Creating a Multiuser 3-D Virtual Environment

Wing Ho Leung and Tsuhan Chen

For many years online text chat software such as ICQ [1] has been used as a means of multiuser interaction. As higher bandwidth is available, multipoint videoconferencing systems such as NetMeeting [2] or CUseeMe [3] that allow people from different geographical locations to see and talk to each other are becoming popular. However, these videoconferencing systems lack the sense of immersion because each participant sees other participants in separate windows, and it is often difficult to tell who is talking to whom. Based on user study, Argyle and Dean [4] suggested that during face-to-face communication people vary physical proximity, eye contact, and other behaviors to optimize an overall level of intimacy. In addition, Tang and Isaacs [5] found that gaze awareness is important because participants can use the visual cue to determine when another participant is paying attention to them. These studies suggest that a more immersive way of multiuser communication is to connect the users in a three-dimensional (3-D) virtual environment so that users feel that they are communicating face to face. Moving from a text-based chat room environment to a 3-D environment creates new challenges in several areas such as computer animation, signal processing, and computer vision. In this article, we introduce immersive interactive technologies used in multiuser 3-D virtual environments. We also survey existing systems, some of which are detailed in other articles in this special issue.

Elements of a Multiuser 3-D Virtual Environment

A virtual environment is a shared space among the users and should be rendered according to the users' perspectives in order to provide users with the sense of immersion. An avatar is often used to represent a user. It can be animated to create body movements, facial expressions, hand gestures, and lip-synchronization. User interfaces such as the visual interface, auditory interface, primary input interface, and tracking interface are practical elements of a multiuser 3-D virtual environment. Other kinds of user interfaces exist, such as the haptic interface or olfactory interface that also enhance immersive interaction among users. However, their underlying technologies are still immature and we believe that more research effort is needed before we can really make use of them in the virtual environment for immersive multiuser interaction. On the other hand, sharing and transmitting multimedia content such as the streaming of video and audio, shared whiteboard, and the sharing and manipulation of 3-D objects are also important components in a multiuser virtual environment.

1053-5888/01/$10.00©2001IEEE

Streaming technology is important for real-time interaction where a user may want to show a videoclip and discuss it with other users at the same time in the virtual room.

Environment Rendition

A realistic rendering of the shared environment provides users with the sense of immersion, and rendering the virtual view in 3-D maintains the spatial relationship between the users and the objects in the environment. There are two approaches to render the environment: computer graphics and image-based rendering. Computer graphics is the traditional approach by modeling the environment with primitive geometric elements such as lines and triangles. On the other hand, with image-based rendering, video or still images of a real-world scene are captured, and the virtual environment is then constructed by applying image-processing techniques to these acquired images.

As an example of the image-based rendering approach, Ichikawa et al. describe in this special issue their work on multimedia ambiance communication, which is a video-based approach used to establish a photo-realistic 3-D image space for rendering the environment [37]. They classify the scene rendering into long-range, mid-range, and short-range views as a layered structure governed by the laws of perceptivity as defined for painting.

Avatar Rendition

In a virtual environment, a user is often represented by an avatar, which is an animated human-like character. The word "avatar" originates from the Hindu religion. It means an incarnation of a deity; hence an embodiment or manifestation of an idea or greater reality. To provide a vivid representation of the users, the avatar's face can be animated to show facial expressions, and the avatar's body can also be animated to perform some action such as gesturing, walking, or dancing.

Body Animation

Body animation has attracted much attention in recent years. The Human Animation Working Group (H-Anim) proposed a standard way of representing humanoids in VRML97 [6]. In terms of computer graphics animation, the work done in MiraLab [7] demonstrates some simple body animations such as walking [8] and grasping [9]. Many body motions result from simultaneous movements of several joints with certain constraints; thus, it is not easy to reproduce realistic complex body animation. Often motion sensors are attached to the user in order to determine the user's posture. For example, during the production of the film *Titanic*, the passengers on the ship are animated using this method.

Two articles in this special issue present techniques for body animation. For realistic rendering of the whole avatar, Tamagawa et al. describe their 2.5-D video avatar used for immersive communication [38]. The surface model of the user who faced in the direction of the camera is formed from the depth information obtained from the stereo cameras. On the other hand, Morishima proposes a realistic human model that even includes a hair-modeling method to make an avatar appear more natural [39].

Face Modeling

It is desirable to construct a realistic 3-D face model customized for a specific user in order to provide a vivid representation of the user in the virtual environment. Laser scanners can be used to obtain the range data of a person's face, and a 3-D face model can be obtained by the triangulation of the range data. However, animating the face model obtained by this method is difficult. As a result, a realistic face model is usually generated by the model-based approach in which a generic face model is deformed to match the user's face image.

In this special issue, Goto et al. and Morishima describe their face modeling work based on the frontal face image and the profile image [39], [40]. Goto et al. present the two steps of their automatic face cloning. These two steps are global matching of a generic face for approximation and recognition of facial features for detail fitting. On the other hand, Morishima demonstrates the improved quality of the 3-D model when multiview face images are used.

Face Animation

There are two approaches for face animation: the flipbook method (key-frame animation) [10] and the wire-frame (parameterized) model approach [11]. The wire-frame method is more computationally intensive than the flipbook method, but it allows more flexible animation [12]. An example of the flipbook method can be found in [13]. On the other hand, a wire-frame model can be two-dimensional (2-D) such as the CANDIDE [14] or 3-D such as FaceWorks [15].

Ekman and Friesen developed the facial action coding system (FACS) [16] to describe all visually distinguishable facial movements. In FACS, the facial movements are based on the combination of action units that control facial expressions. Perlin's responsive face [17] demonstrates a subset of the full range of the facial expressions. Facial expressions can also be reproduced for a realistic 3-D head model by updating the dynamic video texture map [18].

Based on the MPEG-4 standard, Goto et al. describe in this special issue their face animation using facial animation parameters extracted for facial expressions [40]. On the other hand, Morishima describes in this special issue a muscle-based face image synthesis by first using a multilayer back-propagation network to estimate the muscle parameter and then resynthesizing facial expressions based on the facial muscle model [39].

Lip Synchronization

Interaction between audio and visual information in human speech perception has been a fascinating research field for many years. The well-known McGurk effect [19] demonstrates the bimodality of speech perception by showing that, when given conflicting audio and visual cues, a person may perceive a sound that never exists in either modality. To ensure lip synchronization in face animation, there are two approaches. The first approach is using simple energy detection to convert an incoming speech signal directly to an angle for the mouth opening, as shown in Fig. 1. The second approach is analyzing the speech signal to determine which phonemes are present and map them to the basic subset of all visemes, i.e., mouth configurations, that the facial model is capable of generating.

In this special issue, Goto et al. describe their phoneme extraction based on linear predictive analysis and neural networks together with the energy content and the average zero-crossing rate [40]. On the other hand, Morishima describes in this special issue a lip-synchronization method by using a neural network on a frame-by-frame basis to analyze the spoken voice and convert it into mouth shape parameters [39].

Hand Gesture Analysis

Hand gesture is a natural way of human-human interaction. It often complements speech to enhance clarity or to show emotions during communication. For example, a person may point at somewhere to show the direction or wave at someone to say goodbye. In order to keep this natural way of communication in the virtual environment and extend it for human-computer interaction, we need to have a better understanding of hand gestures.

In this special issue, Wu et al. review many recent techniques in human hand modeling and analysis, including some of their own research results [41]. They report that the hand can be modeled in several aspects such as the shape, the kinematical structure, dynamics, and semantics. Hand motion can be captured by finding the global and local hand movements with which the hand posture can be recovered. The meanings of the hand gestures could be interpreted from both temporal gestures and static hand postures.

Hand Gesture Synthesis

Before speaking, the avatar may raise a hand in order to get the attention of other users, as shown in Fig. 2. In this special issue, Wu et al. discuss some realistic hand gesture synthesis methods [41]. A keyframe-based method can be used to animate the hand by interpolating prespecified key-frames corresponding to the hand states. Alternatively, captured motion can be edited to improve the quality of the hand movements.

The hand movements can also be driven by the energy of the user's speech to enhance the speech communication. When the user is speaking at a normal voice level, the avatar can make small hand gestures like any human making a regular speech. Once the user emphasizes certain words by raising the voice or by making sharp bursts of tones, the avatar can make sharp and brisk hand gestures to show the emphasis on those words.

User Interfaces

Visual Interface

The visual interface is the most important component of a virtual environment in giving the users the feeling of presence. The video display corresponds to the position of the user and produces the correct view when the user is navigating in the environment. The most common display device is a color monitor. However, a color monitor by itself does not provide too much immersion to the visual interface because the size of the monitor is limited; thus, the user can still see things in the physical world surrounding the monitor. In order to add more immersion, the monitor output can be projected onto a large screen so that the user can feel that he or she is actually present in the virtual environment. To provide a more 3-D visual experience, shutter glasses can be used to create stereoscopic display. The monitor refresh rate can be doubled (typically to 120 Hz) so that the pair of stereoscopic views is time-multiplexed

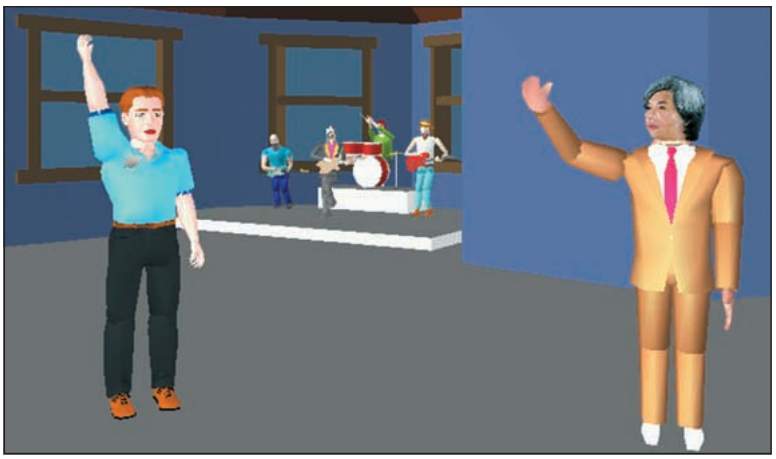
▲ *1. Mouth synchronized with energy of speech.*

to be shown for one eye at a time. Studies have shown that the use of shuttle glass with a projection system provides a significant amount of the sense of immersion [20]. In this special issue, Tamagawa et al. provide examples of display systems (CABIN, CAVE, CoCABIN, UNIVERS, and COSMOS) that can render a wide field of view and stereo images with high resolution [38]. Alternatively, the head mount display (HMD) can be used as the visual interface. One disadvantage of the HMD is that the user may feel uncomfortable wearing this helmet-like device after some time.

In this special issue, Ichikawa et al. discuss their 3-D image space display unit, which uses three projectors and a curved screen [37]. The unit can ensure sufficient brightness at the space corresponding to the user location. The curved screen enables the camera capturing the user's image to be placed to the left of, to the right of, above, or below the screen.

Auditory Interface

The auditory interface is another important component in a multiuser virtual environment, because audio is a very effective means of communication. Simple playback of the audio recorded by the speaker does not render an immersive sound environment, because the user cannot tell where the sound comes from in the virtual environment. The directional sound technology (sometimes referred to as the 3-D sound) can be used to make the auditory interface more realistic and increases the sense of presence in a virtual environment [21]. As a result, a user can hear an increasingly louder voice while moving towards the speaker. Similarly, the voice fades away when the user moves away from the speaker. Moreover, different proportions of sound can be steered among the left and right sound channels according to the listener's and speaker's relative positions so that the listener can feel which person is talking by listening to the direction of the sound. The 3-D sound effect can be achieved simply by weighting the output from the left and right sound output channels or by making use of the head-related transfer functions [22].



▲ *2. Hand gesture synthesis.*

Primary Input Interface

The primary input interface allows the user to input commands directly to the virtual environment. A keyboard is the traditional input device, and the commands are associated with keystrokes. However, initially the user has to spend some time learning the command keystrokes. For this reason, a 2-D pointing device such as a mouse can be used to simplify the control mechanism and increase the sense of immersion. For example, an event can be triggered by a mouse click. A mouse can also be used for navigation in the virtual environment [23]. However, because the mouse is a 2-D pointing device, controlling the navigational direction in a 3-D environment can be confusing. For this reason, 3-D pointing devices [24] can be used instead to facilitate 3-D navigation.

Tracking Interface

In addition to the primary input devices, the level of interaction can be enhanced further if tracking is used to acquire certain user's behaviors automatically. For example, head tracking can be used to determine which direction the user is facing, and eye tracking can be used to determine the user's gaze. Such information can be used to render the user's avatar and allow the user to interact with the environment. Figure 3 shows example results of a face and eye tracking system [25].

In this special issue Goto et al. describe their work on real-time facial feature tracking [40]. During the initialization phase, various parameters containing information for tracking face position and facial features are generated automatically. Then the mouth tracking and eye tracking are performed based on the edge and the gray level information around the mouth and the eyes.

Hand tracking can be used to determine the user's hand gestures and reproduce them in the virtual environment. Hand gestures can also be used as a primary input interface to augment or even replace the keyboard and the mouse. As reviewed by Wu et al. in this special issue, color-based segmentation is an efficient way for hand localization [41]. Wu et al. present two approaches for color-based tracking: nonparametric approaches based on color histograms and parametric approaches based on modeling the color density as Gaussian distribution or Gaussian mixture.

Data Sharing

Streaming of Audio and Video

In a virtual environment, it is essential to have an efficient way of transmitting multimedia information, especially video and audio because they consume a relatively large amount of bandwidth even after compression. It may take minutes or even hours to download a video sequence; thus, users will lose interest if they need to wait for such a long time. This delay can be reduced significantly by

streaming the video and audio content such that the content is sent segment by segment, and upon receiving a segment it is displayed to the user at the same time the next segment is being received. RealPlayer [26] is an example application for streaming video and audio media. It should be noted that multimedia content is not limited to video and audio; often they are combined with other multimedia data such as text, images, and drawings. We have been working on the streaming of presentation slides synchronized with the video and audio captured during a presentation session [27]. The streaming technology is particularly important for real-time interaction in a multiuser virtual environment where a user may want to show a video clip and discuss it with other users at the same time in the virtual room.

Shared Whiteboard

In an office, a whiteboard can be used to sketch a plan, write down reminders, or illustrate an idea during discussion. Some video conferencing applications such as CUseeMe [3] or NetMeeting [2] incorporate an electronic shared whiteboard that facilitates information sharing among users. The shared whiteboard allows multiple users from different locations to do collaborative work. Traditionally, a whiteboard (or blackboard) is used in a classroom for the teacher to write down course materials for the students to learn. As a result, a shared whiteboard can be included in a multiuser virtual environment so that it is suitable for distance learning. In addition, more than one shared whiteboard can appear in a virtual environment, and each whiteboard may be used for a specific function that can be associated with the surroundings. For example, in a virtual office building, users may use the whiteboard for checking important phone numbers or appointment times in their offices, or they may leave a message for other users on the whiteboard in the corridor.

Sharing and Manipulation of 3-D Objects

In addition to 2-D information conveyed by the shared whiteboard, 3-D objects can also be shared in a multiuser virtual environment. For example, in the multiuser networked game Quake, the weapons and the ammunitions (ammos) are examples of 3-D objects shared by multiple players. In the virtual world of the game, the players see the same set of weapons and ammos and a player can pick up an item by passing through it. Other multiuser virtual environments, such as InterSpace [28], allow users to build their own virtual worlds and let other users enter these worlds, by first sending the world information to the other users. Usually these systems contain 3-D objects that are predefined by the central server. It is more appealing, however, if a user is able to share a 3-D object from his or her side to the other users. For example, during a discussion, a user wants to show different components of a car, and then the user can select a 3-D car model stored in his or her computer to be sent to other users so that the 3-D car model appears in the virtual environment seen by all users. Then the user can explain different components of a car by simply referring to the 3-D model. Thus the sharing and manipulation of 3-D objects makes the communication more interactive.

Another issue for sharing and manipulating the 3-D objects is the transmission. The 3-D models usually have one or more attributed data such as normals, colors, or textures. A complex scene or a complex object can have a large size, and it may take a very long time to download. As a result, like video and audio, the 3-D model should be streamed so that a coarse model is seen first and the model is updated for refinement as more information is received. Progressive 3-D streaming as proposed in [29] sends the more important bits to represent the 3-D model and then the less important bits, allowing the user to see the model when only a few bits are received, and the user can stop the download at any time, yet retain the best available quality of the model.

Existing Multiuser 3-D Virtual Environments

A comparison of existing multiuser 3-D virtual environments is shown in Table 1. Each of these systems has its pros and cons. One of the earlier efforts in providing the 3-D virtual environment is CAVE [30]. In CAVE, projection screens are installed on several walls so that 3-D objects and scenes appear to be inside and outside the projection room. However, this system is expensive and requires a lot of space, therefore making it impractical to be used in general. In InterSpace [28] the avatar's face is represented by either a 2-D facial icon or by the video capturing the user's face. As a result, the view is not immersive enough when looking at the avatar in a nonfrontal direction. An emerging trend in other multiuser 3-D virtual environments (ActiveWorld [31], OuterWorld [32], and Cybertown [33]) is to provide e-business services such as online shopping in a 3-D vir-

▲ *3. Face and eye tracking.*

tual shop. For example, after the user enters a fashion shop in a virtual shopping mall, the user can choose different styles of T-shirt to be worn on a virtual mannequin and then look at it from various viewpoints in order to make his or her selection. If the user decides to buy it, then he or she can check out the item and the item will be delivered. While these systems provide an immersive interaction between a user and the virtual environment, they pay less attention to the immersive interaction, especially multimodal interaction, between the users. In OnLive! Traveler, also known as DigitalSpace [34], although the lips are synchronized with speech there is no body associated with the avatar so the avatar's head is floating in the air.

We have been developing a multiuser 3-D virtual prototype environment called networked intelligent virtual environment (NetICE) [35]. Figure 3 is a sample snapshot of NetICE. We focus on immersive technologies that enhance multimodal interaction between the users. We believe that besides speech, behaviors such as eye gaze, hand gestures, lip synchronization, and facial expressions are all important aspects of communication and

Table 1. Comparison Between Existing Multiuser 3-D Virtual Environments.

Multiuser 3-D Virtual Environment	*Avatar Model*	*Face Animation*	*Body Animation*	*Lip Sync*	*3-D Audio*	*Streaming Video*	*Shared Whiteboard*	*Sharing and Manipulation of 3-D Object*	*Tracking*
NetICE	3-D body model	Yes	Moderate	Yes	Yes	Yes	Yes	Yes	Face and eye tracking, ongoing effort for hand tracking
CAVE	Very simple 3-D model	No	Few	No	No	Yes	No	Yes	Head and hand tracking
DIVE [36]	3-D body model	No	Moderate	No	No	Yes	Yes	Yes	N/A
InterSpace	3-D body model with 2-D facial icon or video showing face	Facial expressions shown by 2-D icons	None	No	Only responsive to distance but not directional	Yes	No	No	N/A
OnLive! Traveler	3-D head without body	Yes	N/A	Yes	Yes	No	No	No	N/A
OuterWorld	Small 3-D body model	Yes	Many	No	No	No	No	No	
CyberTown	Small 3-D body model	No	None	No	No	No	No	Yes	N/A
ActiveWorlds	Small 3-D body model	Yes	Many	No	No	No	No	No	N/A

should be performed by the computer using speech processing, animation, and tracking techniques to increase the sense of immersion. Therefore, these are the technologies we have built into NetICE. A demo version of NetICE can be downloaded from http://amp.ece.cmu.edu/.

Several other virtual environments are also presented in this special issue. In the system described by Tamagawa et al. [38], 2.5-D video avatars are mounted between CABIN and COSMOS, which are multiscreen immersive projection displays with five and six screens. These displays were connected by a high bandwidth ATM network. Two stereo cameras are placed in the CABIN and the COSMOS, and the generated 2.5-D video avatars are composed in the shared virtual world on each site.

In [39], Morishima describes a natural communication environment between multiple users in cyberspace by transmission of natural voice and real-time synthesis of an avatar's facial expression. An emotion model is embedded into the system, and gaze tracking and mouth closing point detection are realized.

Potential Applications

Each of the systems described in this article has its own target application. To summarize, here is a list of potential applications that will be offered by multiuser 3-D virtual environments in general.

Virtual Conferencing and 3-D Chat Rooms

People from different geographical locations can have business or personal meetings in the virtual environment. They can use shared whiteboard to exchange their ideas and stream the video and audio for a presentation. Users can also enter 3-D chat rooms in order to gather or make new friends. This is a convenient way of socializing because users do not need to waste time traveling.

Virtual Shopping On-Line

When users enter current virtual shopping malls, they can browse the products by looking at the image or reading the text description of the products. With immersive communication technologies, a virtual salesman can be created to assist potential customers in order to provide a more interactive customer service.

Telemedicine

Doctors can simulate an operation in a virtual emergency room. They can share 3-D objects such as a heart model to be operated on and then exchange opinions. Virtual nurses can be created to calm patients or remind patients to take pills. This reduces the burden of the real nurses, yet the patients will feel that more attention is paid to them by the combination of virtual and real nurses.

Distance Learning and Training

There are already several distance learning programs that are based on the broadcast of the lecture videos. However, while the students can have a good view of the professor, the professor may not be able to see all the students if the lecture is broadcast to many different locations. Besides, it is difficult for the students to tell whether the professor is having eye contact with them. These can be resolved if everyone is placed in a virtual environment so that students may come from different locations, yet they can interact with the professor immersively.

Virtual Collaboration

When people from various locations try to work together to design a car, they can meet in a virtual garage where the engineers build the engines and the designers work on the aesthetic aspect of the car body. Different components of the cars are represented by shared 3-D models, and the two teams can discuss if the design needs to be modified.

Discussion

In this article, we have introduced various immersive interactive technologies and surveyed existing multiuser 3-D virtual environment systems. We believe that in the future, with higher bandwidth, better compression, and faster rendering techniques, more complex 3-D models with detailed textures can be used to render an even more realistic environment. Faster and more reliable face tracking will allow the avatar facial model to be individualized, while less input will be required from the users as the intelligent computer can understand what the users want implicitly. In a few years, a PC will become a ubiquitous platform that is no longer merely a computation tool but will be a common communication tool like the telephone. With the abundance of bandwidth, more and more PCs will be connected together to form communities of different interest groups. Each group will have its own virtual environment where community members can meet, and multiple environments can be hyperlinked together. Eventually, these multiuser 3-D virtual environments will evolve from a human-computer interface to a truly immersive human-to-human interface to provide transparent communication between people.

Acknowledgment

This work was supported in part by IBM University Partnership Program and NSF CAREER Award.

Wing Ho Leung is currently a Ph.D. student in electrical and computer engineering at Carnegie Mellon University. He received the B. Eng. degree in electrical engineering in 1998 from McGill University, Canada, and an M.S. in electrical and computer engineering in 2000 from Carnegie Mellon University. He was with the eMedia group at IBM T.J. Watson Research Center during the summers of

1999 and 2000. His research interests include multimedia signal and image processing, information retrieval, and 3-D virtual conferencing applications. His current research focus is on image compression for image-based rendering applications and retrieval for vector-based graphics.

Tsuhan Chen received the Ph.D. degree in electrical engineering from the California Institute of Technology in 1993. From 1993 to 1997, he was with AT&T Bell Laboratories, Holmdel, New Jersey, and later at AT&T Labs-Research, Red Bank, New Jersey, as a Senior Technical Staff Member and then a Principle Technical Staff Member. Since 1997, he has been with the Electrical and Computer Engineering Department, Carnegie Mellon University, as an Associate Professor. His research interests include multimedia signal processing and communication, audio-visual interaction, video coding and multimedia standards. He cofounded and chaired the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He is as Associate Editor for *IEEE Transactions on Image Processing* and *IEEE Transactions on Multimedia*. He serves on the Steering Committee of *IEEE Transactions on Multimedia* and the Editorial Board of *IEEE Signal Processing Magazine*. He is a technical co-chair of the First IEEE Conference on Multimedia and Expo. He has written many technical papers and holds seven U.S. patents. He is a recipient of the National Science Foundation CAREER Award.

References

[1] ICQ. Online chat software. [Online]. Available: http://www.icq.com/

[2] Microsoft. NetMeeting 3 videoconferencing software. [Online]. Available: http://www.microsoft.com/windows/netmeeting/

[3] CUseeMe Networks. Videoconferencing software. [Online]. Available: http://www.cuseeme.com/

[4] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, vol. 28, pp. 289-304, 1965.

[5] J.C. Tang and E.A. Isaacs, "Why do users like video? Studies of multimedia-supported collaboration," *Computer Supported Cooperative Work (CSCW)*, vol. 1, pp. 163-193, 1993.

[6] Humanoid Animation Working Group (H-Anim). [Online]. Available: http://www.ece.uwaterloo.ca/~h-anim/

[7] MiraLab, University of Geneva. [Online]. Available: http://miralabwww.unige.ch/index.html

[8] R. Boulic, N.M. Magnenat-Thalmann, and D. Thalmann, "A global human walking model with real time kinematic personification," *The Visual Computer*, vol. 6, no. 6, pp. 344-358, 1990.

[9] R. Mas-Sanso and D. Thalmann, "A hand control and automatic grasping system for synthetic actors," in *Proc. Eurographic'94*, pp. 167-178.

[10] P. Griffin and H. Noot, "The FERSA project for lip-sync animation."

[11] K. Waters and T.M. Levergood, "DECface: An automatic lip-synchronization algorithm for synthetic faces," DEC Cambridge Research Lab. Tech. Rep., Sept. 1993.

[12] T. Chen and R. Rao, "Audio–visual integration in multimodal communication," *Proc. IEEE*, vol. 86, pp. 837-852, May 1998.

[13] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *ACM SIGGRAPH 97*, pp. 353-360.

[14] M. Rydfalk, "CANDIDE: A parameterized face," Linkoping Univ. Rep. LiTH-ISY-I-0866, Oct. 1987.

[15] Digital Equipment Corporation. FaceWorks facial animation software. [Online]. Available: http://interface.digital.com/

[16] P. Ekman and W. Friesen, *The Facial Action Coding System*. San Francisco, CA: Consulting Psychologists, 1978.

[17] K. Perlin. Responsive Face facial animation demo. [Online]. Available: http://www.mrl.nyu.edu/perlin/facedemo/

[18] W.H. Leung, B. Tseng, Z.-Y. Shae, F. Hendriks, and T. Chen, "Realistic video avatar," in *IEEE Int. Conf. Multimedia and Expo.*, New York, July 2000, pp. 631-634.

[19] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, Dec. 1976.

[20] C. Ware, K. Arthus, and K.S. Booth, *Fish Tank Virtual Reality. In Interchi '93: Bridges Between Worlds*. Addison-Wesley: Reading, MA, pp. 37-42.

[21] C.M. Hendrix, "Exploratory studies on the sense of presence in virtual environments as a function of visual and auditory display parameters," Master's thesis, Human Interface Technol. Lab., Washington Technol. Center, Univ.Washington, 1994.

[22] W.G. Gardner and K.D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc.*, vol. 97, pp. 3907-3908, June 1995.

[23] M. Chen, S.J. Mountford, and A. Sellen, "A study in interactive 3-D rotation using 2-D control devices," in *Proc. ACM Siggraph'88*, pp. 121-129.

[24] K. Hinckley, J. Tulio, R. Pausch, D. Proffitt, and N. Kassell, "Usability analysis of 3D rotation techniques," in *Proc. ACM Symp. User Interface Software and Technology*, 1997, pp. 1-10.

[25] Advanced Multimedia Processing Lab., Carnegie Mellon University. Face Tracking. [Online]. Available: http://amp.ece.cmu.edu/

[26] RealPlayer. Streaming media application. [Online]. Available: http://www.real.com/

[27] Advanced Multimedia Processing Lab., Carnegie Mellon University. Streaming media player. [Online]. Available: http://amp.ece.cmu.edu/

[28] Nippon Telegraph and Telephone (NTT). InterSpace multiuser 3D virtual environment. [Online]. Available: http://www.ntts.com/ispace.html

[29] M. Okuda and T. Chen, "Joint geometry/texture progressive coding of 3D models," in *IEEE Int. Conf. Image Processing*, Vancouver, Canada, Sept. 2000, pp. 632-635.

[30] Electronic Visualization Lab., Univ. Illinois, Chicago. CAVE projection-based virtual reality system. [Online]. Available: http://www.evl.uic.edu/pape/CAVE/

[31] Activeworlds.com, Inc. ActiveWorlds multiuser 3D virtual environment. [Online]. Available: http://www.activeworlds.com/

[32]. OuterWorld. Multiuser 3D virtual environment, see Ray Studios, LLC. [Online]. Available: http://www.outerworlds.com/crown/

[33] Cybertown. Multiuser 3D virtual environment, blaxxun interactive. [Online]. Available: http://www.cybertown.com/

[34] OnLive! Traveler multiuser 3D virtual environment. [Online]. Available: http://www.onlive.com/

[35] W.H. Leung, G. Goudeaux, S. Panichpapiboon, S.-B. Wang, and T. Chen, "Networked collaborative environment (NetICE)," in *IEEE Int. Conf. Multimedia and Expo.*, New York, July 2000, pp. 1645-1648.

[36] Swedish Institute of Computer Science (SICS). DIVE multiuser 3D virtual environment. [Online]. Available: http://www.sics.se/dce/dive/dive.html

[37] T. Ichikawa, K. Yamada, and T. Kanamaru, "Multimedia ambiance communication," *IEEE Signal Processing Mag.*, vol. 18, pp. 43-50, May 2001.

[38] K. Tamagawa, T. Yamada, T. Ogi, and M. Hirose, "Developing a 2.5D video avatar," *IEEE Signal Processing Mag.,* vol. 18, pp. 35-42, May 2001.

[39] S. Morishima, "Face analysis and synthesis," *IEEE Signal Processing Mag.,* vol. 18, pp. 26-34, May 2001.

[40] T. Goto, S. Kshirsagar, and N. Magnenat-Thalmann, "Automatic face cloning and animation," *IEEE Signal Processing Mag.,* vol. 18, pp. 17-25, May 2001.

[41] Y. Wu and T.S. Huang, "Hand modeling, analysis, and recognition," *IEEE Signal Processing Mag.*, vol. 18, pp. 51-60, May 2001.