICE.

References

[1] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man–machine interface," *IEEE J. Select. Areas Commun*., vol. 9, no. 4, pp. 594-600, 1991.

[2] S. Morishima, "Virtual face-to-face communication driven by voice through network," in *Workshop on Perceptual User Interfaces*, 1997, pp. 85-86.

[3] N.I. Baddler, C.B. Phillip, and B.L. Webber, *Simulating Humans, Computer Graphics Animation and Control*. Oxford, U.K.: Oxford Univ. Press, 1993.

[4] N.M. Thalmann and P. Kalra, "The simulation of a virtual TV presenter," in *Computer Graphics and Applications*. Singapore: World Scientific, 1995, pp. 9-21.

[5] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents," in *Proc. SIGGRAPH'94*, pp. 413-420.

[6] E. Vatikitotis-Bateson, K.G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Proc. ICSLP-96*, pp. 1485-1488.

[7] K. Waters and J. Frisbie, "A coordinate muscle model for speech animation," in *Proc. Graphics Interface'95*, pp. 163-170.

[8] P. Ekman and W.V. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists, 1978.

[9] S. Nakamura, "HMM-based transmodal mapping from audio speech to talking faces," in *IEEE Int. Workshop on Neural Networks for Signal Processing*, 2000, pp. 33-42.

[10] I. Essa, T. Darrell, and A. Pentland, "Tracking facial motion," in *Proc. Workshop on Motion and Nonrigid and Articulated Objects*, 1994, pp. 36-42.

[11] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans*., vol. E-74, no. 10, pp. 3474-3483, 1991.

[12] S. Morishima, *Modeling of Facial Expression and Emotion for Human Communication System Displays*. Amsterdam, The Netherlands: Elsevier, 1996, pp. 15-25.

[13] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proc. SIGGRAPH'95*, pp. 55-62.

[14] H. Sera, S. Morishima, and D. Terzopoulos, "Physics-based muscle model for mouth shape control," in *Proc. Robot and Human Communication*, 1996, pp. 207-212.

[15] S. Suzuki and H. Ando, "Unsupervised classification of 3D objects from 2D view," in *Advances in Neural Information Processing System 7*. Cambridge, MA: MIT Press, 1995, pp. 949-956.

[16] Y. Katayama and K. Ohyama, "Some characteristics of self-organizing back propagation neural network," in *Annu. Convention Rec. IEICE,* 1989, pp. SD-1-14.

[17] H. Kobayashi and F. Hara, "Analysis of the neural network recognition characteristics of 6 basic expressions," in *Proc. IEEE ROMAN'94*, pp. 222-227.

[18] M. Rosenblum, Y. Yacoob, and L. Davis, "Human emotion recognition from motion using a radial basis function network architecture," in *Proc. Workshop on Motion and Non-Rigid and Articulated Objects*, 1994, pp. 43-49.

[19] N. Thalman, T. Kurihara, and D. Thalman, "An integrated system for modeling, animating and rendering hair," in *Eurographics'93*, pp. 211-221.

[20] L.-H. Chen, S. Saeyor, H. Dohi, and M. Ishizuka, "A system of 3D hair style synthesis based on wisp model," *Visual Computer*, vol. 15, no. 4, pp. 159-170, 1998.

[21] K. Kishi and S. Morishima, "Modeling and animating human hair," *Trans. IEICE*, vol. J83-D-2, no. 12, pp. 2716-2724, 2000.