Low-Cost Telepresence for Collaborative Virtual Environments

Seon-Min Rhee, Remo Ziegler, *Student Member*, *IEEE*, Jiyoung Park, Martin Naef, *Member*, *IEEE*, Markus Gross, *Senior Member*, *IEEE*, and Myoung-Hee Kim, *Member*, *IEEE*

Abstract—We present a novel low-cost method for visual communication and telepresence in a CAVE™-like environment, relying on 2D stereo-based video avatars. The system combines a selection of proven efficient algorithms and approximations in a unique way, resulting in a convincing stereoscopic real-time representation of a remote user acquired in a spatially immersive display. The system was designed to extend existing projection systems with acquisition capabilities requiring minimal hardware modifications and cost. The system uses infrared-based image segmentation to enable concurrent acquisition and projection in an immersive environment without a static background. The system consists of two color cameras and two additional b/w cameras used for segmentation in the near-IR spectrum. There is no need for special optics as the mask and color image are merged using image-warping based on a depth estimation. The resulting stereo image stream is compressed, streamed across a network, and displayed as a frame-sequential stereo texture on a billboard in the remote virtual environment.

Index Terms—Remote systems, segmentation, stereo, virtual reality, teleconferencing.

1 Introduction

The introduction of telepresence features into collaborative applications promises improved productivity as well as significantly higher user acceptance of long distance collaboration (Fig. 1). In addition to having a voice connection, rendering a high-quality image of the remote user in realtime provides a feeling of co-presence [22]. In particular, applications in immersive virtual environments are prime candidates as the integration of the user representation into the virtual shared world provides increased social proximity, compared to a separate video. Practical implementations of telepresence systems in immersive environments, however, pose some severe difficulties, namely, the lighting requirements for video acquisition contradict with the needs for high quality projection. Strong lighting and uniform backgrounds would ease acquisition and segmentation of the user, however, for projection, the user has to stand in a dark room, surrounded by projection surfaces, such as in a CAVE™ [3].

- *S.-M. Rhee and J. Park are with the Visual Computing and Virtual Reality Laboratory, Department of Computer Science and Engineering, Ewha Womans University, 405-1, Ewha-SK Telecom Bldg., 11-1, Daehyun-dong, Seodaemun-gu, 120-750, Seoul, Korea. E-mail: {blue, lemie}@ewhain.net.*
- *R. Ziegler and M. Gross are with the Computer Graphics Laboratory, Department of Computer Science, ETH Zurich, ETH Zentrum, IFW D 27.1, Haldeneggsteig 4, CH-8092, Zurich, Switzerland. E-mail: {remo.ziegler, grossm}@inf.ethz.ch.*
- *M. Naef is with the Glasgow School of Art, Digital Design Studio, House for an Art Lover, 10 Dumbreck Road, Glasgow G41 5BW UK. E-mail: mnaef@acm.org.*
- *M.-H. Kim is with the Visual Computing and Virtual Reality Laboratory, Department of Computer Science and Engineering/Center for Computer Graphics and Virtual Reality, Ewha Womans University, 400, Ewha-SK Telecom Bldg., 11-1, Daehyun-dong, Seodaemun-gu, 120-750, Seoul, Korea. E-mail: mhkim@ewha.ac.kr.*

Manuscript received 7 Nov. 2005; revised 6 Feb. 2006; accepted 3 May 2006; published online 8 Nov. 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-0152-1105.

Numerous approaches, leading from complex and expensive systems such as that in the blue-c [8], to simpler systems using stereo cameras and IR-lights in combination with beam-splitters in [2], have been tried to solve the problem. However, very high cost, static background, long processing times or difficult calibration are limiting factors of previous systems.

The work presented in this paper demonstrates a way of extending existing spatially immersive displays with off-the-shelf components. The robustness of the segmentation and simplicity of the used image-warping approach lead to a real-time system with a high quality user representation while keeping the costs low.

1.1 Contributions

We present a novel system for concurrent projection and image acquisition inside a spatially immersive display. It provides a low-cost solution based on commodity hardware by combining segmentation in the near-infrared spectrum, image-warping to match the IR and color information from different cameras, and depth extraction through stereo matching to calculate disparities. The robustness of the segmentation and simplicity of the image-warping approach lead to a high quality user representation.

The main contributions of this work can be summarized as follows:

- We introduce a practical implementation for image segmentation in the near-infrared spectrum, which works in a standard CAVE™-like environment with active, concurrent projection.
- An image-warping algorithm and a stereo depth detection method enable us to combine the mask from the near-IR segmentation with a color image, relaxing the strict alignment requirements from traditional approaches employing beam splitters.
- Since no interpolation between different views is necessary using a stereo image stream, we maintain a very high texture quality.

Fig. 1. An example of a running collaboration with a snapshot of both sides.

- The acquisition and transmission system is integrated into an existing collaborative virtual reality toolkit.

The remainder of this paper is organized as follows: In Section 2, we discuss different approaches in related work. Section 3 gives a system overview of hardware and software components. We will describe some experiments including IR lamp positioning and reflection property characterization to generate good infrared reflective images for a robust user silhouette in Section 4. Section 5 explains silhouette fitting methods to generate stereo images for a stereo-video avatar, followed by a description of the integration of this avatar into the virtual world in Section 6. Finally we present results including performance analysis in Section 7 and conclude in Section 8.

2 Related Work

In this section, we briefly review some of the existing methods for acquisition and reconstruction of 3D shapes in general as well as acquisition of different representations for supporting telepresence in Collaborative Virtual Environments (CVEs).

2.1 Fast Acquisition and Reconstruction

The main problem when acquiring and reconstructing 3D objects for collaborative environments is to fulfill real-time constraints. In 1994, Laurentini [9] presented the "visual hull" which is created from several silhouette images. Based on the idea of visual hulls, Matusik et al. presented two very fast methods to create an image-based visual hull (IBVH) taking advantage of epipolar geometry in [13], and a polyhedral visual hull constructing a triangular surface representation in [12]. In the past, infrared-based structured light [4] has been successfully applied for reconstruction of 3D objects in a short amount of time. Narayanan et al. [15] combined the depth map and intensity image of each camera view at each time instant to form a Visible Surface Model. In 1998, Pollard and Hayes [19] rendered real scenes from new viewpoints using depth map representations by morphing live video streams.

2.2 Simultaneous Acquisition and Displaying

One of the key issues of these methods is the segmentation of foreground and background, which has been the topic of many papers to date. Telepresence can be provided by a video avatar, which is generated from live video streams from single or multiple cameras. The TELEPORT system presented in [7] uses 2D video avatars generated by a delta-keying technique. Raskar et al. [20] calculated depth information using a structured light technique in the Office of the Future. In the past few years, video avatar generation techniques based on multiple cameras (such as the National Tele-Immersion Initiative [21]) have been developed. Ogi et al. [16], [18] developed a 2.5D video avatar consisting of a triangulated and textured depth image. Using a Triclops Color Stereo Vision system made by Point Grey Research Inc., they apply a depth segmentation algorithm and use blue screens to improve the silhouettes by chroma-keying which is not feasible in a running CVE system. Ogi et al. [17] take the basis of segmentation by stereo cameras and blue screening to a further step and generate video avatars as plane, depth, and voxel-models according to their embedding needs. A very similar approach to segmentation has been chosen by Suh et al. [23], but the constraint of a constant background color is very restrictive.

Subramanian et al. [26] reconstruct a view dependent textured head model using an additional tracker device with complementary background subtraction while approximating the body by a bounding box. In [1] researchers from INRIA are applying a realtime reconstruction technique based on the visual hull algorithm developed at MIT assuming a static background.

Our work has been inspired by the blue-c system [8]. This system exploited and extended the time-slicing concept of active stereo projection to control the lighting and shutdown the projection within a short time interval to enable traditional segmentation methods. In contrast to our approach, the blue-c acquires and transmits a dynamic 3D video object [24]. It requires significantly higher processing power and custom built hardware to control the lighting and projection system.

2.2.1 Thermal Keying and IR-Segmentation

The systems mentioned above all control the background of the user to be acquired in the visible spectrum. A very different technique presented by Yasuda et al. [27] introduced a thermal vision camera in order to realize a human region segmentation based on natural heat emission of the body. However, the system fails to segment cold objects held by the user. An IR segmentation approach using a beam-splitter was presented in the Lightstage system [5] and also more recently in [2]. Instead of choosing an IR-reflective material to bounce the light back, infrared translucent material is lit from behind in [4]. In a system presented in [10] a special retroreflective background augmenting reflections of IR-light is chosen.

Instead of requiring precise mechanical alignment needed when using beam-splitters, we align the IR and color images using an image warping step, which is less error-prone. It also works well in a typical running CAVE^{TM}-like environment without specially treated screen material. Furthermore, we avoid complicated 3D geometry while preserving a sensation of depth by introducing a stereo billboard.

3 System Overview

Stereoscopic displays play an important role in virtual reality by increasing the feeling of immersion. Instead of

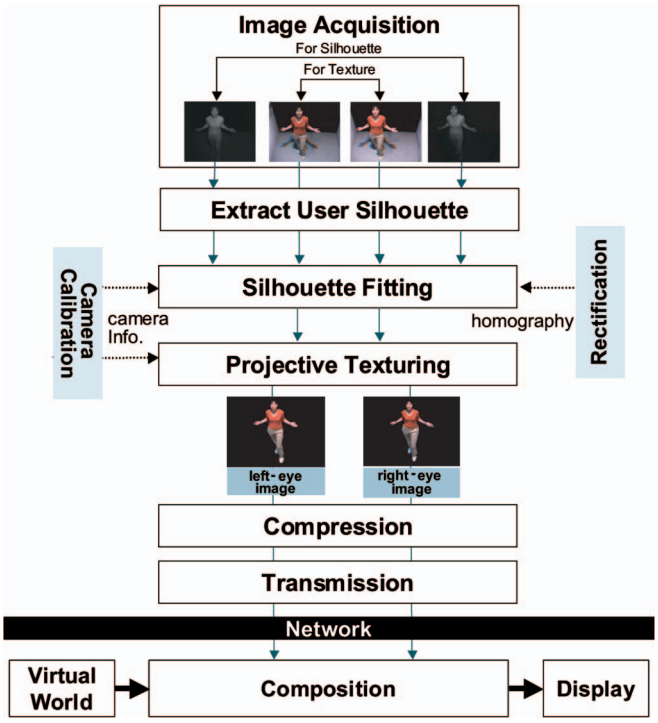


Fig. 2. System overview: the processing pipeline for the creation of stereo video avatars based on IR segmentation.

acquiring 3D geometry from the user, which is then rendered for both eyes separately onto the stereoscopic display, we directly capture images from two texture cameras placed at a distance corresponding roughly to binocular disparity.

In order to embed the captured collaborating party in the virtual world a segmentation of the texture images is necessary. We therefore use two camera pairs to generate a segmented, stereoscopic image sequence. Each pair consists of a grayscale camera with an attached IR bandpass filter and a color camera with an IR cutoff filter. We extract the user silhouette mask using the grayscale camera in the near infrared spectrum. (The silhouette mask will be referred to as "silhouette" throughout the remainder of the paper.) The silhouette can then be used as a mask on the texture image which is captured by the color camera, referred to as "texture camera."

In Sections 3.1 and 3.2, we describe the software and hardware components in more detail.

3.1 Software Components

Fig. 2 depicts the pipeline to generate the segmented stereo image stream to support telepresence. The sequential nature of the algorithm is reflected in the software components. Silhouettes of the user can be found for each eye by comparing the current IR image to a previously learned background image. Silhouettes are then fitted to the texture information in order to mask out the user from the background. Finally, the texture image is warped by simulating a backprojection onto the billboard.

The resulting images are compressed and transmitted to the collaborating site. There, the images are mapped onto a billboard that is placed into the virtual scene. Each step of the pipeline is explained in more detail in Sections 4, 5, 6, and 7.

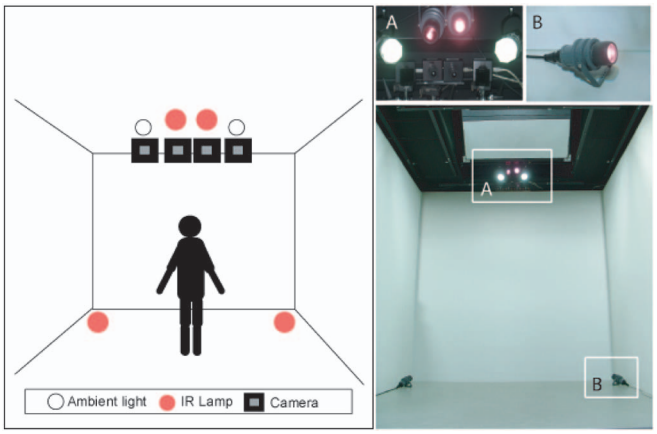


Fig. 3. Sketched and photographic view of the hardware setup, showing four IR lamps, four cameras, and two ambient lights. (a) Scaled view of one of the four used IR lights. (b) Scaled view of the specific arrangements of the components.

3.2 Hardware Components

The four sided CAVE^{TM}-like environment used for our experiments consists of a front, left, right, and a floor part, all measuring 2.4m $\times$ 2.4m, and four CRT projectors (BARCO-GRAPHICS 808s). Four cameras, four IR, and two ambient lights are arranged as depicted in Fig. 3.

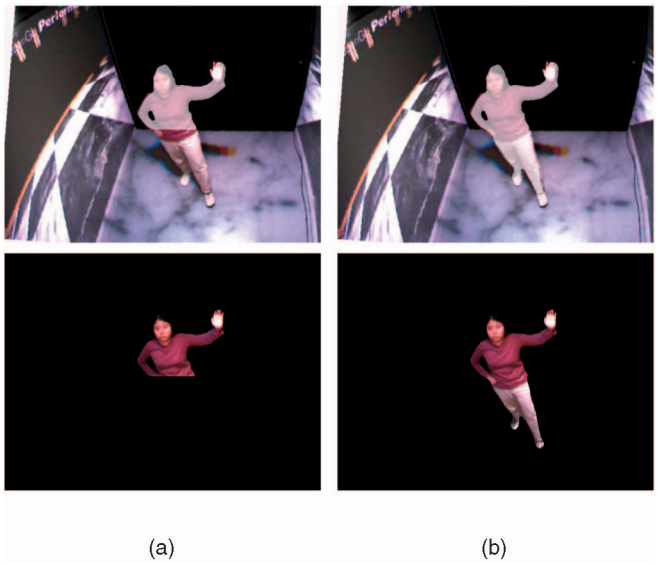
The four Dragonfly IEEE-1394 cameras from Point Grey Research Inc. are placed on the top of the front screen. Of these, two are texture cameras that are mounted in the center and two are silhouette cameras affixed on either side of the texture cameras. All four cameras are attached to a single computer to simplify the synchronization. We avoid occlusion of the projection by mounting the cameras on top of the front screen. Four 35W IR lamps with a light spectrum peaking at 730nm are carefully positioned to illuminate the user. Two of them are mounted on top next to the cameras, pointing roughly in the same direction as the cameras. Additional IR lights are placed in the left and right front corners on the floor. They point towards the middle of the front screen to diffuse the light and illuminate the user indirectly from the bottom. The advantages of this setup are discussed in detail in Section 4.1.

The projection environment has to be rather dark since we use CRT projectors. We add two lights in the visible spectrum to illuminate the user to yield a better texture quality. These lights must have a fairly narrow light cone to avoid illuminating the projection screen. The spotlights are mounted above the cameras on either side of the IR-lights.

4 Extract User Silhouette

To deal with the dynamic lighting environment caused by concurrent projection we use infrared reflective images as the input to a traditional background subtraction algorithm. The near-infrared illumination in our environment is static since the projectors only emit light in the visible spectrum and no other dynamic sources are present.

The physical setup of our CAVE^{TM}-like environment does not allow uniform illumination of the background without illuminating the person as well. If we restrict the IR illumination to the back wall we would only get half of the silhouette due to the camera pointing down on the user (see Fig. 4). We therefore illuminate the person from the front



(a) (b)

Fig. 4. Influence of IR back lighting versus front lighting. (a) Simulated mask and segmented image when using IR back lighting techniques. (b) Generated mask and segmented image when using IR front lighting techniques.

and use a background subtraction method optimized for infrared light based on [11].

Near-IR images only contain a single intensity channel. Careful placement of the light sources is therefore critical in order to achieve a robust segmentation, as compared to color image segmentation which typically relies on the hue and saturation channels as well. Furthermore, choosing the right materials for clothes and the background is important. A short analysis of typical, relevant materials is included in the following sections. Since there is no interference between the IR light and the texture cameras due to the integral IR cut-off filters, we can do the analysis independently of the projected content.

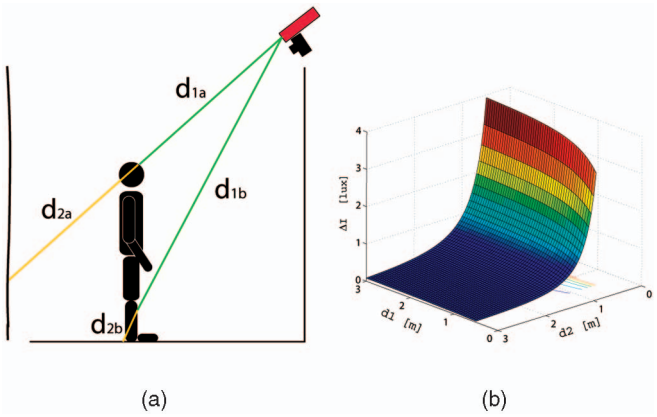
4.1 Infrared Lamp Positioning

The lights are positioned as close as possible to the IR-cameras to minimize shadows and self-shadowing artifacts.

In our setup, we use one IR-light above each IR-camera and one IR-light in each front corner of the CAVE™-like environment (Fig. 3). The two lights on the floor illuminate the legs indirectly by bouncing the light off the front wall (see Fig. 5a), while the two lights above are used to additionally illuminate the top part of the body (see Fig. 5b). Fig. 5c shows the result when using all the lights simultaneously.

(a) (b) (c)

Fig. 5. Segmentation depending on different setups of the IR lights. (a) IR lights only on the floor. (b) IR lights only on top of the front wall. (c) Combination of the two previous setups.



(a) (b)

Fig. 6. Intensity depending on distance from IR light source to object. (a) Illustration of possible distances. (b) Graph showing intensity difference depending on d_1 and d_2.

In order to get a good silhouette we have to maximize the intensity difference ΔI between the foreground captured at runtime and the initial background. In addition to careful positioning of the IR lights, we benefit from the intensity fall-off relative to the distance squared. Fig. 6 illustrates the rapid intensity difference when increasing the distance between foreground and background d_2.

Assuming a Lambertian surface for the background, surface orientation did not cause problems in our experiments, and specular highlights of the foreground improve segmentation further. However, it can create problems with hair under certain circumstances as discussed in Section 4.2.3.

4.2 Material, Skin and Hair

Most CAVE™-like environments are open to the back. By hanging a curtain with low IR-reflection as a back wall, unwanted reflection caused by common lab material can be removed.

The foreground material mostly consists of cloth, human skin, and hair. The cloth can of course be picked for specific reflection properties, whereas the skin and hair have set constraints.

4.2.1 *Characterize Cloth*

IR reflection is not related to visible colors, but to the combination of paint and fabrics (see Figs. 7a, 7b, and 7c).

Due to the difficulty of the IR reflectance characterization, we built a simple setup to measure the reflection properties of different materials (Fig. 7d). Using this simple scanning device, we can tell quickly if the cloth worn leads to good silhouettes. We capture an IR-image of the screen material and the curtain in the background at a distance d_c from the IR light and compare them to the intensity of the material captured at the same distance d_c.

The conservative way is to check if the difference between the foreground intensity I_F $(0 \leq I_F \leq 1)$ and the background intensity I_B $(0 \leq I_B \leq 1)$ exceeds a threshold T.

$$T < I_F - I_B. \tag{1}$$

More cloth can be accepted if we take the intensity fall-off into account. We can consider the worst case scenario for the segmentation by calculating the difference between the intensity of the foreground at maximum distance from the

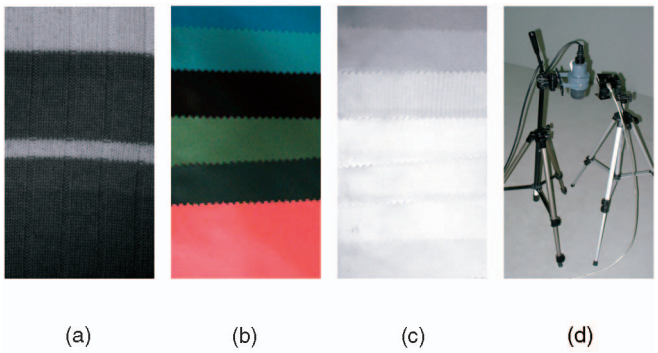


Fig. 7. IR reflection properties of different materials. (a) Part of a sweater consisting of one single material, but colored with different dyes. (b) Texture image of different color patches of the same material (c) and their IR reflection. (d) System used to scan material.

light d_{fmax} and the intensity of the background at minimum distance d_{bmin} from the light, where $d_{fmax} < d_{bmin}$ (see (2)).

$$T < \frac{I_F \cdot d_c^2}{d_{fmax}^2} - \frac{I_B \cdot d_c^2}{d_{bmin}^2}. \quad (2)$$

After scanning different cloth and various other materials, we found the majority to be reflective enough in order to be segmented correctly by our algorithm. Most people will therefore be able to use our system without changing clothes. We provide a jacket for those rare cases where the user wears nonreflective cloth.

4.2.2 Characterize Skin

If the IR reflection of skin is measured at the same distance and illumination as the background material, the segmentation does not perform well as can be seen in Fig. 8a. However, during collaboration, the user has a certain distance to the background material, which leads to a perfect segmentation of skin. In Fig. 8b, the hand is at 1 meter from the IR light source, while the background is captured at 1.8 meters. The scaled difference of the skin and the background shows that skin can be segmented reliably with a projection wall as backdrop.

4.2.3 Characterize Hair

Hair is the most difficult part to be recognized as foreground as it does not reflect near-IR light well. The reflection property of different hair colors is very similar as shown in Fig. 9, with a tendency of blonde reflecting slightly better than black hair. Only a slight enhancement

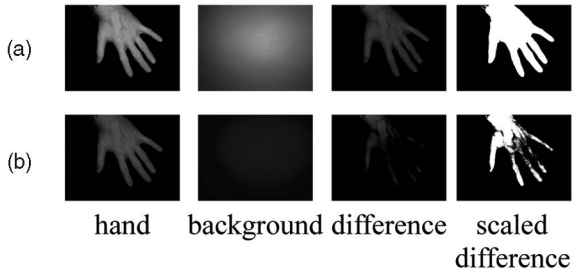


Fig. 8. Segmentation of skin. (a) Hand and Background are at the same distance from the IR light source. (b) Hand is at 1 meter and Background at 1.8 meter from the IR light source.

Fig. 9. Different hair and skin color captured with a color and an IR camera.

can be achieved by applying hair spray or gel as it often enhances the specular component only, which is already strong. We therefore darken the background around the head region instead of trying to brighten up the hair. Using a non IR reflective curtain as backdrop leads to good results for hair segmentation even for black hair (Fig. 9).

5 Stereo Image Generation

Each silhouette is fitted to the texture image using estimated disparity before being applied as a mask leading to a segmented texture image. After masking, we generate projective texturing images to reduce foreshortening.

5.1 Disparity Estimation

We define the centroid of a user in each silhouette image as corresponding points. Knowing the two centroids, we can estimate the depth of the point in 3D space corresponding to the centroid by using the triangulation algorithm described in [25]. After rectification using [6], all epipolar lines become parallel to rows of the image (Fig. 10), which reduces the searching space for corresponding points between the silhouette and the texture image to one dimension.

By approximating the person by a plane and assuming an upright position, the varying depth can be approximated linearly. Consequently, we can calculate disparity between silhouette and texture images with (3), where b is the length of the baseline of two cameras, f is the focal length and z_h is the estimated depth of the user increasing from head to feet.

$$d_h = \frac{b \cdot f}{z_h}. \quad (3)$$

By calculating the disparity for each image pair independently over time, we can observe jittering artifacts due to differences in silhouette extractions, which lead to varying depth estimations. In order to minimize jittering artifacts, but still keep a good texture fitting when the captured person is moving, we apply a Kalman filter to the estimated 3D position.

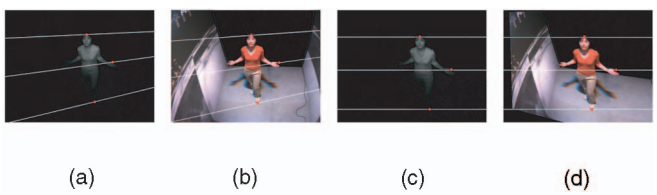


Fig. 10. Epipolar lines for color and texture images (a) and (b) before rectification and (c) and (d) after rectification.

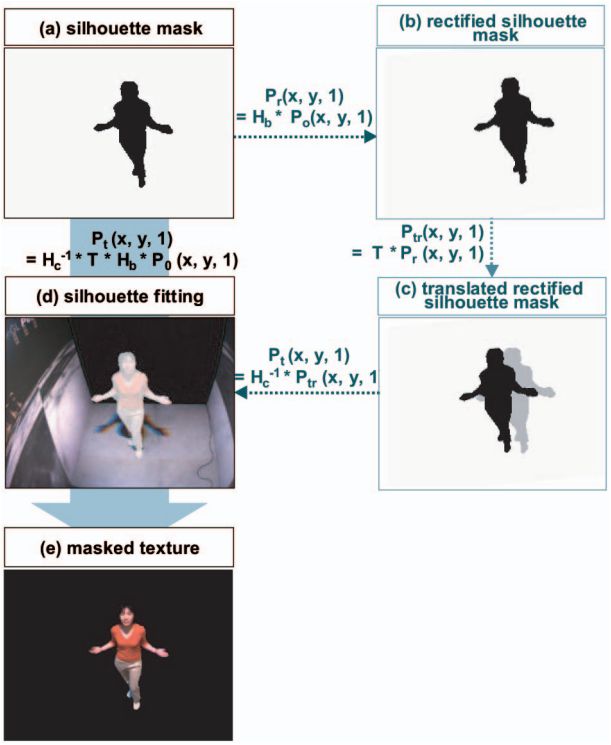


Fig. 11. Silhouette fitting process: $\mathrm{P}_o(\mathrm{x}, \mathrm{y}, 1)$ is a pixel position in the silhouette mask, $\mathrm{P}_c(\mathrm{x}, \mathrm{y}, 1)$ is a pixel position in texture image, H_c, H_b is a homography for the rectification of the texture and silhouette image, respectively.

5.2 Masking

The silhouette image is rectified by the homography H_b, translated by T depending on the disparity d_h and transformed by the inverse homography H_c^{-1} resulting in a mask for the texture image (see Fig. 11). H_c is the homography which would be needed to rectify the texture image.

For this to be a good approximation we make the following assumptions: both cameras lie close to each other, the user is approximated by a plane, and we do not have significant distortion differences between the cameras. Applying the disparity correction to the silhouette leaves the implicit captured depth cues of the texture cameras unchanged.

5.3 Projective Texturing

Due to the positioning of the capturing cameras above the front wall, we get images at an oblique angle, causing foreshortening. Repositioning of the cameras at eye-level is impossible due to occlusions of the screen. Applying projective texturing onto a stereo video billboard allows rendering of the captured user along a vertical parallax and therefore from eye-level while minimizing foreshortening effects (see Fig. 12b).

Simplifying the geometry of the user by a plane yields to some distortion during reprojection. However, it is restricted to areas with high variance of the distance of the true 3D geometry of the user to the plane. We name this distance z_P (see Fig. 13a). The resulting distortion y_P is given as

$$y_P = \frac{y_C \cdot z_P}{z_C}, \tag{4}$$

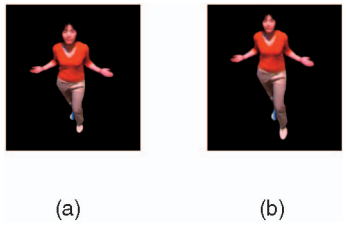


Fig. 12. Projective texturing. (a) Before and (b) after.

with z_C being the distance in z-direction from the camera center to a point P_i on the user and y_C being the distance between the capturing camera C_{Cap} and the rendering camera C_{Ren}. The introduced error is limited to the vertical parallax and has no influence on the horizontal parallax and also no influence on the disparity values as shown in the Section 5.4. Therefore, projective texturing is a valid solution for minimization of foreshortening.

5.4 Disparity Analysis

Omitting a detailed 3D geometry and approximating the user by a plane requires a thorough analysis of the final disparity visible to the user in the CAVE$^{\mathrm{TM}}$-like setup. Since there are only two cameras and no detailed geometry reconstruction, the representation can only be rendered correctly from a vertical parallax including the position of the capturing camera C_{Cap}. In order to assure the quality of the disparity through the whole pipeline for this viewing range, steps potentially having an effect on it will be analyzed in more detail. Masking of the textures, projective texturing and, finally, rendering the stereo billboard in the virtual scene are the three critical steps.

Since masking the textures using the silhouette images is not modifying the disparities but only setting parts of the

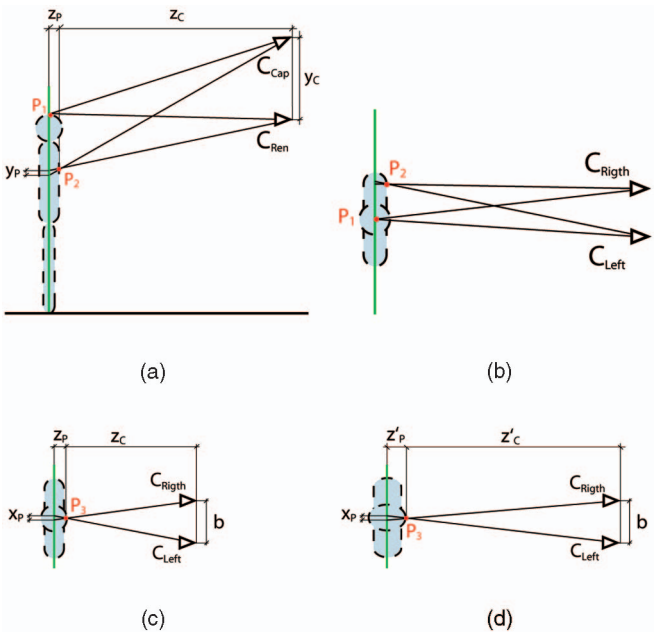


Fig. 13. Different drafts of the setup configuration are depicted (a) from the side and (b) from the top considering reprojection onto the green billboard. In (c) and (d), a rendering setup for two different viewing positions are shown.

TABLE 1
Informal System Evaluation: 3D Appearance (cm)

	Best range	Acceptable distance	No difference
Experts	132-147	92-225	79
Non Experts	141-167	110-205	102

texture image to transparent, the correctness of the disparities is trivial.

Furthermore, the influence of the projective texturing step on the disparity is analyzed. Points of the user lying on the same plane as the billboard (e.g., P_1 in Figs. 13a and 13b) are reprojected to the original place in 3D space. This will lead to the correct disparity independent of camera distance to the billboard. For the points lying distance z_P away from the billboard (e.g., P_2 in Figs. 13a and 13b), the reprojection of the left and right eye images lead to two distinct points. The intersection of the rays from the camera centers of C_{Left} and C_{Right} to the reprojected points lies at a distance z_P from the billboard as long as the cameras lie $z = z_P + z_C$ from the billboard. Note that the height of the position of the camera does not have an influence on z_P, but only on the distortion y_P.

We assume the same baseline b as well as the same focal length f for the cameras for the rendering C_{Ren} as for the cameras in the capturing C_{Cap} setup. Furthermore, the distance of corresponding points on the billboard x_P (see Figs. 13c and 13d) is independent of the movement of the local user during rendering, leading to

$$\frac{x_P}{z_P} = \frac{b}{z_C} \quad and \quad \frac{x_P}{z'_P} = \frac{b}{z'_C}. \tag{5}$$

Therefore, the perceived depth of every point of the remote user z_P is changing proportionally to z, where $z = z_P + z_C$.

Consequently, the remote user appears correct if the distance z_{Cap} from the remote user to the capturing cameras and the distance z_{Ren} from the local user to the billboard are the same. We performed an informal system evaluation with 16 people of whom four are experts with stereoscopic displays and 12 are nonexperts. The collaborating person was captured at a distance of 1.4m. The users were put in a CAVE^{TM}-like system with one mono billboard and one stereo billboard representing the same collaborator. They were asked to choose the best range to see the correct 3D appearance, the distance which was acceptable for a 3D representation, and the distance from where they could not distinguish between the mono billboard and the stereo billboard. Results are shown in Table 1. Some of the users mentioned that the perceived depth does not continuously increase when standing further and further away, but rather leads to the same depth perception as for the mono billboard. This is similar to viewing a real 3D object from far away, where parallax information becomes a weaker depth cue than self-shadowing or occlusion. If a user stands closer than 90cm to the front wall not only will the depth cue given through disparity almost get lost, but she will also be badly captured by the cameras. Therefore this position would make little sense during a collaboration.

This informal evaluation confirms the introduction of an additional depth cue by using the stereo billboard. It would require detailed user studies in order to elaborate which aspects of the stereo billboard are most advantageous for copresence. This will be subject for future work.

TABLE 2
Processing Time (s) per Stereo Image for Different Resolutions

Resolution	*640x480*	*320x240*
Acquisition	0.015	0.014
Background subtraction	0.059	0.013
Masking	0.047	0.012
Projective texture generation	0.047	0.011
JPEG compression	0.055	0.001
Sum	0.223	0.051

6 Integration

The generated stereo images of the video avatar are transmitted to the collaborative party to be integrated into the virtual world. In order to send the images in realtime, we apply a JPEG compression and transfer them as a packet with additional header information using the TCP/IP protocol. The resulting stereo image is rendered onto a billboard to represent the user inside the virtual world. The basic features required for rendering a video stream are provided by the blue-c API, namely, its 2D video service [14]. The original 2D video service was enhanced with stereo capabilities to render different images for both eyes. If the user is moving relative to the billboard, the billboard is rotated in order to be perpendicular to the line between the camera and the billboard center.

Blending the user seamlessly into the virtual environment requires an alpha channel to stencil out the background. In order to avoid the overhead of sending an additional image mask for transparency, we define a color range appearing transparent. As opposed to defining a single matte color, this approach permits the encoding of different alpha values. It has proven to be robust enough against JPEG compression artifacts for our purposes.

7 Results and Applications

In this section, first, we will show a performance analysis for the generation of video avatars and their transmission over the network, second, we will discuss the texture quality, and third, we will present results and possible applications.

7.1 Performance

For the performance analysis we consider two parts. One is the processing time of the captured images and the other is the transmission of those images over the network. For the performance analysis of the rendering part, we refer to [14]. By the term "frame," we are in fact referring to two images, one per eye.

Included in the processing time of the captured images are all the steps from the acquisition to the projective texture generation (see Table 2). Parallel processing or a multi computer setup have not yet been exploited, even though the system would lend itself naturally to such a performance optimization.

Processed images have to be transmitted over a network to the other collaborating party, where bandwidth limitations can cause a bottleneck. During our experiments, we sent video avatars generated at ETH Zürich in Switzerland over an intercontinental link to a CAVE^{TM}-like system at

Fig. 14. For reasons of quality verification all the snapshots have been taken in the mono mode of our environment. (a) and (b) High quality stereo video avatar composition in a Chess application. (c) Realtime mirroring of user in running CAVE^{TM}-like system. (d) Collaboration example of two users over an intercontinental link.

Ewha Womans University in Korea. We measured a total available network bandwidth of up to 161.25KB/sec leading to a transmission time of 0.094s for a 640×480 frame. Considering the processing time is 2.4 times longer, the bottleneck does not lie in the transmission of the data.

Doing the processing and the transmission of the images in parallel lets us update video avatars in the virtual world in realtime (19fps for 320×240). When increasing the resolution to 640×480 the framerate drops to 4.5fps.

7.2 Texture Quality

The visual quality of a fixed viewpoint stereo video billboard depends on the texture quality and the correct disparity treatment due to the lack of a precise geometry. Two steps in the pipeline can cause a decrease of texture quality, namely projective texturing and JPEG compression. Projective texturing can introduce distortion depending on the variance in distance from the true 3D geometry of the user to the billboard as described in Section 5.3. This problem limits the possibility of using our representation for collaboration, where exact pointing gestures are required. However, the high frequency details of the texture are mostly preserved since there is no blending between different viewpoints, therefore, a high-quality texture can be generated. In the future, further improvements could be achieved by substituting the JPEG compression by MPEG4.

7.3 Visual Results and Application

Our acquisition of a stereo video avatar is very fast since no stereo-algorithm, visual hull or any other geometrical reconstruction is necessary. Despite our simple representation the ability to perceive collaborators in 3D under the constraints mentioned in Section 5.4 is possible due to stereo video images on the one hand, and due to high texture quality maintaining additional visual depth cues (e.g., self-shadowing) on the other. Additionally, the captured texture quality facilitates nonverbal communication cues. The system can be integrated in an up to six sided CAVE^{TM}-like environment without encountering problems of realtime segmentation. Therefore, the representation can be seamlessly blended into a virtual collaboration environment.

In Fig. 14 and Fig. 15, we integrated our system in a Chess application from [14] for test purposes and experienced realtime collaboration over an intercontinental link. In the future, applications like teleconferencing, virtual architecture guidance, collaborative planning, and even collaborative games could be envisaged. This lightweight and cheap extension to virtual environments, delivering high texture quality at fast frame rates, is ideal for many collaborative environments even beyond CAVE^{TM}-like systems.

Fig. 15. A sequence of snapshots from our running collaboration system showing visions of possible future work. (a), (b) and (c) Direct collaboration between two parties. (d) User mirrored into virtual world captured in stereo mode. (e) and (f) Future collaboration environment, where more than one person per location can join a meeting in a virtual world.

8 Conclusion and Future Work

We created a simple and cheap (US $6,000) setup based on infrared and texture cameras, which does not rely on calibration-critical optical components such as semitransparent mirrors thanks to a texture warping algorithm that matches the images from different cameras.

Our approach works very well for face to face collaboration settings and also more generally for collaborations allowing a fixed viewing range of the collaborating party, since a per pixel disparity appearing correct to the user (see Section 5.4) is equivalent to per pixel depth information. As soon as the viewpoint of the user changes along a horizontal parallax, our system lacks geometry or additional cameras to which the viewpoint could be switched. However, we achieved real time collaboration with high texture quality and accurate depth along the vertical parallax through the captured position, and limited depth accuracy when moving towards or from the object. Furthermore, our system performs very well even if the user is surrounded by a running CAVE^{TM}-like environment. These criteria appeared to be more important for achieving an increased feeling of copresence in a face to face collaboration than an accurate geometrical representation of the user, which is cumbersome and very expensive to achieve in a running immersive environment [8]. In the future, detailed user studies will be necessary in order to elaborate which aspects of our representation have most effect on the feeling of copresence.

Currently, our hardware installation is limited to a single stereo acquisition pair and, therefore, to a single viewing direction. Adding additional cameras around the user would allow switching between streams for additional viewing directions. The IR-based segmentation mask would also enable us to perform a full 3D object reconstruction, similar to Matusik et al.'s polygonal visual hull or the blue-c 3D video fragment system. Future work will also include optimization of the IR lighting to increase the robustness of the segmentation. In particular, the IR diffusion properties of projection screens could be further exploited to get a uniform dark background, which would make a characterization of hair and cloth redundant. Our work could be extended to multiple users per collaboration environment like shown in Fig. 15e and Fig. 15f, which would further enhance group collaborations. In order to provide a complete collaboration environment, we will integrate audio capabilities in future versions.

Acknowledgments

The authors would like to thank Tim Weyrich, Miguel A. Otaduy, Daniel Cotting, and Filip Sadlo for constructive discussions influencing their work and Hyeryung Kwak, Juhaye Kim, and Ji-Hyun Suh for generating final images. This work was supported in part by the Korean Ministry of Information and Communication (MIC) under the Information Technology Research Center (ITRC) Program and in part by the Korea Science and Engineering Foundation (KOSEF) under the International Research Internship Program.

REFERENCES

[1] J. Allard, E. Boyer, J.-S. Franco, G. Raffin, and C. Menier, "Marker-Less Real Time 3D Modeling for Virtual Reality," *Immersive Projection Technology,* 2004.

[2] L. Chong, H. Kaczmarski, and C. Goudeseune, "Video Stereoscopic Avatars for the Cave Virtual Environment," *Proc. IEEE Int'l Symp. Mixed and Augmented Reality,* 2004.

[3] C. Cruz-Neira, D.J. Sandin, and T.A. DeFanti, "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the Cave," *Proc. 20th Ann. Conf. Computer Graphics and Interactive Techniques,* pp. 135-142, 1993.

[4] J.W. Davis and A.F. Bobick, "Sideshow: A Silhouette-Based Interactive Dual-Screen Environment," Technical Report 457, Mass. Inst. of Technology (MIT) Media Lab, 1998.

[5] P. Debevec, A. Wenger, C. Tchou, A. Gardner, J. Waese, and T. Hawkins, "A Lighting Reproduction Approach to Live-Action Compositing," *ACM Trans. Graphics,* vol. 21, no. 3, pp. 547-556, 2002.

[6] A. Fusiello, E. Trucco, and A. Verri, "A Compact Algorithm for Rectification of Stereo Pairs," *Machine Vision and Applications,* vol. 12, no. 1, pp. 16-22, 2000.

[7] S.J. Gibbs, C. Arapis, and C.J. Breitender, "Teleport—Towards Immersive Copresence," *Multimedia Systems,* vol. 7, no. 3, pp. 214-221, 1999.

[8] M. Gross, S. Wuermlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. K-Meier, T. Svoboda, L. Gool, S. Lang, K. Strehlke, A.V. Moere, and O. Staadt, "Blue-C: A Spatially Immersive Display and 3D Video Portal for Telepresence," *ACM Trans. Graphics,* vol. 22, no. 3, pp. 819-827, 2003.

[9] A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 2, pp. 150-162, Feb. 1994.

[10] S.-Y. Lee, I.-J. Kim, S.C. Ahn, H. Ko, M.-T. Lim, and H.-G. Kim, "Real Time 3D Avatar for Interactive Mixed Reality," *Proc. ACM SIGGRAPH Int'l Conf. Virtual Reality Continuum and its Applications in Industry (VRCAI '04),* pp. 75-80, 2004.

[11] W. Matusik, "Image-Based Visual Hulls," master's thesis, Mass. Inst. of Technology, 2001.

[12] W. Matusik, C. Buehler, and L. McMillan, "Polyhedral Visual Hulls for Real-Time Rendering," *Proc. 12th Eurographics Workshop Rendering,* pp. 115-125, 2001.

[13] W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan, "Image-Based Visual Hulls," *Proc. SIGGRAPH '00,* pp. 369-374, 2000.

[14] M. Naef, O. Staadt, and M. Gross, "Blue-C API: A Multimedia and 3D Video Enhanced Toolkit for Collaborative VR and Telepresence," *Proc. ACM SIGGRAPH Int'l Conf. Virtual Reality Continuum and Its Applications in Industry (VRCAI '04),* pp. 11-18, 2004.

[15] P.J. Narayana, P. Rander, and T. Kanade, "Constructing Virtual Worlds Using Dense Stereo," *Proc. Int'l Conf. Computer Vision (ICCV '98),* pp. 3-10, 1998.

[16] T. Ogi, T. Yamada, M. Hirose, M. Fujita, and K. Kuzuu, "High Presence Remote Presentation in the Shared Immersive Virtual World," *Proc. IEEE Virtual Reality Conf. (VR '03),* 2003.

[17] T. Ogi, T. Yamada, Y. Kurita, Y. Hattori, and M. Hirose, "Usage of Video Avatar Technology for Immersive Communication," *Proc. First Int'l Workshop Language Understanding and Agents for Real World Interaction,* 2003.

[18] T. Ogi, T. Yamada, K. Tamagawa, M. Kano, and M. Hirose, "Immersive Telecommunication Using Stereo Video Avatar," *Proc. Virtual Reality 2001 Conf. (VR '01),* 2001.

[19] S. Pollard and S. Hayes, "View Synthesis by Edge Transfer with Application to the Generation of Immersive Video Objects," *Proc. ACM Symp. Virtual Reality Software and Technology,* pp. 91-98, 1998.

[20] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays," *Proc. SIGGRAPH '98,* pp. 179-188, 1998.

[21] A. Sadagic, H. Towles, J. Lanier, H. Fuchs, A. Van Dam, K. Daniilidis, J. Mulligan, L. Holden, and B. Zeleznik, "National Tele-Immersion Initiative: Towards Compelling Tele-Immersive Collaborative Environments," *Proc. Medicine Meets Virtual Reality Conf.,* 2001.

[22] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder, "Small Group Behaviour in a Virtual and Real Environment: A Comparative Study," *Presence: Teleoperators and Virtual Environments,* vol. 9, no. 1, pp. 37-51, 2000.

[23] Y. Suh, D. Hong, and W. Woo, "2.5D Video Avatar Augmentation for VR Photo," *Proc. International Conf. Artificial Reality and Telexistence (ICAT '02),* 2002.

[24] S. Wuermlin, E. Lamboray, O. Staadt, and M. Gross, "3D Video Recorder: A System for Recording and Playing Free-Viewpoint Video," *Computer Graphics Forum,* vol. 22, no. 2, pp. 181-193, 2003.

[25] G. Taubin, "Camera Model and Triangulation," Note for EE-148: 3D Photography, CalTech, 2001.

[26] S. Subramanian, V. Rajan, D. Keenan, A. Johnson, D. Sandin, and T. DeFanti, "A Realistic Video Avatar System for Networked Virtual Environments," *Proc. Seventh Ann. Immersive Projection Technology Symp. (IPT '02),* 2002.

[27] K. Yasuda, T. Naemura, and H. Harashima, "Thermo-Key: Human Region Segmentation from Video," *IEEE Computer Graphics and Application,* vol. 24, no. 1, pp. 26-30, Jan./Feb. 2004.

Seon-Min Rhee received the BS and MS degrees in computer science and engineering from Ewha Womans University at Seoul, Korea in 1999 and 2001, respectively. She is currently a PhD student in the Department of Computer Science and Engineering, Ewha Womans University. Her research interests are 3D video and human computer interaction.

Remo Ziegler received the MSc degree in computer science in 2000 from the Swiss Federal Institute of Technology ETHZ in Zürich, Switzerland, and worked during the following four years as a research assistant at Mitsubishi Electric Research Laboratories MERL in Boston, Massachusetts, and NICTA, a national research Laboratory in Sydney, Australia. He recently started working on his PhD degree at the Computer Graphics Laboratory at ETHZ. His main interests lie in 3D scanning, reconstruction, and projective technology. He is student member of the IEEE and a member of the ACM.

Jiyoung Park received the BS and MS degrees in computer science and engineering from Ewha Womans University, Seoul, Korea in 2002 and 2004, respectively. She is currently a PhD student in the Department of Computer Science and Engineering, Ewha Womans University. Her research interests include computer graphics, virtual reality, and medical image processing.

Martin Naef received the MSc degree in computer science from the Swiss Federal Institute of Technology, Zürich (ETHZ), and the PhD degree from the Computer Graphics Laboratory at ETHZ for his work on the blue-c Application Programming Interface. He is currently working at the Digital Design Studio of the Glasgow School of Art in Scotland. He is a member of the IEEE, the IEEE Computer Society, and the ACM.

Markus Gross received the MS degree in electrical and computer engineering and the PhD degree in computer graphics and image analysis, both from the University of Saarbrucken, Germany. He is a professor of computer science and director of the Computer Graphics Laboratory of the Swiss Federal Institute of Technology (ETH) in Zürich. From 1990 to 1994, Dr. Gross worked for the Computer Graphics Center in Darmstadt, where he established and directed the Visual Computing Group. His research interests include point-based graphics, physics-based modelling, multiresolution analysis, and virtual reality. He has been widely publishing and lecturing on computer graphics and scientific visualization, and he authored the book *Visual Computing* (Springer, 1994). Dr. Gross has taught courses at major graphics conferences including ACM SIGGRAPH, IEEE Visualization, and Eurographics. He is the associate editor of the *IEEE Computer Graphics and Applications* and has served as a member of international program committees of many graphics conferences. Dr. Gross has been a papers cochair of the IEEE Visualization 1999, the Eurographics 2000, and the IEEE Visualization 2002 Conferences. He chaired the papers committee of ACM SIGGRAPH 2005. He is a senior member of the IEEE and the IEEE Computer Society.

Myoung-Hee Kim received the BA degree from Ewha Womans University, Seoul, Korea in 1974, the MS degree from Seoul National University in 1979, and the PhD degree from Gottingen University, Germany in 1986. She is currently a professor of computer science and engineering and director of the Visual Computing and Virtual Reality Laboratory and director of the Center for Computer Graphics and Virtual Reality at Ewha Womans University. She has served as a conference cochair of Pacific Graphics (2004) and Israel-Korea Binational Conference on Geometrical Modeling and Computer Graphics. She is the President of Korea Computer Graphics Society (KCGS) and the Vice President of Korea Society for Simulation (KSS). Her research interests include computer graphics, virtual reality, computer simulation, and medical image visual computing. She is a member of the IEEE.