Virtual Metamorphosis

Jun Ohya, Jun Kurumisawa, and Ryohei Nakatsu
ATR Media Integration & Communications Research Laboratories

Kazuyuki Ebihara
Victor Company of Japan

Shoichiro Iwasawa
Seikei University

David Harwood and Thanarat Horprasert
University of Maryland

The virtual metamorphosis system lets people change their forms into any other form in a virtual scene. To realize these changes, a computer vision system estimates facial expressions and body postures and reproduces them in a computer graphics avatar in real time. We introduce three systems in order of their development: the Virtual Kabuki system, Networked Theater, and "Shall We Dance?"

We all want to change our appearance in some way. For example, makeup helps enhance appearance, although it primarily relates to the face. Many people try to change their moods by changing their accessories, clothes, and hairstyles. Some people even resort to plastic surgery to make permanent changes.

We often observe behaviors that stem from a more radical desire for metamorphosis. For example, amusement parks and sightseeing areas sometimes have standing character boards of monsters or animation characters with their face areas cut out. Children and adults alike enjoy putting their faces into the cutouts and having their photographs taken (Figure 1). People simply enjoy changing their forms.

People aren't satisfied with just static metamorphosis, however. Children often pretend to act as their favorite movie heroes, monsters, and animation characters. A child's desire for dynamic metamorphosis apparently persists into adulthood. For example, MGM Studios in Santa Monica, California, provides an attraction that lets participants (mostly adults) appear in the scenes of popular TV shows. This experience excites the participants, and the audience also enjoys watching.

Figure 1. Metamorphosis using a standing character board.

Virtual metamorphosis

Recently, the importance of communications between remotely located people has escalated due to the trend toward more collaboration and the increasing time, energy, and expense associated with transportation. Telephones have provided the major means of communication so far. However, for natural human communication, visual information also proves important. For this purpose, telecommunications manufacturers have developed video phones and videoconferencing systems. Yet users find it difficult to overcome the feeling that they're situated at distant locations. In addition, users hesitate to have their faces appear on the receiver's display. Developing technologies that enable a metamorphosis in appearance may offer the key to making remote visual communications widely used—and more fun.

ATR Media Integration and Communications Research Laboratories (ATR MIC) have targeted communication environments in which remotely located people can communicate with each other via a virtual scene. A possible application of such an environment includes a virtual metamorphosis system. Realizing such a system requires estimating the facial expressions and body postures of the person to be "morphed" using noncontact and real-time methods and reproducing the information in the form the user desires. We emphasize that facial expressions and body postures should be estimated in a noncontact fashion, more specifically, by computer-vision-based technologies.

Taking these requirements into consideration, ATR MIC developed three virtual metamorphsis

1070-986X/99/$10.00 © 1999 IEEE

systems. In the first, called the Virtual Kabuki system,[1] users can change their form to a Kabuki actor's form in virtual Kabuki scenes. This system estimates a person's facial expressions and body postures in real time from the facial images acquired by a small charge-coupled device (CCD) camera and from the thermal images acquired by an infrared camera, respectively. A Silicon Graphics Onyx Reality Engine II reproduces the estimated facial expressions and body postures in real time in a Kabuki actor's 3D model.

For its second virtual metamorphosis system, ATR MIC developed the PC-based Networked Theater, in which multiple sites link to networks so that multiple persons can metamorphose in one common virtual scene. In Networked Theater, PCs handle the computer graphics rendering processes, and a CCD camera estimates body postures. Moreover, background scenes can be created with edited video sequences and/or computer graphics images.

Since body postures must be estimated in 2D in Networked Theater, ATR MIC addressed real-time 3D posture estimation in a collaborative project with the University of Maryland Institute of Advanced Computer Studies (UMIACS), College Park, Maryland and the Massachusetts Institute of Technology's Media Lab. In the "Shall We Dance?" system, body postures of persons at multiple sites can be estimated in 3D in real time using normal CCD cameras and ordinary PCs.

Virtual Kabuki system

As shown in Figure 2, the Virtual Kabuki system consists of three main processing modules:

1. creation of 3D models of Kabuki actors,
2. real-time, passive (noncontact) estimation of facial expressions and body postures of the person targeted to metamorphose into the Kabuki actor, and
3. real-time reproduction of the estimated facial expressions and body postures in the Kabuki actor's model.

Modeling Kabuki actors

Prior to metamorphosis, we model the Kabuki actors' forms using wireframe models to represent the 3D shapes. We map color textures corresponding to the color data of costumes and skin to the wireframe models. We can deform the models by moving the wireframe models' vertices according to the estimated posture data. We can also map the color textures to the deformed models. As shown in Figure 2, the Kabuki actors' models are stored in graphics workstations like the SGI Onyx Reality Engine II.

Real-time estimation

Conventional VR systems use contact methods to measure facial expressions and body postures.[2] Although quite accurate, these cumbersome methods have limited applications. For this reason, ATR MIC developed passive methods to estimate facial expressions and body postures in real time using computer-vision-based techniques.

Previous computer-vision-based methods classified facial expressions from a video sequence,[3,4] but couldn't estimate the facial expression at each time instant. We developed a method to estimate deformations of facial components such as the eyes and mouth using a frequency domain transform.[5] The person who will metamorphose into a Kabuki actor wears a helmet to which we attach a small CCD camera pointed at the face and mounted on a chin-level bar perhaps four inches away. The camera stays at the same position relative to the face regardless of head movement.

In the face image acquired by the camera, a window overlays each facial component. Each window is converted to frequency domain data by discrete cosine transform (DCT), which proves a very efficient computation. More specifically, each window divides into sub-blocks, where a sub-block consists of 8 by 8 pixels. Then the system applies DCT to each sub-block and calculates the summations of DCT energies in the horizontal, vertical, and diagonal directions in each sub-block (Figure

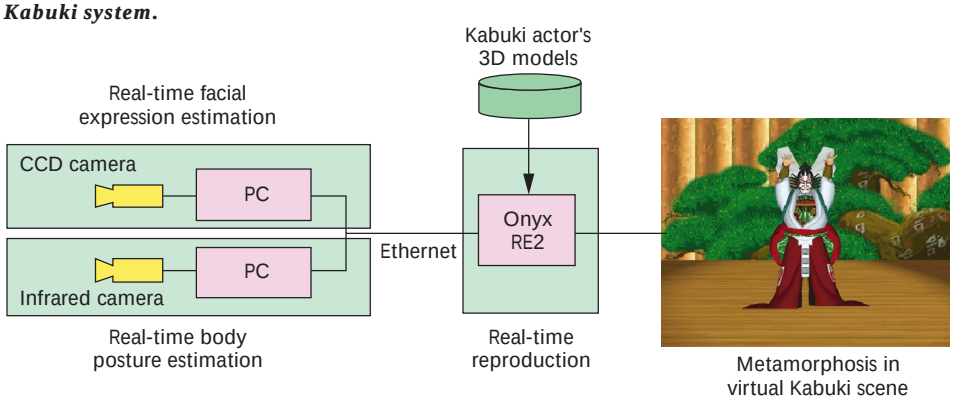


Figure 2. Virtual Kabuki system.

3). To obtain a DCT feature for a direction (among the three directions), the system adds the summations in that direction. Why do the DCT features prove useful? As Figure 4 illustrates, changes in the DCT features represent shape changes in the facial components. In Figure 4, closing the eye decreases the value of the horizontal feature and increases the vertical feature. The DCT features for the three directions reproduce the facial deformation in the face model.

Many computer-vision-based approaches proposed body posture estimation,[6-9] but real-time, stable estimation of different postures failed. The Virtual Kabuki system estimates the person's body posture from a thermal image (Figure 5, left column) acquired by an infrared camera.[10] We use thermal images because a human body's silhouette can be robustly extracted from the background, independent of changes in lighting conditions and costume colors. First, the system locates some significant points such as the top of the head and the fingertips by analyzing the extracted silhouette's contour. In addition, to reproduce the entire posture of a human body, we need to estimate the locations of main joints such as the elbows and knees. These areas prove difficult to estimate by a simple analysis of the silhouette contour because the main joints don't always produce salient features on the contour.

Therefore, we developed a learning-based method that estimates the joints' positions using the located positions of the significant points. That is, we use polynomials constructed in advance by a learning procedure to calculate the main joints' coordinates from the coordinates of the located significant points. Since estimating the values of the polynomials' coefficients is a combinatorial optimization problem, we used a genetic algorithm[11] to determine the coefficient values from a sample data set. In the middle column of Figure 5, small squares indicate the located positions of the significant points and joints.

Real-time reproduction

To reproduce facial expressions in the Kabuki actor's 3D model, we used the estimated DCT features to deform the face model. A previously developed method for synthesizing facial expressions based on a physics model[12] showed good performance, but the system didn't run in real time. Therefore, ATR MIC developed a real-time reproduction method based on anatomy for artists, or plastic anatomy.[13] An artistic principle for deformation, anatomy for artists—proposed in Italy in the 15th century—makes works such as paintings and sculptures look realistic.[14]

Our proposed method consists of a modeling and a reproduction process. The modeling process generates the 10 reference facial expressions an artist chooses in a 3D face model, where the 3D displacement vector from a neutral facial expres-

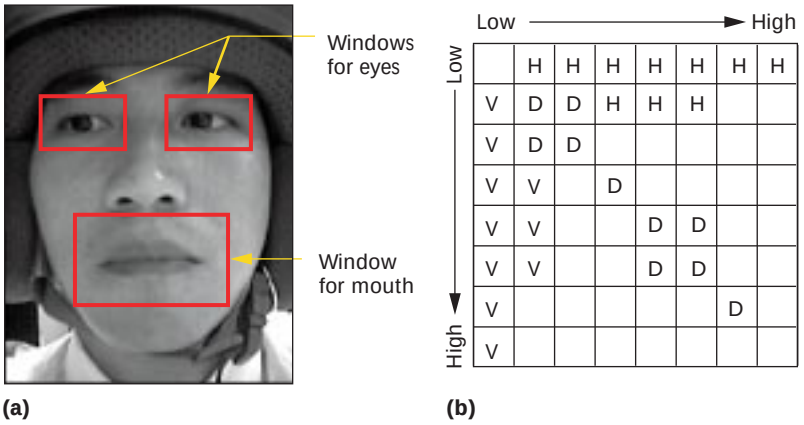


Figure 3. DCT-based facial expression esimation. (a) Windows for facial expression estimation. (b) DCT feature calculation for each sub-block in a window. DCT features for the horizontal, diagonal, and vertical directions result from adding the DCT energies of the H, D, and V pixels, respectively.

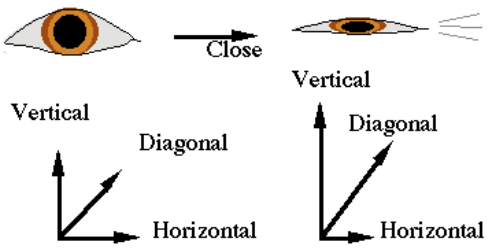


Figure 4. Changes in DCT features according to shape changes in facial components.

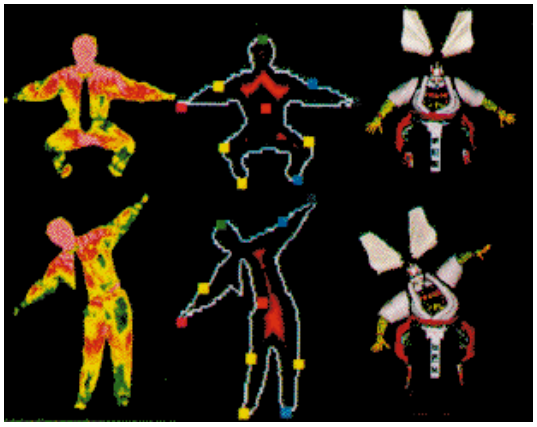


Figure 5. Body posture estimation and reproduction: thermal images (left), human body posture estimation (middle), and posture reproduction (right).

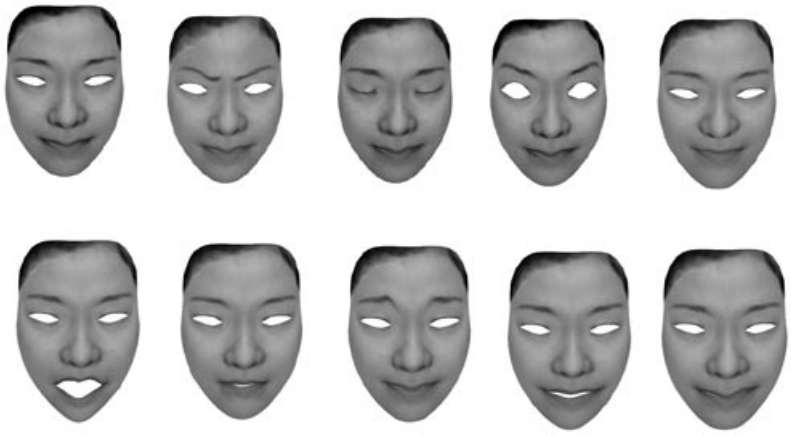


Figure 6. Reference facial expressions created by an artist.

sion is recorded at each vertex of the face model for each reference expression (Figure 6). The artist chooses the reference expressions according to the principle of anatomy for artists. Therefore, some expressions include deformations that humans cannot actually display.

To generate intermediate facial expressions requires mixing the 10 reference facial expressions. Thus, at each vertex of the 3D face model, the artist determines the mixing rate of the 3D displacement vectors of each reference facial expression for each image of a sample image set containing many different facial expressions. For each determination the artist compares the real expression to the generated expression. Subsequently, at each vertex a linear combination of the 10 reference displacement vectors is constructed. Each coefficient (mixing rate) of the linear combination is represented by a linear combination of the DCT features. A genetic-algorithm-based learning procedure determines the coefficients of the linear combination of the DCT features. In the reproduction process, we input the DCT features obtained from the estimation process into the linear combination for the mixing rates to calculate each vertex's displacement vector. Figure 7 shows some facial expression reproductions.

Figure 7. Examples of facial expression reproduction.

To reproduce the estimated 2D body postures in a 3D Kabuki actor's model, we used a genetic-algorithm-based learning procedure to construct in advance polynomials that convert the 2D coordinates of the significant points and joints to 3D coordinates. The estimated 2D coordinates inserted into the polynomials reproduce 3D postures in the Kabuki actor's model reasonably well. The right column of Figure 5 shows the reproductions.

We implemented the estimation and reproduction processes as shown in Figure 2. To estimate facial expressions, we used a Gateway 2000 PC. An infrared camera (Nikon Thermal Vision Laird3) captured thermal images for estimating body postures. The system estimated human body postures from the thermal images on an SGI Indy workstation. We used an SGI Onyx Reality Engine 2 to render a Kabuki actor's 3D model and a Kabuki scene, and a large projector screen to dis-

Figure 8. Scenes of the Virtual Kabuki system. The infrared camera observes the person, and the estimated postures are reproduced in the Kabuki actor's model displayed in the projector screen (top row). The camera fixed to the helmet observes the face, and the estimated facial expressions are reproduced in the Kabuki actor's face model (bottom row).

play the images. Figure 8 shows some scenes of the Virtual Kabuki system. In this implementation, the process runs at approximately 20 frames per second for facial expression estimation and body posture estimation, and 12 fps for rendering. Although the speed for the estimations is good, the rendering speed needs improvement. ATR MIC presented a live demonstration to many people at the Digital Bayou in Siggraph 96, held in New Orleans in August 1996.

Networked Theater

The Virtual Kabuki system has the following limitations:

1. only one person at one site can join,
2. it requires expensive facilities,
3. the person needs to wear a helmet-mounted camera, and
4. the background scene stays fixed.

To overcome these limitations, ATR MIC constructed the next version of metamorphosis systems, called Networked Theater. The final system image of the Networked Theater will link multiple sites, including offices and homes, through networks so that many people can metamorphose in a common virtual scene. Since we hope the system will have home-use applications, we used ordinary PCs and normal CCD cameras. Additionally, we implemented functions for editing video image sequences so that users can display arbitrary background scenes. Using the editing functions, Networked Theater users can produce video programs and movies in which they appear as actors or actresses. Here, we describe the new technologies in the Networked Theater.

Human-body silhouette extraction

Since the infrared camera used in the Virtual Kabuki system was quite expensive, we replaced it with a normal CCD camera. We faced the challenge of extracting a human body silhouette from the background regardless of changes in lighting conditions and background. ATR MIC developed a method[15] that uses a CCD camera to obtain images whose pixels have *YIQ* values, where *Y* is intensity, and *I* and *Q* represent colors. Before extracting the silhouette, we acquired the mean *YIQ* values of each background pixel and the threshold values for the *Y*, *I*, and *Q* components from training image sequences on the background. Based on the mean and the threshold values, we classified each pixel as the human silhouette or background. We also achieved stable silhouette extraction with this approach.

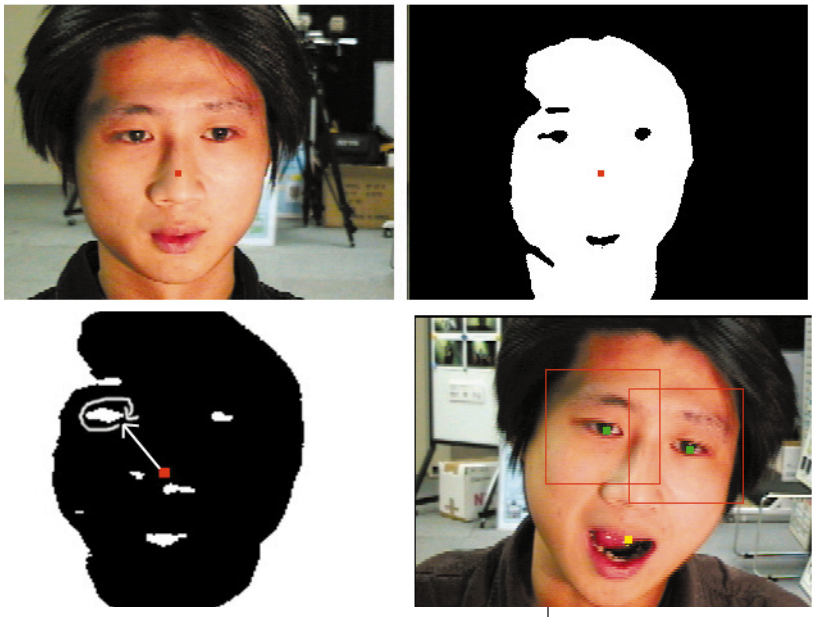


Figure 9. Automatic face tracking. Upper left and right: the centroid of the skin color area is located. Lower left: the computer measures the distances between facial components in advance. Lower right: windows are applied to facial components.

Automatic face tracking

We replaced the helmet-mounted camera worn by the participant in the Virtual Kabuki system with a camera separated from the person. For the DCT-based estimation of deformations of facial components, the face should ideally be observed continually from the front. Therefore, our system uses a gaze-controllable camera whose viewing direction can be computer controlled. More specifically, we use the Sony EVI-30, which a PC can pan and tilt. The camera automatically tracks the participant's face, and the participant's facial expressions can be estimated in the same manner as that of the helmet method. In the tracking method, the system learns the facial skin color and geometrical relationship between facial components in advance. These data control the camera so that the centroid of the skin color area stays at the center of each frame. Then the system applies windows to the facial components using the geometrical relationship so that deformations of facial components can be estimated with the DCT-based approach. Figure 9 shows a face tracking and window application.

Editing video sequences

In the Virtual Kabuki system users can create the background scene with computer graphics soft-

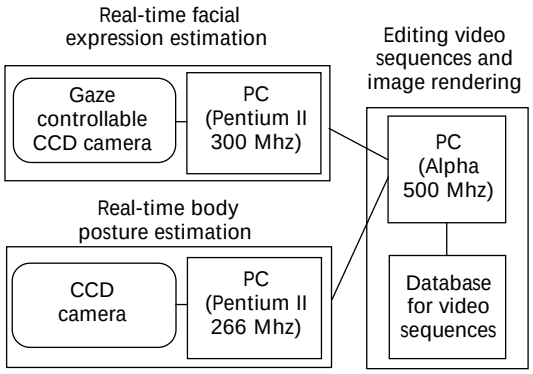


Figure 10. Block diagram of the Networked Theater.

ware. We often want to use real images as well as computer graphics-based images for background scenes because computer graphics images sometimes look artificial. Also, creating computer graphics images generally proves expensive in both time and cost. Therefore, we developed a method to use video sequences for background scenes. In our system, we divided all video sequences into short segments and stored them. The database management commands—based on scenarios—can connect short video segments or merge short segments with computer graphics images. This lets users generate several background scenes.

Implementation and performance

As shown in Figure 10, we implemented the described methods and functions on a PC-based system. The face tracking and facial expression estimations ran at 15 fps, and the silhouette extraction and posture estimation ran at 16 fps. A DEC Alpha handled the editing functions and rendering edited image sequences at 20 fps, displaying up to eight avatars simultaneously. Figure 11 shows some scenes in which the avatars appear in different backgrounds edited by the system.

Shall We Dance?

Although the Networked Theater permits estimating human postures in real time from the monocular images acquired by a CCD camera, the posture estimation occurs only in 2D. The "Shall We Dance?" project—a real-time 3D computer vision system for detecting and tracking human movement—provides a person with control over the movement of a virtual computer-graphics character. The project lets remotely located people control 3D avatars in one virtual space so that they (actually the avatars) can dance together realistically.

While the entertainment industry uses active systems and marker systems[16,17] for motion capture, mass-market applications can't rely on such systems either because of cost or the impracticality of people entering an environment while fitted with active devices or special reflectors. Due to these restrictions, a vision-based motion-capture system that doesn't require contact devices would have significant advantages. Many researchers have tried to detect and track human motion.[6,18-21] To the best of our knowledge, none of their systems can detect and track a full 3D human body's motion in real time. Our approach applies an inexpensive and completely unencumbered computer-vision system to the motion-capture problem. Our system runs on a network of dual Pentium 400 PCs at 20 to 30 fps (depending on the size of person observed).

System overview

Figure 12 shows the block diagram of the system. Multiple motion-capture systems can connect to a common "graphical reproduction" system to let many people interact in the virtual world.

Our motion-capture system uses four to six color CCD cameras to observe a person. Each camera attaches to a PC running an extended version of the University of Maryland's W4 system,[22] which detects people and locates and tracks body parts. The system performs background subtraction, silhouette analysis, and template matching to locate the 2D positions of salient body parts such as the head, torso, hands, and feet in the image. A central controller obtains the 3D posi-

Figure 11. Scenes from the Networked Theater. A CCD camera captures the person's motion and reproduces it in the red character displayed on the screen.

tions of these body parts using triangulation and optimization processes. The extended Kalman filters and kinematics constraints of human motion developed at MIT (and described elsewhere[23]) smooth the motion trajectories and predict the body part locations for the next frame. We then feed this prediction back to each instance of W4 to help in 2D localization. The graphic reproduction system uses the body posture output to render and animate the cartoon-like character.

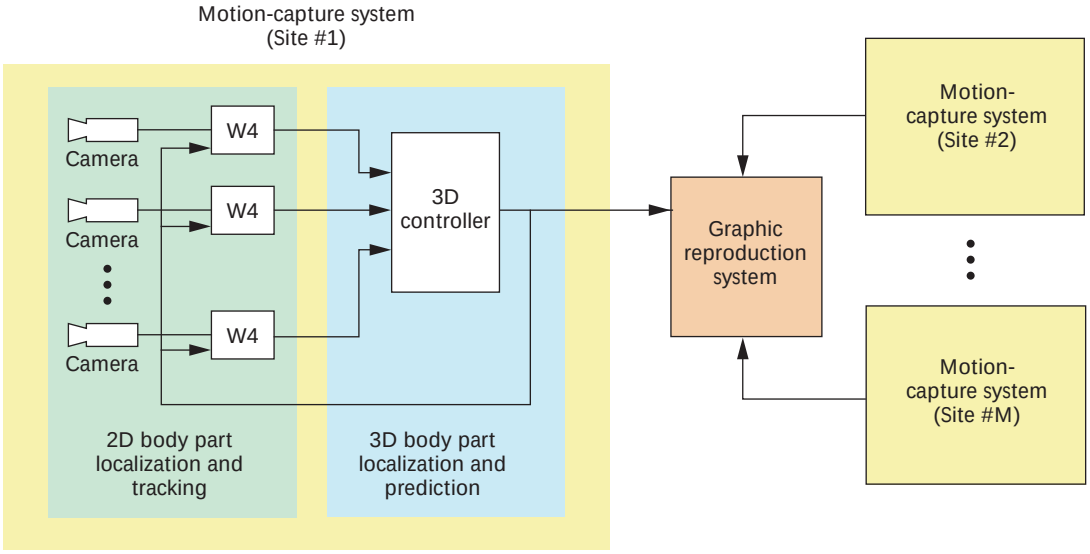


Figure 12. Block diagram of the "Shall We Dance?" real-time 3D motion-capture system.

W4 system

W4 combines shape analysis and robust tracking techniques to detect people and locate and track their body parts (head, hands, feet, and torso). The original version of W4 was designed to work only with visible monochromatic video sources taken from a stationary camera, but the current version now handles color images. W4 consists of five computational components:

- modeling backgrounds,
- detecting foreground objects,
- estimating motion of foreground objects,
- tracking and labeling objects, and
- locating and tracking human body parts.

The system statically models the background scene. For each frame in the video sequence, the system segments the moving objects from the background with a new pixel classification method[24] that uses a geometric cardboard model of a person in a standard upright pose to model the body's shape and locate the head, torso, hands, legs, and feet. The system then tracks these parts using silhouette analysis and template matching, which uses color information. After predicting the locations of the head and hands using the cardboard model, the system verifies and refines their positions using dynamic template matching, then updates these templates. Figure 13 (next page) illustrates W4's person detection and body part localization algorithm.

3D posture estimation and dynamic modeling

Integrating the location data from each image permits estimating 3D body postures. First, we calibrate the cameras to obtain their internal and external parameters. For each frame in the sequence, each instance of W4 not only sends the body part location data but also a corresponding confidence value that indicates to the central controller the level of confidence of its 2D localization for each particular part. The controller then computes the 3D localization of the body part by performing triangulation over the set of 2D data with higher confidence values than a specified threshold. In this process we treat each body part separately. That is, at a certain frame, we can obtain the 3D position of the right hand and the left hand from triangulating the cameras' different subsets.

To constrain the body motion and smooth the motion trajectory, we employed a model of human body dynamics[23] developed by MIT's Media Lab. However, the framework was computationally expensive, especially when applied to the whole body. Thus we experimented with a computationally light-weight version that uses several linear Kalman filters in tracking and predicting the locations of the individual body parts. These predictions are then fed back to the W4 sys-

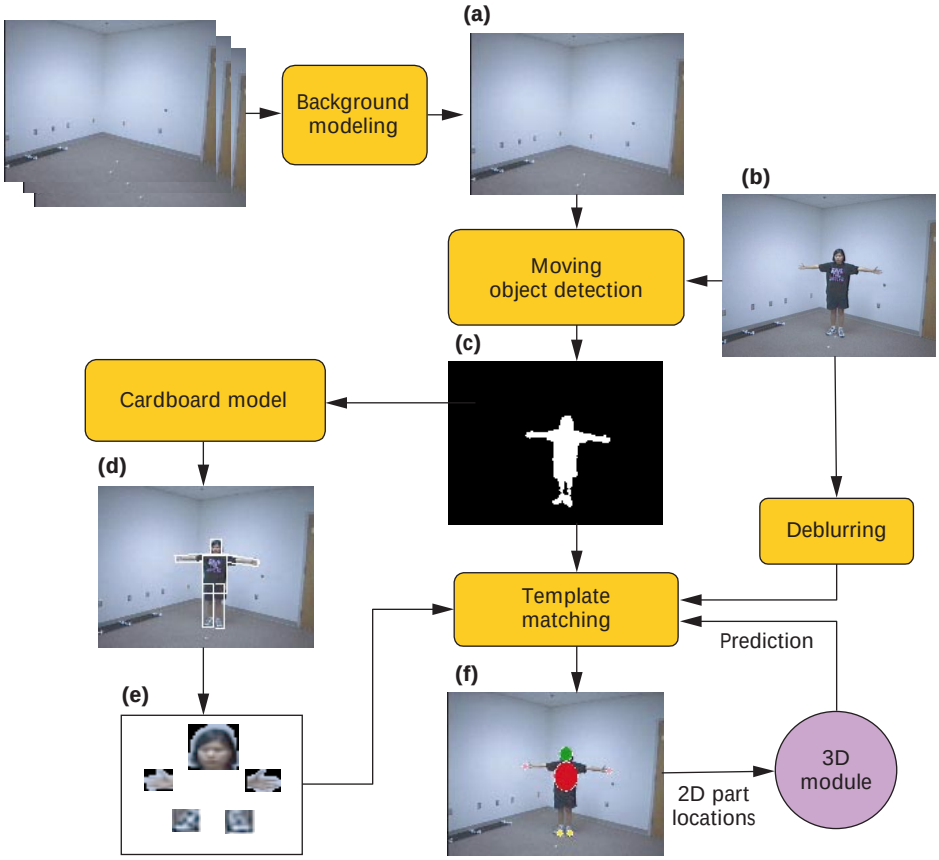


Figure 13. The 2D body-part localization process diagram. First, the system models the background scene (a). For each frame in the video sequence (b), the algorithm segments the moving object from the background using a new method of pixel classification (c). Based on the extracted silhouette and the original image, the system analyzes the cardboard model (d) and creates salient body part templates (e). Finally, the system locates these parts (head, torso, hands, and feet) using a combined method of shape analysis and color matching (f).

tems to control their 2D tracking. We found this system worked well and required much less hand-tuning than the fully dynamic model.

Figure 14 demonstrates four keyframes of a video sequence captured in our lab. We used four cameras in this sequence with only a single motion-capture system to control the movement of both computer graphics characters.

We recently demonstrated the system at Siggraph 98's Emerging Technologies exhibit. The demonstration consisted of two sites of 3D motion capture. The two sites sent the estimated 3D body-posture parameters to a central graphics reproduction system so that two people could dance in the virtual space. Figure 15 shows a snapshot of our demonstration area. The person shown controlled the movement of the character dressed in a tuxedo. A person at the other site controlled the sumo wrestler cartoon. In this demonstration, a person entered the exhibit area and momentarily assumed a fixed posture that allowed the system to initialize (that is, locate the person's head, torso, hands, and feet). The participants could then "dance" freely in the area. The trajectories of their body parts controlled the animation of the cartoon-like character. Whenever the tracking failed, the person could reinitialize the system by assuming a fixed posture at the center of the demonstration area.

Lessons learned and future work

This article introduced three systems developed for virtual metamorphosis, a concept originally proposed by ATR MIC. Virtual metamorphosis systems capture the facial expressions and body postures of many people to control avatars in a virtual space. Doing this requires computer-vision-based technologies for estimating facial expressions and body postures in real time. In addition, inexpensive implementations based on PCs and normal CCD cameras are important for eventual home use. The three systems we developed taught us progressively more about how to achieve these goals.

1. Exhibiting the Virtual Kabuki system to many people demonstrated the effectiveness of the concept of virtual metamorphosis. However, the system required the single participant to

wear a helmet, was expensive, and produced 2D results.

2. The Networked Theater cost less and allowed more than one person to join the system. People didn't have to wear a helmet and could edit background scenes on a PC.

3. The "Shall We Dance" project, our most recent system, demonstrates real-time "3D" estimation of human body postures using computer-vision technologies.

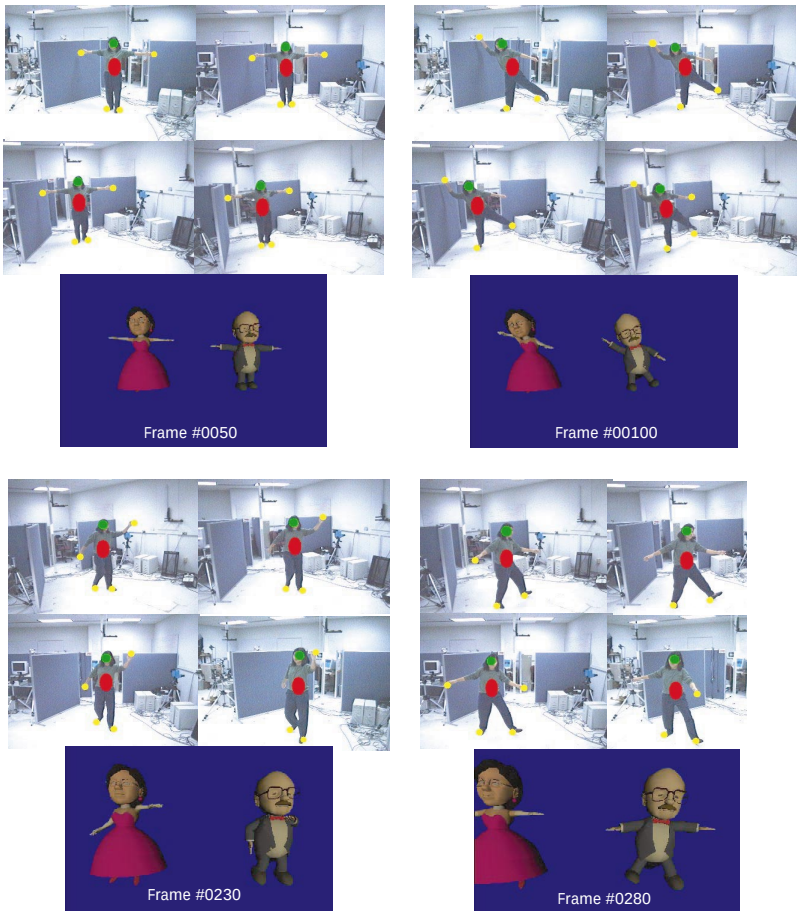


Figure 14. Some keyframes of a video sequence captured with our modified human body posture estimation system.

However, we still need to improve 3D posture estimation. We should be able to deal with arbitrary postures, including crossed hands in front of the body. Interactions between avatars in virtual spaces also need development. In addition, we would like to work with acoustic information. Finally, we hope to edit and use background scenes more effectively.
MM

Acknowledgments

We thank Tatsumi Sakaguchi and Katsuhiro Takematsu of ATR Media Integration & Communications Research Laboratories, Kyoto, Japan and Masanori Yamada of NTT Human Interface Laboratories (previously at ATR) for their cooperation and support in developing the Networked Theater

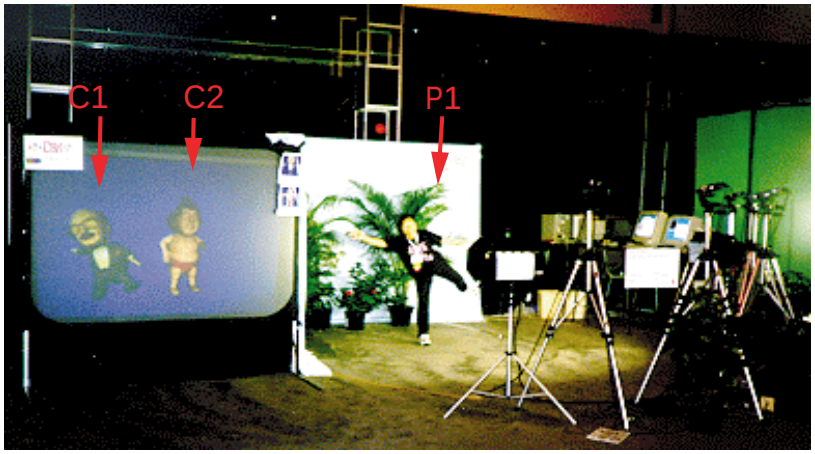


Figure 15. A snapshot of our demonstration area at Siggraph 98. The motion of the person (P1) in the dancing area controlled the movement of the tuxedo-wearing cartoon character (C1). A person in another dancing area controlled the movement of the sumo wrestler cartoon (C2).

and the "Shall We Dance?" system. We would like to thank Ismail Haritaoglu of the University of Maryland, College Park, Maryland, for his help in integrating his W4 system into the "Shall We Dance?" project, and Larry S. Davis of the University of Maryland for his guidance and support. Finally, we gratefully acknowledge the support of Chris Wren and Alex Pentland of the MIT Media Lab in developing the "Shall We Dance?" project.

References

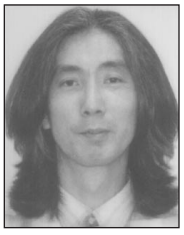
1. J. Ohya et al., "Virtual Kabuki Theater: Towards the Realization of Human Metamorphosis Systems," *Fifth IEEE Int'l Workshop on Robot and Human Communication (RO-MAN 96)*, IEEE Press, Piscataway, N.J., Nov. 1996, pp. 416-421.
2. J. Ohya et al., "Virtual Space Teleconferencing: Real-Time Reproduction of 3D Human Images," *J. of Visual Communication and Image Representation*, Vol. 6, No. 1, Mar. 1995, pp. 1-25.
3. M.J. Black et al., "Tracking and Recognizing Rigid and Nonrigid Facial Motions Using Local Parametric Models of Image Motion," *Proc. Fifth Int'l Conf. on Computer Vision*, IEEE CS Press, Los Alamitos, Calif., June 1995, pp. 374-381.
4. L.A. Essa et al., "Facial Expression Recognition Using a Dynamic Model and Motion Energy," *Proc. Fifth Int'l Conf. on Computer Vision*, IEEE CS Press, Los Alamitos, Calif., June 1995, pp. 360-367.
5. K. Ebihara, J. Ohya, and F. Kishino, "Real-Time Facial Expression Detection Based on Frequency Domain Transform," *Proc. Visual Communications and Image Processing 96*, SPIE Vol. 2727, SPIE Press, Bellingham, Wash., Mar. 1996, pp. 916-926.
6. C.R. Wren et al., "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1998, pp. 780-785.
7. D. Terzopoulos and D. Metaxas, "Dynamic 3D Models With Local and Global Deformations: Deformable Superquadrics," *IEEE Trans. PAMI*, Vol.13, No.7, 1991, pp. 703-714.
8. Y. Kameda, M. Minoh, and K. Ikeda, "Three-Dimensional Pose Estimation of an Articulated Object from its Silhouette Image," *Proc. Asian Conference on Computer Vision*, Institute of Electronics, Information, and Communication Engineers, Tokyo, 1993, pp. 612-615.
9. S. Kurakake and R. Nevatia, "Description and Tracking of Moving Articulated Objects," *Proc. 11th Int'l Conf. on Pattern Recognition*, Int'l Assoc. of Pattern Recognition, Surrey, UK, Vol. I, 1992, pp. 491-495.
10. S. Iwasawa et al., "Real-Time Estimation of Human Body Posture from Monocular Thermal Images," *Proc. 1997 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., Jun. 1997, pp. 15-20.
11. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, Mass., 1989.
12. D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," *IEEE Trans. PAMI*, Vol.15, No. 6, June 1993, pp. 569-579.
13. K. Ebihara et al., "Real-Time Facial Expression Reproduction Based on Anatomy for Artists for Virtual Metamorphosis Systems," *Trans. the Institute of Electronics, Information, and Communication Engineers*, D-II, Vol. J81-D-II, No.5, May 1998, pp. 841-849 (in Japanese).
14. Y. Nakao and M. Miyanaga, *Atlas of Anatomy for Artists*, Nanzan-do, Japan, Oct. 1986.
15. M. Yamada, K. Ebihara, and J. Ohya, "A New Robust Real-Time Method for Extracting Human Silhouettes from Color Images," *Proc. Third IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., Apr. 1998, pp. 528-533.
16. B. Delaney, "On the Trail of the Shadow Woman: The Mystery of Motion Capture," *IEEE Computer Graphics and Applications*, Vol. 18, No. 5, Sept./Oct. 1998, pp. 14-19.
17. D.J. Sturman, "Computer Puppetry," *IEEE Computer Graphics and Applications*, Vol. 18, No. 1, Jan./Feb. 1998, pp. 38-45.
18. D.M. Gavrila and L.S. Davis, "Toward 3D Model-Based Tracking and Recognition of Human Movement," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Computer Science Dept. Multimedia Laboratory, Univ. of Zurich, Zurich, 1995
19. C. Bregler and J. Malik, "Tracking People With Twists and Exponential Maps," *Proc. Computer Vision and Pattern Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., 1998, pp. 8-15.
20. I.A. Kakadiaris and D. Metaxas, "Model-Based Estimation of 3D Human Motion With Occlusion Based on Active Multiviewpoint Selection," *Proc. Computer Vision and Pattern Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., 1998, pp. 81-87.
21. Q. Cai and J.K. Aggarwal, "Tracking Human Motion Using Multiple Cameras," *Proc. Int'l Conf. on Pattern Recognition*, Int'l. Assoc. of Pattern Recognition, Surrey, UK, August 1996.
22. I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People," *Proc. Third IEEE Int'l*

Conf. on Automatic Face and Gesture Recognition, IEEE Computer Society Press, Los Alamitos, Calif., April 1998, pp. 222-227.

23. C.R. Wren and A. Pentland, "Dynamic Modeling of Human Motion," *Proc. Third IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, IEEE Computer Society Press, Los Alamitos, Calif., April 1998, pp. 22-27
24. T. Horprasert et al., "Real-Time 3D Motion Capture," *Proc. Workshop on Perceptual User Interfaces*, M. Turk, ed., Nov. 1998, pp. 87-90, http://research.microsoft.com/PUIWorkshop/Papers/Horprasert.pdf/.

Jun Ohya is a department head of ATR Media Integration & Communications Research Laboratories. His research interests includes computer vision, computer graphics and virtual reality. He received BS, MS, and PhD degrees in precision machinery engineering from the University of Tokyo, Japan, in 1977, 1979, and 1988, respectively. In 1979 he joined NTT Research Laboratories and transferred to ATR in 1992. He stayed at the University of Maryland's Computer Vision Laboratory as a visiting research associate from 1988 to 1989.

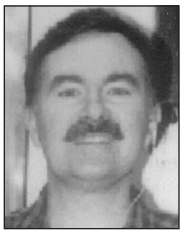
Kazuyuki Ebihara is a researcher at the Victor Company of Japan. He originally joined Victor Company of Japan in 1985, where he developed signal processing technologies for broadcasting systems. He transferred to ATR in 1994 and was engaged in real-time human image analysis and synthesis methods. He received a BS in computer science from Kanazawa Institute of Technology, Japan, in 1982, and a PhD in electronics and information systems engineering from Osaka University, Japan, in 1998.



Jun Kurumisawa received a BA from the National University of Fine Arts and Music, Tokyo, Japan, in 1983. After working at several production companies as a computer graphics artist, he joined ATR Media Integration & Communications Research Laboratories in 1996. His interest includes constructing interactive artistic systems and integration of fine arts and computer science technologies.

Shoichiro Iwasawa is a PhD student at Seikei University, Japan. He earned BS and MS degrees in electrical engineering and electronics from Seikei University, Japan, in 1994 and 1996, respectively. From 1996 to 1997 he worked at ATR. His research interests include computer vision and computer graphics.

Ryohei Nakatsu is the president of ATR Media Integration & Communications Research Laboratories. His research interests include emotion recognition, nonverbal communications, and integration of multimodalities in communications. He received BS, MS, and PhD degrees in electronic engineering from Kyoto University in 1969, 1971, and 1982 respectively.



David Harwood is a senior member of the research staff of the Institute for Advanced Computer Studies at the University of Maryland, College Park with more than forty publications in the field of computer image analysis.

Thanarat Horprasert is currently a PhD candidate in computer science working with Larry Davis at University of Maryland, College Park. Her research interests include human motion understanding and real-time vision. She received a BS in computer engineering from Chulalongkorn University, Bangkok, Thailand, in 1992, and an MS in computer science from the University of Southern California, Los Angeles, in 1995. She is a member of the Upsilon Pi Epsilon Computer Science Honor Society and the Phi Kappa Phi National Honor Society.

Readers may contact Ohya at ATR Media Integration and Communications Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan, e-mail ohya@mic.atr.co.jp.