Animatronic Shader Lamps Avatars

Peter Lincoln Greg Welch Andrew Nashel Andrei State Adrian Ilie Henry Fuchs

The University *of* North Carolina *at* Chapel Hill
Department of Computer Science*

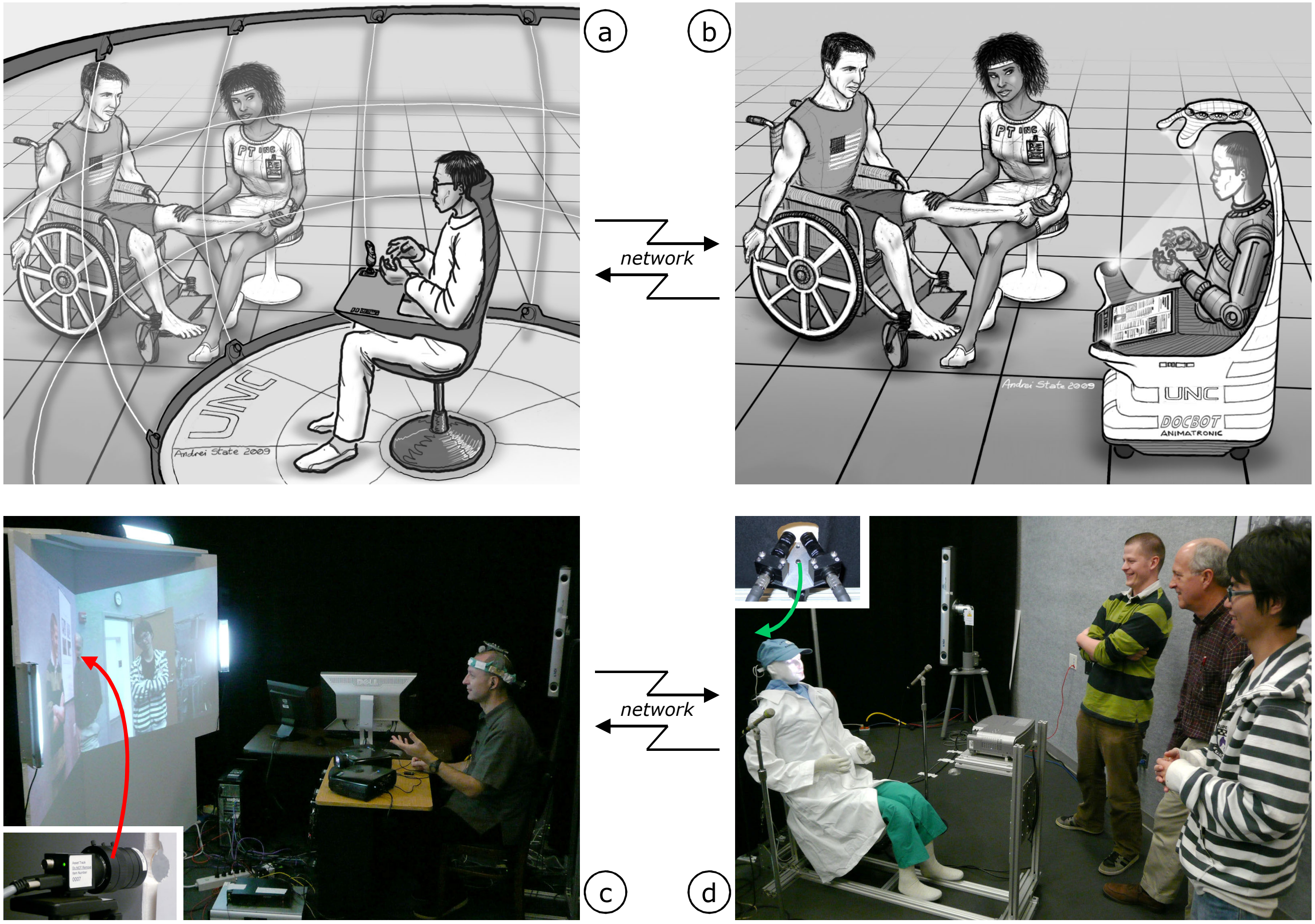

Figure 1: The upper images conceptually illustrate one possible use of animatronic Shader Lamps Avatars (SLA): full-duplex telepresence for medical consultation. The physician in (a) interacts with a remote patient and therapist in (b) by means of a camera-equipped SLA. The SLA allows the physician to *see* and be *seen* by the patient and therapist. The lower two figures show our current bi-directional proof-of-concept prototype. The user in (c) wears a tracking system and is imaged by a video camera (inset and red arrow). In (d) we show the avatar of the user, consisting of a Styrofoam head mounted on a pan-tilt unit and illuminated by a projector. The setup in (c) also includes a two-projector panoramic view of the avatar site, acquired by two co-located cameras mounted above the Styrofoam head in (d) (inset and green arrow)

Abstract

Applications such as telepresence and training involve the display of real or synthetic humans to multiple viewers. When attempting to render the humans with conventional displays, non-verbal cues such as head pose, gaze direction, body posture, and facial expression are difficult to convey correctly to all viewers. In addition, a framed image of a human conveys only a limited physical sense of presence—primarily through the display's location. While progress continues on articulated robots that mimic humans, the focus has been on the motion and behavior of the robots rather than on their appearance.

*e-mail: {plincoln, welch, nashel, andrei, adyilie, fuchs}@cs.unc.edu, phone: 919-962-1700, fax: 919-962-1799

We introduce a new approach for robotic avatars of real people: the use of cameras and projectors to capture and map both the dynamic motion and the appearance of a real person onto a humanoid animatronic model. We call these devices *animatronic Shader Lamps Avatars* (SLA). We present a proof-of-concept prototype comprised of a camera, a tracking system, a digital projector, and a life-sized Styrofoam head mounted on a pan-tilt unit. The system captures imagery of a moving, talking user and maps the appearance and motion onto the animatronic SLA, delivering a dynamic, real-time representation of the user to multiple viewers.

Index Terms: H.4.3 [Information Systems Applications]: Communications Applications—Computer conferencing, teleconferencing, and videoconferencing H.5.1 [Multimedia Information Systems]: Animations—Artificial, augmented, and virtual realities I.3.7 [Computer Graphics]: Three Dimensional Graphics and Realism—Virtual Reality; I.3.8 [Computer Graphics]: Applications;

1 Introduction

The term "telepresence" describes technologies that enable activities as diverse as remote manipulation, communication, and collaboration. Today it is a moniker embraced by companies building commercial video teleconferencing systems and by researchers exploring immersive collaboration between one or more participants at multiple sites. In a collaborative telepresence system, each user needs some way to perceive remote sites, and in turn be perceived by participants at those sites. In this paper we focus primarily on the latter challenge—how a user is *seen* by remote participants, as opposed to how he or she *sees* the remote participants.

There are numerous approaches to visually simulating the presence of a remote person. The most common is to use 2D video imagery; however, such imagery lacks a number of spatial and perceptual cues, especially when presented on static displays. If the user gazes into the camera, then all participants think the user is looking at them individually; if instead the user gazes elsewhere, no one thinks the user is gazing at them, but each may think the user is gazing at a neighboring participant. These 2D displays can be augmented with pan-tilt units in order to provide some amount of gaze awareness [24, 15]; however, the same shared eye gaze issue continues to apply as in the static case. Even with 3D captured or rendered imagery and 3D or view-dependent displays, it is difficult to convey information such as body posture and gaze direction to multiple viewers. Such information can single out the intended recipient of a statement, convey interest or attention (or lack thereof), and direct facial expressions and other non-verbal communication. To convey that information to specific individuals, each participant must see the remote person from his or her own viewpoint.

1.1 Providing Distinct Views

Providing distinct, view-dependent imagery of a person to multiple observers poses several challenges. One approach to is to provide separate tracked and multiplexed views to each observer, such that the remote person appears in one common location. However, approaches involving head-worn displays or stereo glasses are usually unacceptable, given the importance of eye contact between all (local and remote) participants.

Another approach is to use *multi-view* displays. These displays can be realized with various technologies and approaches, however each has limitations that restrict its utility:

- "Personal" (per-user) projectors combined with retroreflective surfaces at the locations corresponding to the remote users [21, 22]. Advantages: arbitrary placement of distinct viewing zones. Limitations: awkward to achieve stereo; each projector needs to remain physically very close to its observer.
- Wide-angle lenticular sheets placed over conventional displays to assign a subset of the display pixels to each observer [17, 27]. Advantages: lateral multi-view with or without stereo. Limitations: difficult to separate distinct images; noticeable blurring between views; fixed viewing positions; approach sometimes trades limited range of stereo for a wider range of individual views.
- High-speed projectors combined with spinning mirrors used to create 360-degree light field displays [14, 13]. Advantages: lateral multi-view with stereo. Limitations: small physical size due to spinning mechanism; binary/few colors due to dividing the imagery over 360 degrees; no appropriate image change as viewer moves head vertically or radially.

1.2 Eye Contact

Eye contact is an essential ingredient of human interaction [3] and as such merits special attention in teleconferencing applications. Conventional teleconferencing systems based on video cameras and video displays generally do not offer eye contact due to the inherent difficulty of physically co-locating the display showing the remote participant(s) and the camera(s) capturing imagery of the local participants. High-end products such as Cisco Telepresence [33] alleviate this problem through a display-camera setup that keeps the distance between the acquisition camera and the screen location showing the remote participant's eyes at a minimum. Other solutions include optical beam splitters that virtually co-locate camera and display [34], and even automatic, real-time manipulation of remote users' video images, aiming to re-orient the remote user's eyes and face towards the camera [5]. The addition of stereoscopy and/or head tracking further increases the complexity of such approaches.

Our approach (Figure 1) makes the approach inherently asymmetric: while the human participants can obviously look the SLA in the eyes, the SLA can only appear to be making eye contact with those participants if correctly matched imagery acquired from the SLA's point of view is displayed at the SLA user's location. "Correctly matched" implies imagery that is presented to the SLA user in such a way that when the user looks at a distant human participant's image—whether by directly facing that participant's image or merely out of the corner of an eye—the SLA user's head and eye poses are remapped onto the SLA such as to recreate at the distant location the geometry of eye contact [30] between the SLA and the targeted human participant. Furthermore, "correctly matched" also requires that the imagery for the SLA user be acquired from the points of view of the SLA's eyes. One way to accomplish this is to mount miniature video cameras within the SLA's eyes. While we do not do that (yet), we developed a preliminary approximate approach, described in Section 3.2.

1.3 Shader Lamps Avatars (Overview)

The approach we describe here is to use cameras and projectors to capture and map both the dynamic motion and the appearance of a real person onto a human-shaped display surface. We call these devices animatronic *Shader Lamps Avatars* (SLA) [18]. The approach intrinsically provides depth cues, distinct views, and improved gaze cues. This one-to-many approach also scales to any number of observers, who do not need to be head-tracked. To convey appearance, we capture live video imagery of a person, warp the imagery and use Shader Lamps techniques [4, 25, 26] to project it onto the human-shaped display surface. As a result, all observers view the remote user from their own perspectives. To convey motion and orientation we track the user and use animatronics to update the pose of the display surface accordingly, while continually projecting matching imagery.

A fundamental limitation of this approach is that it does not result in a general-purpose display—it is a *person* display. More general multi-view displays [13, 17] can—and often are—used to display artifacts like coffee cups and pieces of paper along with the remote person. However, to use such displays for multi-viewer teleconferencing, one needs either many cameras (one per view) or real-time 3D reconstruction.

This paper presents an implemented prototype Animatronic SLA telepresence system. This implemented system is one step along a path towards a fully usable and flexible system. Figure 1 shows conceptual sketches and real results from our current proof-of-concept prototype. Our method and prototype are described in detail in Sections 3 and 4. In Section 5 we present results, followed by details of our experience with a public demonstration of the system in Section 6, and in Section 7 we conclude with thoughts on the current state of our work and discuss future possibilities.

2 Related Work

There has been prior work related to our SLA ideas. These works include both commercialized and academics systems, which are each composed of projective surfaces, animatronic objects, tactile surfaces, cameras, and/or synthetic sources. The relevant works are organized by major categories below.

2.1 3D-Surface Projective Systems

Fixed-surface projective systems includes those consisting of moving or static fixed-shape surfaces and projectors that provide an appearance for that surface. Some of the most visible work in projective avatars has been in theme park entertainment, which has been making use of projectively illuminated puppets for many years. The early concepts consisted of rigid statue-like devices with external film-based projection, examples of which include the head busts at the *Haunted Mansion* ride at Disneyland. More recent systems include animatronic devices with internal (rear) projection such as the animatronic Buzz Lightyear that greets guests as they enter the *Buzz Lightyear Space Ranger Spin* attraction in the Walt Disney World Magic Kingdom. While our current SLA prototype uses front projection, using similar internal projection would reduce the overall footprint, making it less intrusive and potentially more practical.

In the academic realm, *Shader Lamps*, introduced by Raskar et al. [26], use projected imagery to illuminate physical objects, dynamically changing their appearance. In this system, the virtual and physical objects have the same shape. The authors demonstrated changing surface characteristics such as texture and specular reflectance, as well as dynamic lighting conditions, simulating cast shadows that change with the time of day. The concept was extended to *Dynamic Shader Lamps* [4], whose projected imagery can be interactively modified, allowing users to paint synthetic surface characteristics on physical objects. Shader Lamps-illuminated objects have the main advantage in that they can be viewed by multiple unencumbered participants in an accurate manner on all surfaces covered by the projected imagery. Our prototype makes significant use of Shader Lamps techniques.

Hypermask [35] is a system that dynamically synthesizes views of a talking, expressive character, based on voice and keypad input from an actor wearing a mask onto which the synthesized views are projected. While aimed at storytelling and theatrical performances, it deals with many of the issues we discuss here as well, such as the construction of 3D models of human heads and projecting dynamic face imagery onto a moving object (in this case, the mask). Unlike Shader Lamps, however, the projection surface differs from the projected object, which can distort the appearance and perceived shape when viewed off-angle.

2.2 Animatronic Systems

There are many humanoid animatronic systems in production or in existence as research systems. These systems typically take on a singular fixed identity. Future versions of the technology we introduce here will require complex humanoid animatronics (robots) as "display carriers," which can be passive (projectively illuminated, as shown here) or active (covered with flexible, self-illuminated display surfaces such as the ones currently under development in research labs at Philips, Sony and others) to support switching between multiple users' appearances.

Significant work in the area of humanoid robots is being conducted in research labs in Japan. In addition to the well-known Honda ASIMO robot [9], which looks like a fully suited and helmeted astronaut with child-like proportions, Shuuji Kajita at Japan's National Institute of Advanced Industrial Science and Technology has recently demonstrated a robot with the proportions and weight of an adult female, capable of human-like gait and equipped with an expressive human-like face [2]. Other researchers have focused on the subtle, continuous body movements that help portray lifelike appearance, on facial movement, on convincing speech delivery, and on response to touch. Work led by Hiroshi Ishiguro [12] at Osaka University's Intelligent Robotics Laboratory stands out, in particular the lifelike *Repliee* android series [7] and the *Geminoid* device. They are highly detailed animatronic units equipped with numerous actuators and designed to appear as human-like as possible, thanks to skin-embedded sensors that induce a realistic response to touch. The *Geminoid* is a replica of Hiroshi Ishiguro himself, complete with facial skin folds, moving eyes, and implanted hair—yet still not at the level of detail of the "hyper-realistic" sculptures and life castings of (sculptor) John De Andrea [6], which induce a tremendous sense of realism despite their rigidity. Geminoid is teleoperated, and can thus take the Ishiguro's place in interactions with remote participants, much like the technology we advocate here. While each of these systems can take on a *single* human's appearance to varying degrees of realism, they are limited in their flexibility in who can legitimately teleoperate the system.

On the other hand, the Takanishi Laboratory's WD-2 robot [16] is capable of changing shape to produce multiple expressions and identities. The WD-2 also uses rear-projection to texture a real user's face onto the robot's display surface. The robot's creators are interested in behavioral issues and plan to investigate topics in human-geminoid interaction and sense of presence. The flexibility in appearances of which the WD-2 is capable would make it quite useful for a telepresence system, as it could theoretically take on the shape of its user. Unfortunately, in its current state, the shape-changing apparatus is too bulky for use as a head atop a mobile body. However, one can anticipate the eventual miniaturization of the equipment, making this a potentially useful addition to an SLA.

When building animatronic avatars, one is inevitably faced with the challenge of mapping human motion to the animatronic avatar's motion. The avatar's range of motion, as well as its acceleration and speed characteristics, will generally differ from a human's; with current state-of-the art in animatronics, they are a subset of human capabilities. Hence one has to "squeeze" the human motion into the avatar's available capabilities envelope, while striving to maintain the appearance and meaning of gestures and body language, as well as the overall perception of resemblance to the imaged person. In the case of our current prototype, we are for now concerned with the mapping of head movements; previous work has addressed the issue of motion mapping ("retargeting") as applied to synthetic puppets. Shin et al. [29] describe on-line determination of the importance of measured motion, with the goal of deciding to what extent it should be mapped to the puppet. The authors use an inverse kinematics solver to calculate the retargeted motion. They also introduce filtering techniques for noisy input data (not an issue with our current tracker, but possibly with alternative, tetherless vision-based methods). Their work is geared towards complete figures, not just a single joint element as in our prototype, but their methods could be applied to our system as well.

The TELESAR 2 project led by Susumu Tachi [32, 31] integrates animatronic avatars with the display of a person. In contrast to the

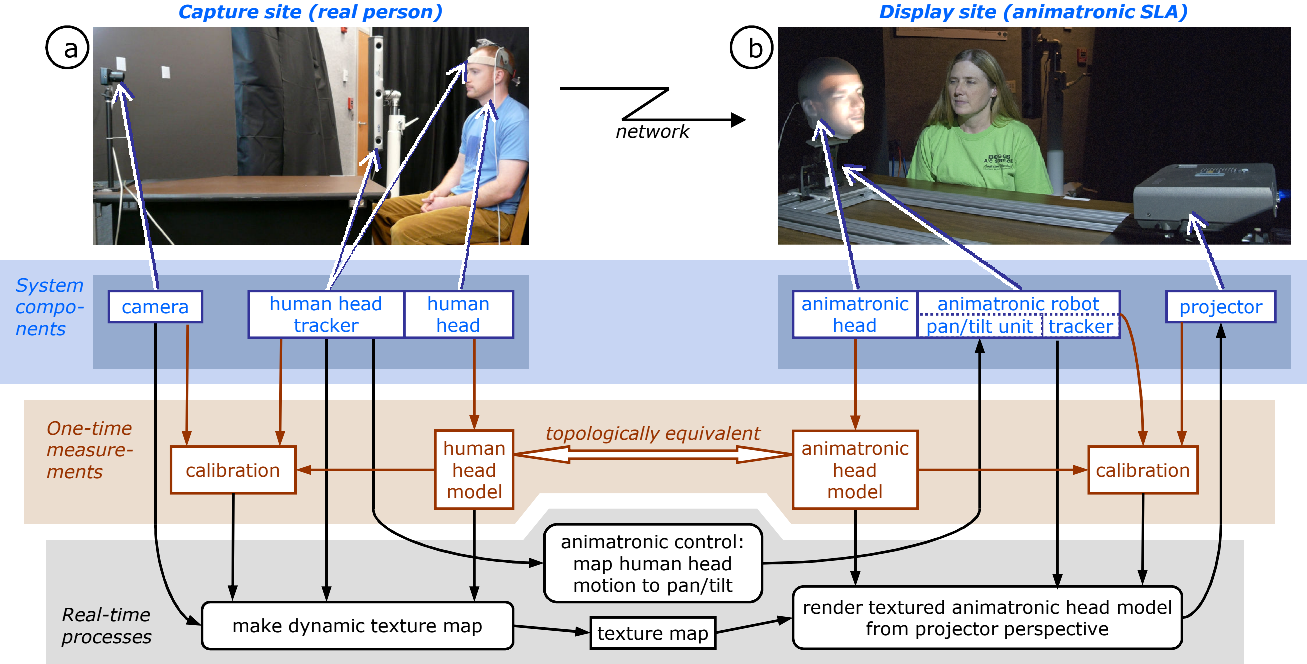

Figure 2: Proof-of-concept implementation and diagram of our Shader Lamps Avatar (SLA) user capture and display. At the *capture site* shown in (a), a camera captures a person, also tracked using a headband. At the *display site* shown in (b), a projector displays images onto an avatar consisting of a Styrofoam head placed on an animatronic robot. The diagram in the lower part of the figure highlights the system components and the processes involved.

other work in this subsection, the robot-mounted display surfaces do not mimic human face or body shapes; the three-dimensional appearance of the human is recreated through stereoscopic projection. The researchers created a roughly humanoid robot equipped with remote manipulators as arms, and retro-reflective surfaces on face and torso, onto which imagery of the person "inhabiting" the robot is projected. The retro-reflective surfaces and the multiple projectors enable multiple fixed viewing positions with distinct views of the user. However, a very large number of projectors would be required to provide a full 360°view for participants. The robot also contains cameras; it is controlled by a human from a remote station equipped with multi-degree-of-freedom controls and monitors displaying imagery acquired by the robot's cameras. The work is part of an extensive project that aims to enable users to experience "telexistence" in any environment, including environments that are not accessible to humans.

3 Design

In this section we describe the overall design of our proof-of-concept system. The system is composed of two main functions and corresponding channels: the capture and presentation of the avatar's user, and the capture and presentation of the avatar's site.

3.1 User Capture and Presentation

The components of our proof-of-concept system, as shown in Figure 2, are grouped at two sites: the *capture site* and the *display site*. The capture site is where images and motion of a human subject are captured. In addition to a designated place for the human subject, it includes a camera and a tracker, with a tracker target (a headband) placed onto the human's head, as shown in Figure 3 (a). We currently use a single 1024 × 768 1/3" CCD color camera running at 15 FPS for capturing imagery. The focus, depth of field, and field of view of the camera have been optimized to allow the subject to comfortably move around in a fixed chair. The NDI Optotrak system is currently used for tracking. Future systems may choose to employ vision-based tracking, obviating the need for a separate tracker and allowing human motion to be captured without cumbersome user-worn targets.

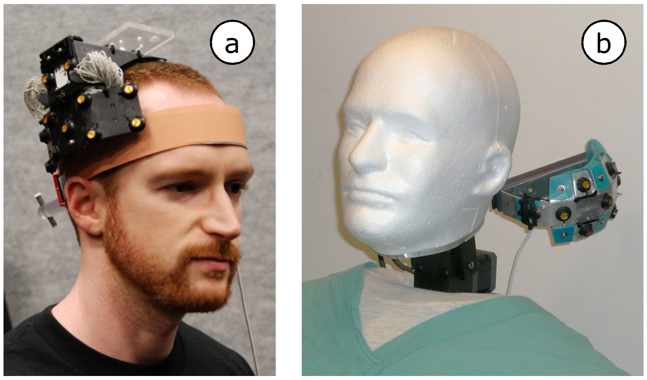

Figure 3: Active IR-LED tracker targets. (a) Headband tracker placed on a human head. (b) Tracker tool attached to the back of the avatar's head, which is mounted on a pan-tilt unit, shown in its reference pose (zero pan and tilt).

The display site includes a projector, the avatar, and a tracker with a tracker target mounted on the avatar as shown in Figure 3 (b). The avatar consists of an Styrofoam head that serves as the projection surface. The avatar head is a generic commercially available male Styrofoam head. The avatar is mounted on a pan-tilt unit that allows moving the head to mimic the movements of the human at the capture site. The pan-tilt unit in use is capable of rotating at 300°per second; however, in order to ensure smooth motion,

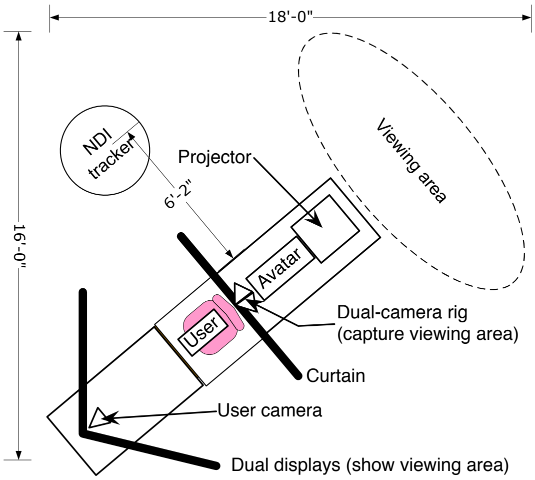

Figure 4: Full-duplex configuration of the prototype system. The back-to-back setup was chosen to primarily to suit the capabilities of the NDI tracker while both presenting the avatar to the viewers, and allowing the viewers to step to the side to see the human user/inhabiter.

the speed is limited to 100°per second. This head and pan-tilt unit are mounted above a dressed torso with fixed arms and legs. The 1024×768 60Hz DLP projector is mounted approximately 1 meter in front of the avatar and is configured to only project upon the visual extent, including range of motion, of the mounted avatar; the projector's focus and depth of field are sufficient to cover the illuminated half of the avatar. Instead of a tracker, future systems may choose to use position-reporting features of more sophisticated pan-tilt units in order to derive the pose of the Styrofoam head.

3.2 Site Capture and Presentation

We initially developed our prototype system with capture and display sites colocated within our lab (see Figure 4). In order to progress towards a realistic full-duplex tele-conferencing system (our main application focus), we incorporated all image and sound transmission paths needed for the two sites to operate at a large distance from one another. As shown in Figure 1 (c), the capture site is equipped with a panoramic dual-projector setup; the two projectors are connected to a dual-camera rig mounted just above the avatar's head at the display site (d). The fields of view of the camera rig and of the projection setup are matched, aligning the gaze directions of the human user at the capture site and of the avatar at the remote site. That is, if the human user turns his or her head to face a person appearing 15 degrees to the right on the projective display, the slaved avatar head will also turn by 15 degrees to directly face that same person. This allows for approximately correct gaze at both sites in the horizontal direction—the SLA's gaze towards the remote participants at the display site, and the remote participants' gazes (appearing in the panoramic dual-projector imagery) towards the human user at the capture site.

To achieve correct SLA gaze in the vertical direction we first ensure that the SLA's eyes appear to have the correct vertical elevation when the human user is looking at a remote participant's image at the capture site. We achieve this by vertically adjusting the projected panoramic imagery, which provides the human user's visual target at the capture site. The dual cameras at the display site however, which are used to capture the remote participants, are mounted *above* the SLA head, and therefore the remote people appear to be looking down when shown at the capture site, even if they are gazing at the SLA. An optimized future design could make use of cameras mounted within the avatar's eye location (as mentioned in Section 1.2), or re-orient remote participant's eyes and/or faces through image manipulation methods [5].

The second subsystem required for full-duplex operation consists of a set of audio components for sound transmission. The display site is equipped with two stereo microphones that pick up ambient sound and conversation, amplified and transmitted into ear buds for the capture site user. That user wears a lapel microphone, whose amplified signal is transmitted to a single speaker located close to the avatar's head at the display site. Together with the core elements described above, these additional components turn our experimental system into a rudimentary yet full-fledged SLA telepresence prototype.

4 Method

In this section we explain the methods we employ in our proof-of-concept system. We begin by describing one-time operations such as calibration and model construction. We continue with the adjustments performed before each run and finish by describing the real-time processes that take place during the use of the system.

4.1 One-time Operations

One-time operations are performed when the system components are installed. They include camera and projector calibration, as well as head model construction and calibration.

4.1.1 Camera and Projector Calibration

To calibrate the intrinsic and extrinsic parameters of the camera at the capture site, we use a custom application [11] built on top of the OpenCV [23] library. Our custom application, in order to compute the camera's intrinsic parameters, makes use of the standard OpenCV camera calibration procedure, which processes a set of images containing checkerboards of known physical sizes. As a slight variant on the standard techniques, in order to ensure that the computed extrinsic parameters are in the same space as the tracker's coordinate frame, we use a probe to capture the 3D points of one of the fixed checkerboard positions and use those points as the input to the extrinsic parameters calibration of the OpenCV library. In the case of our system, these techniques result in a reprojection error on the order of a pixel or less.

We calibrate the projector at the display site using a similar process. Instead of capturing images of the checkerboard pattern, we place the physical checkerboard at various poses inside the projector's field of view, and use our custom application to render and manually adjust the size and location of a virtual pattern until it matches the physical pattern. By using these virtual patterns and another set of tracker probe positions as input to our custom calibration application, we produce the projector's intrinsic and extrinsic parameters, the latter in the tracker's coordinate space.

4.1.2 Head Model Construction

We built our human and avatar 3D head models using FaceWorx [19], an application that uses two images of a person's head (front and side view) and manual identification of distinctive features such as eyes, nose and mouth to produce a textured 3D model. The process consists of importing a front and a side picture of the head to be modeled and adjusting the position of a number of given control points overlaid on top of each image—see Figure 5 (a,e). The program provides real-time feedback by displaying the resulting 3D model as shown in Figure 5 (b,f). A key property of all FaceWorx models is that they have the same topology, only the vertex positions differ. This allows a straightforward mapping from one head model to another. In particular, we can render the texture of a model onto the shape of another. In Figure 5, the projection-ready model (i) is obtained using the shape from the avatar head (h) and the texture from the human head (c).

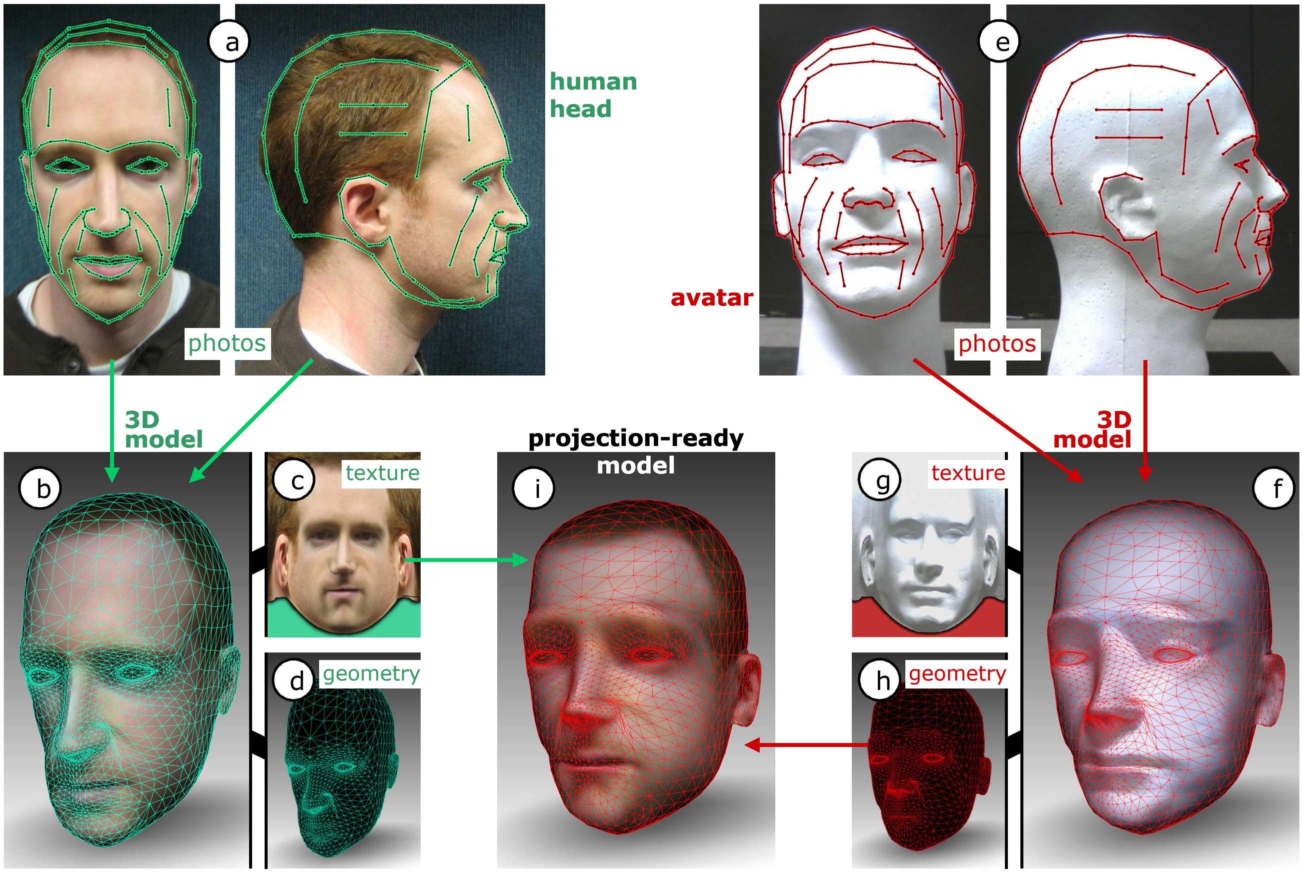

Figure 5: Head model construction and mapping. FaceWorx [19] is used to move control points in photographs showing the fronts and sides of heads (a,e), resulting in 3D models (b,f), which are comprised of texture (c,g) and geometry (d,h). The final model (i) is built using the texture of the human head (c) and the geometry of the avatar head (h).

4.1.3 Head Model Calibration

Capturing the human head model and rendering the animatronic head model on the Styrofoam projection surface requires finding their poses in the coordinate frames of the trackers at each site. Both the human's and the avatar's heads are assumed to have static shape, which simplifies the calibration process. The same procedure can be used for both the human's and avatar's heads. The first step in this calibration is to find the relative pose of each head model with respect to a *reference coordinate frame* which corresponds to a physical tracker target rigidly attached to each head being modeled. We use a tracker probe to capture about 4 or 5 3D points corresponding to salient face features on each head, and compute the offsets between each captured 3D point and the 6D pose of the reference coordinate frame. Next, we use a custom GUI to manually associate each computed offset to a corresponding 3D vertex in the FaceWorx model. We then run an automatic optimization process to compute the 4×4 homogeneous transformation matrix that minimizes the error in the mapping between the 3D point offsets and the corresponding 3D vertices in the FaceWorx model. The calibration transformation matrices obtained through the optimization process are constrained to be orthonormal. This transformation represents the relative pose and scale of the model with respect to the reference coordinate frame. At run-time we multiply the computed matrix by the matrix that characterizes the pose of the reference coordinate frame in the tracker's coordinate frame to obtain the complete live transformation. The quality of the calibration matrix can be qualitatively evaluated by running the system and is more dependent on the accuracy of the model than the accuracy of the probed positions.

4.2 Per-run Calibrations

The headband used to track the human head is assumed to be rigidly mounted onto the head. Alas, each time the user dons the headband, the pose (position and orientation) is slightly different. Although a complete calibration prior to each run would ensure the best results, in practice small manual adjustments are sufficient to satisfy the above assumption. Only two small adjustments are required for each run of the system.

The first adjustment consists of aligning the poses of the pan-tilt unit and of the human head. We ask the human to rotate his or her head and look straight at the camera, and capture a *reference pose*. We set this pose to correspond to the zero pan and zero tilt pose of the pan-tilt unit—see Figure 3 (b), which positions the Styrofoam head as if it were directly facing the projector. Given the physical calibration of the human user's viewing area displays (see Figure 4), this ensures that the human user's gaze matches the avatar's gaze.

The second small adjustment is required only if the user has removed the headband between head calibration and system execution. We perform additional manual adjustments to the headband by asking the user to rotate and shift the headband to ensure that the projections of salient face features in the projected image are aligned with the corresponding features on the animatronic head; these features include the positions of the eyes, tip of the nose, and edges of the mouth. In essence, these shifting operations restore the headband to its originally calibrated position on the human's head. Realigning the pan-tilt and human poses one more time restores the gaze alignment and completes the per-run calibrations.

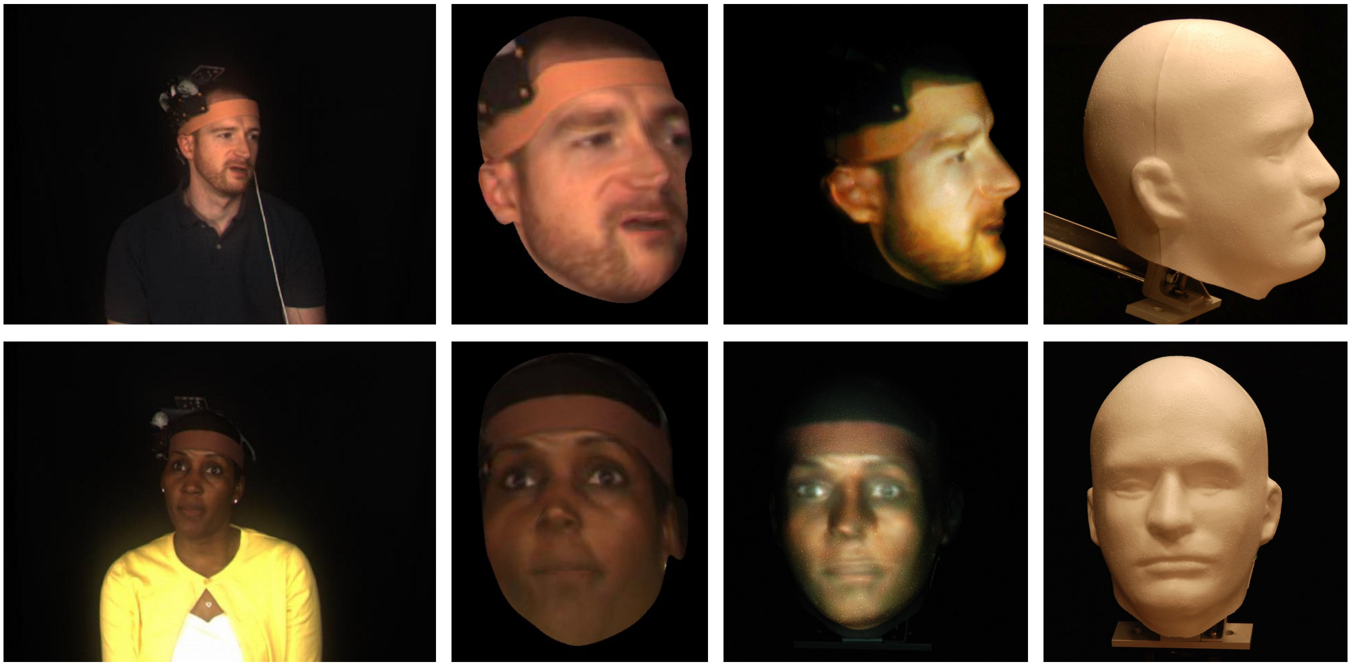

Figure 6: Humans and avatars as seen from different viewpoints. Column 1 shows the live camera images; column 2 shows the warped head models; column 3 shows photos of the models projected onto the avatar; column 4 shows the un-illuminated Styrofoam head in poses matching the column 3 images. In row 1, the photos in columns 3 and 4 are taken from the left side of the projector; in row 2, these photos are taken from behind the projector.

4.3 Real-time Processes

Once the system is calibrated, it becomes possible for the avatar on the display side to mimic the appearance and motion of the person on the capture side. In this section we describe the real-time processes that implement this function.

4.3.1 Animatronic Control

Given a pose for a human head tracked in real time and a reference pose captured as described in Section 4.2, it is possible to compute a relative orientation. This orientation constitutes the basis for the animatronic control signals for the avatar. The pose gathered from the headband is a 4×4 orthonormal matrix consisting of rotations and translations from the tracker's origin. We use a decomposition of the rotation components of the matrix to compute the roll, pitch, and yaw of the human head. The relative pitch and yaw of the tracked human are mapped to the pan and tilt capabilities of the pan-tilt unit and transformed into commands issued to the pan-tilt unit. Using this process, the avatar emulates a subset of the head *motions* of its human "master;" roll and translation motions are discarded.

4.3.2 Dynamic Texturing

Given a calibrated input camera, a tracked human, and a calibrated 3D model of the human's head, we compute a texture map for the human head model. This is achieved through *texture projection*—the imagery of the camera is mathematically mapped to the surface of the human head model, as though the camera were a digital projector and the head the projection surface. To map the texture onto the avatar's head model, which is a different shape, some processing is required. We use custom OpenGL vertex and pixel shaders to render a live textured model of the human or avatar head in real time from any point of view on a standard display.

In the case of the physical avatar, however, it is desirable to compute a texture map using the calibrated model of the human head and project the resulting live imagery onto the calibrated model of the avatar head. Although the two heads have different shapes, both heads are modeled in FaceWorx and thus have the same topology. That similar topology enables us to perform the warping operation shown in Figure 5 to transform the texture projection to target the avatar's head. Through OpenGL vertex and pixel shaders, it is possible to perform this warp entirely on the GPU. Essentially these shaders perform texture projection with one major difference: we use the vertex coordinates and pose of the tracked and calibrated human head model for computing texture look-up coordinates, and we use the vertex coordinates and pose of the tracked and calibrated avatar head model for computing the location to draw the head. Given an arbitrary projection matrix, it is possible to render a textured model of the avatar from any perspective, using a live texture from camera imagery of the human head. By selecting the perspective of the calibrated projector, the live texture is projected upon the tracked animatronic head, and the model shape is morphed to that of the animatronic head model. Using this process, the animatronic head emulates the *appearance* of its human counterpart.

5 Results

The overall result of the system is the presentation of a physical proxy for a live human. Currently the avatar can present elements of a user's facial appearance and head motion. See Figure 6.

Visual appearance is generated through the use of a single camera and single projector and thus is limited to certain perspectives. In particular, high-quality imagery is limited to the front of the face. Surfaces not facing the camera or projectors, such as the top or sides of the head, are not well covered when the user is facing the camera or when the avatar is facing the projector. As in-person communication is generally face-to-face, it is reasonable to focus visual attention onto this component. Since the human's facial features are mapped to the avatar's corresponding features by taking advantage of the identical topology of their 3D models, the avatar can present the human's eyes, nose, mouth, and ears in structurally appropriate positions. The quality of this matching is demonstrated in Figure 6. As both relationships (camera/human and projector/avatar) are approximately the same in terms of direction, the imagery is generally appropriate, and the features well matched. As the user moves, the tracker and the camera imagery update correspondingly to project the proper texture on the virtual model of the head, thus maintaining proper eye contact from the target participant's perspective and good gaze awareness from the other participants' perspectives.

Using the pan-tilt unit, the avatar is also capable of movement that matches the yaw and pitch components of the human's head motion, within the limits of the pan-tilt unit and tracker. Because the human's features are texture-mapped to the corresponding locations of the avatar, all observers at the display site can both see a representation of the avatar's user and accurately assess the direction the user is looking. However, humans are capable of accelerating faster than the available pan-tilt unit's configured maximum speed of 100°/sec. This limiting factor and the pan-tilt unit's response delay can result in the avatar's head motion lagging behind the most recently reported camera imagery and corresponding tracker position. Deliberate head motions, such as gazing, nodding, or indicating no, can be matched, and mismatched orientations between the human and avatar for a given camera frame can be handled by the rendering algorithm. Unfortunately, extremely fast periodic head motions can result in truncated amplitude. It is possible that this lag issue could be mitigated by a more responsive pan-tilt unit, good-quality predictive filtering on the expected pan-tilt unit's motions, or a higher-level intended-behavior analysis of the human's motion. Motions that go beyond panning or tilting, such as cocking one's head or stretching one's neck would require a motion platform with additional degrees of freedom.

Fortunately, the capture and playback sides of the system can be decoupled—the motion of the avatar need not match that of the human user to show appropriate imagery. Because the texture produced by the input camera is displayed on the avatar via projective texturing of an intermediate 3D model, the pose of the avatar is independent of the human's pose. The image projected on the avatar is dependent on the avatar's model and the current pose of the pan-tilt unit. As such, the motion of the avatar can be disabled or overridden and the facial characteristics of human and avatar would still match to the best degree possible. However, if the relative orientations of human and camera at the capture site and of avatar and projector at the display site are significantly different, the quality of the projective texture may be degraded due to missing visual information. For example, if the person looks significantly to one side, away from the capture camera, and the avatar faces the projector, then part of the projected surface cannot be seen by the camera and can result in incorrect imagery. This issue could resolved with additional cameras and/or projectors that would capture and/or project with better coverage of the two heads.

6 Demonstration at ISMAR 2009

On October 19–20, 2009, we demonstrated the full-duplex prototype SLA system at the 2009 International Symposium on Mixed and Augmented Reality (ISMAR 2009) in Orlando, FL. As described in Section 3.2 and illustrated in Figure 4, the capture and display sites were set up in a back-to-back configuration, separated by a large curtain. As a result, the capture site was not directly visible to casual visitors, who were thus interacting primarily with the SLA on the display side. The visitors could, however, step to the side to look behind the curtain and see the human inhabiter.

We demonstrated the system for a total of three hours on two separate days. On the first day, the SLA was inhabited for approximately two hours by co-author Henry Fuchs, someone we expected to be visibly recognizable to many of the visitors. For the second day, we hired a professional comedian (Brian Bradley) to inhabit the SLA. The idea was to try someone who was less likely to be visibly recognizable, but was skilled at personal interactions in a public setting, and likely to be engaging (humorous) in doing so.

Neither person had any significant experience "in the avatar" before, and both had to get used to the system and its restrictions (e.g., head motion limits), which they did quickly. Both inhabiters managed to engage many walk-up visitors in exchanges that ranged from a few seconds to several minutes, at times with lively back-and-forth talking. One exchange between the avatar of professional comedian (Brian Bradley) and some visitors is given below.

> **Visitor:** [a bit indecipherable, but apparently a comment about not being a real human]
> **SLA:** Ha ha, wow, [rolling head while laughing] you're not exactly the Avon lady yourself! [nodding toward the visitor] You have dark secrets in your bag I'm sure. [nodding affirmatively]
> **Visitor:** You're a little creepy. [looking around the sides of the SLA]
> **SLA:** [shaking head] I'm not creepy! [looking at visitor] I'm very nice.
> **SLA:** [looking up at another visitor] What's your name?
> **Visitor:** Karen.
> **SLA:** Hi Karen. See-more here. Hi Ladies! [looking around and nodding]
> **Visitors:** Hi.
> **SLA:** How are you? [lifting and tilting head toward another group of visitors—Karen follows the SLA gaze]

A subsequent exchange was as follows.

> **SLA:** What I hear from Karen is that I'm creepy! [looking around at three visitors]
> **Visitor:** [visitors laugh]
> **SLA:** Uh, well [looking around]—a little can—just a few—uh—a couple molecules of creepy is enough to give me self-esteem issues. [looking downward sadly]

As was the case in the above exchange, several of the conversations involved more than one visitor, requiring the human user (and hence the SLA) to alternately look at one visitor, then at the other as the human user was addressing each visitor in turn. We observed that as the SLA was changing gaze direction in this way, the visitors appeared to naturally follow its gaze, and assess who among the bystanders had become the SLA's new eye contact partner. Following someone else's gaze in this way is a natural group interaction behavior [8] and we were encouraged that our SLA and the full-duplex setup appeared to support it.

We also noticed apparent emotional connections with the SLA. For example, one visitor made a joking comment about how his (the visitor's) chest hurt, asking whether the "doctor" (the SLA) could tell him what was wrong. The SLA (comedian), looking at the visitor, responded that the primary cause was likely related to the visitor's sweater, which (the comedian said) went out of style about 20 years ago. The visitor in turn looked down at the sweater, and walked away with a bit of a dejected look. As in other exchanges, nearby people were looking back and forth between the SLA and the visitor. In this particular case, when the SLA made the "out of style" comment about the visitor's sweater, other nearby visitors looked back at the SLA making comments questioning the nature of the insult, and offering verbal sympathy for the visitor.

Most visitors commented on the SLA's appearance in some way. Some reacted in a quizzical fashion, realizing that the avatar was not real, and yet seemed intrigued by its presence. Some commented that the avatar was "a little eerie," and some appeared reluctant to interact with it, for whatever reason. (Some people would normally be reluctant to interact face-to-face with a *real human* comedian in a public setting, for example if they were embarrassed.) On the other hand, many visitors appeared to fully engage their own bodies, using head motion, changing body position and posture, and hand gestures that seemed as natural as if the SLA had been a real

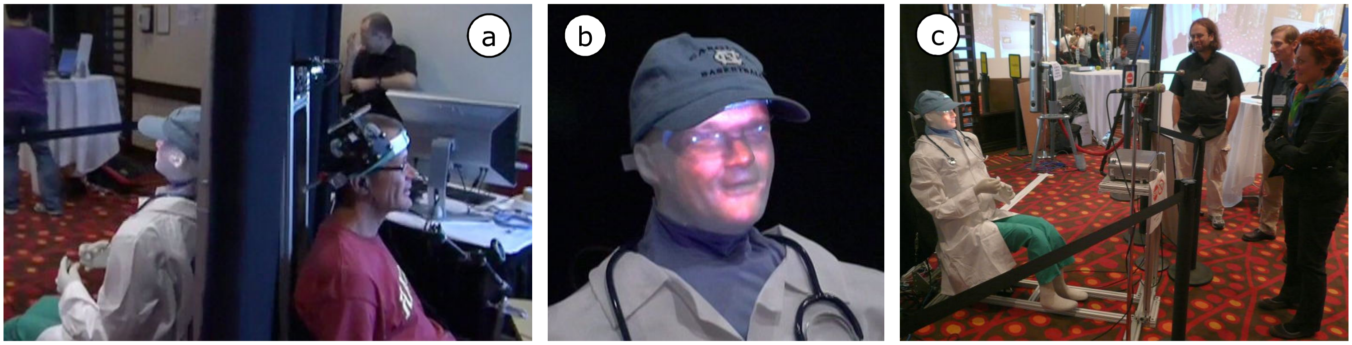

Figure 7: Experimental setup of the prototype system as presented at ISMAR 2009. (a) shows the SLA and the professional comedian (Brian Bradley) back-to-back—the comedian's head is optically tracked and his appearance is captured by a camera, (b) shows a closeup of the SLA with the comedian's dynamic appearance, and (c) attendees conversing with the comedian's by means of the SLA. See also Figure 4.

person in front of them. Some would reach out and point to specific parts of the SLA body, asking for example "Can you move your hands?" In the future it would be interesting to compare such SLA interactions with the same using a 2D video of the inhabiter.

Some of the visitors initially thought the avatar behavior was synthetic (an automated character) until we encouraged them to talk to it. Naturally, the conversations with the researcher focused more on technology, whereas the interactions with the comedian were driven by jokes. Some visitors used the terms "uncanny" as well as "uncanny valley," with the latter obviously referring to the notion that an avatar (any synthetic artifact) that has some human-like features, but not quite human behavior may, at some point, begin to appear uncanny even as its creators strive to make the features and behavior more realistic [20]. Nevertheless, all of the "uncanny valley" quoters proceeded to engage the avatar without reserve.

Overall we were encouraged by what we saw during this opportunity. It seems that the overall approach shows promise for the telepresence application it was conceived for.

7 Conclusions and Future Work

We introduced animatronic Shader Lamps Avatars (SLAs), described a proof-of-concept prototype system, and presented preliminary results. We are currently exploring passive vision-based methods for tracking the real person's head [1, 10, 28] so that we can eliminate the separate tracking system. We also hope to add additional cameras and projectors. Both will involve the dynamic blending of imagery: as the real person moves, textures from multiple cameras will have to be dynamically blended and mapped onto the graphical model, and as the physical avatar moves, the projector imagery will have to be dynamically blended (intensity and color) as it is projected. We are also considering methods for internal projection. In terms of the robotics, we will be exploring possibilities for more sophisticated animation, and more rigorous motion retargeting methods [29] to address the limitations of the animatronic components (range and speed of motion, degrees of freedom) while still attempting human-like performance. Some of the filtering techniques in [29] could be useful if we use vision-based face tracking as mentioned. We are also exploring possible avatar head shapes in terms of the acceptability of a generic head compared to a copy of the user's head, or some canonical average head. Finally, together with collaborators at the Naval Postgraduate School, we are undertaking a series of human subject evaluations related to gaze.

While our current prototype supports only rudimentary full-duplex communications by means of the dual camera/projector setup described above, we envision a generous full-duplex capability via the use of multiple cameras associated with the SLA and a seamless surround display associated with the user. For example, outward-facing cameras could be mounted in a canopy over the SLA to provide remote imagery for the user as depicted in Figure 1 (b) and (a) respectively. If these outward-facing cameras are mounted close to the head, then the vertical disparity between where the participants are looking, namely the avatar's eyes, and the avatar user's viewpoint would be minimized, helping maintain good eye contact for the avatar's user. The optimal location for full two-way eye contact would place the capture cameras inside of the avatar's eyes. However, given that the avatar's head moves, one would have to remap the camera imagery back to its geometrically correct location on display surface at the avatar user's location. Figure 8 shows a preliminary demonstration of a panoramic camera and a surround display that could be used for viewing the avatar's surroundings. Figure 8 also illustrates the asymmetric one-to-many nature of the telepresence avatar paradigm.

We are working on an SLA mounted on a mobile platform, with outward-facing cameras. We envision an inhabiter exploring remote facilities such as hospitals, factories and shopping centers, while interacting with multiple remote individuals—both *seeing* and being *seen*. For some disabled individuals, this could provide a "prosthetic presence" that is otherwise unattainable. SLAs may also be useful as role players in immersive training environments for medicine and defense, robotic teachers that visually transform between historians and historic individuals, or personal robotic companions that take on different real or synthetic appearances during live interactions. In fact SLAs could some day support the limited integration of a virtual "second life" into our "first lives"—allowing people to visit remote real places, using a real or alternate persona, as if they (or their persona) were really there.

Acknowledgements

We thank Herman Towles for his insightful suggestions, technical help, and advice. John Thomas provided mechanical and electronic engineering assistance. David Harrison set up our full-duplex audio subsystem. Dorothy Turner became our first non-author SLA user (Figure 5, bottom half of image set). Tao Li helped set up the ISMAR demonstration. Donna Boggs modeled as the avatar's interlocutor (Figure 2). We thank Chris Macedonia, M.D. for inspiring us by expressing his desire to visit his patients in remote hospitals and other medical facilities with a greater effectiveness than is possible with current remote presence systems, and for offering the term "prosthetic presence." We are grateful to comedian Brian Bradley for his appearance as a prosthetic physician at our ISMAR 2009 booth, and we thank all ISMAR participants who visited our booth and engaged both the avatar and the researchers with questions and suggestions. This work was supported in part by the Office of Naval Research (award N00014-09-1-0813, "3D Display and Capture of Humans for Live-Virtual Training," Dr. Roy Stripling, Program Manager), the National Science Foundation (Grant No. CNS-0751187), and the Intelligence Advanced Research Projects Activity (Cooperative Agreement No. FA8750-08-2-0209).

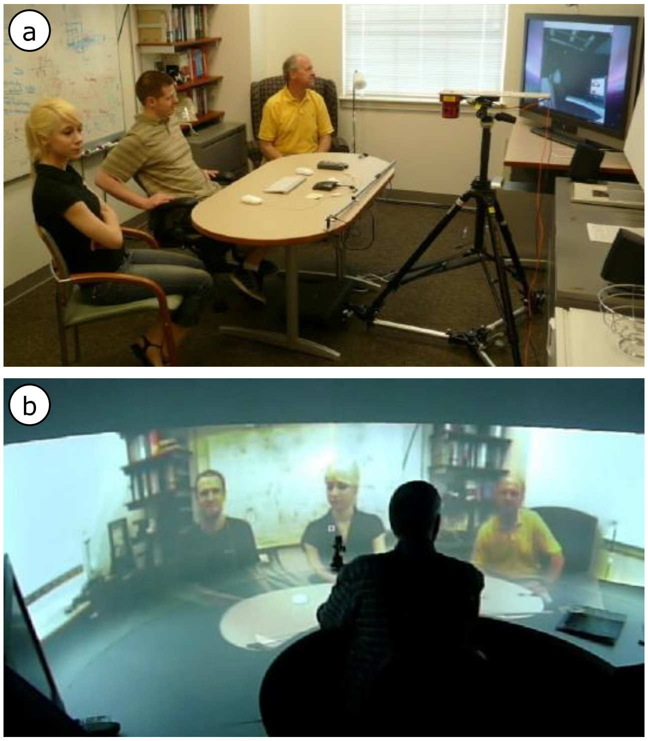

Figure 8: Mock-up of remote panoramic video for avatar control. A PointGrey Ladybug camera is used to capture panoramic imagery of a remote scene in (a). The real-time panoramic video is mapped to a projector-based 270° surround display as shown in (b). The Ladybug would eventually be mounted above the SLA.

References

[1] J. Ahlberg and R. Forchheimer. Face tracking for model-based coding and face animation. *International Journal of Imaging Systems and Technology*, 13(1):8–22, 2003.

[2] AIST. Successful development of a robot with appearance and performance similar to humans. http://www.aist.go.jp/aist_e/latest_research/2009/20090513/20090513.html, September 2010.

[3] M. Argyle and M. Cook. *Gaze and mutual gaze / Michael Argyle and Mark Cook*. Cambridge University Press, Cambridge, Eng. ; New York :, 1976.

[4] D. Bandyopadhyay, R. Raskar, and H. Fuchs. Dynamic shader lamps: Painting on real objects. In *Proc. IEEE and ACM international Symposium on Augmented Reality (ISAR '01)*, pages 207–216, New York, NY, USA, October 2001. IEEE Computer Society.

[5] A. Criminisi, J. Shotton, A. Blake, and P. Torr. Gaze manipulation for one-to-one teleconferencing. *Computer Vision, IEEE International Conference on*, 1:191, 2003.

[6] J. L. DeAndrea. AskART. http://en.wikipedia.org/wiki/John_De_Andrea, September 2010.

[7] R. Epstein. My date with a robot. *Scientific American Mind*, June/July:68–73, 2006.

[8] J. K. Hietanen. Does your gaze direction and head orientation shift my visual attention? *Neuroreport*, 10(16):3443–3447, 1999.

[9] Honda Motor Co., Ltd. Honda Worldwide - ASIMO. http://world.honda.com/ASIMO/, May 2009.

[10] T. S. Huang and H. Tao. Visual face tracking and its application to 3d model-based video coding. In *Picture Coding Symposium*, pages 57–60, 2001.

[11] A. Ilie. Camera and projector calibrator. http://www.cs.unc.edu/~adyilie/Research/CameraCalibrator/, May 2009.

[12] H. Ishiguro. Intelligent Robotics Laboratory, Osaka University. http://www.is.sys.es.osaka-u.ac.jp/research/index.en.html, May 2009.

[13] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. In *SIGGRAPH '09: ACM SIGGRAPH 2009 papers*, pages 1–8, New York, NY, USA, 2009. ACM.

[14] A. Jones, I. McDowall, H. Yamada, M. Bolas, and P. Debevec. Rendering for an interactive 360° light field display. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, volume 26, pages 40–1 – 40–10, New York, NY, USA, 2007. ACM.

[15] D. Kerse, H. Regenbrecht, and M. Purvis. Telepresence and user-initiated control. In *Proceedings of the 2005 international conference on Augmented tele-existence*, page 240. ACM, 2005.

[16] T. Laboratory. Various face shape expression robot. http://www.takanishi.mech.waseda.ac.jp/top/research/docomo/index.htm, August 2009.

[17] P. Lincoln, A. Nashel, A. Ilie, H. Towles, G. Welch, and H. Fuchs. Multi-view lenticular display for group teleconferencing. *Immerscom*, 2009.

[18] P. Lincoln, G. Welch, A. Nashel, A. Ilie, A. State, and H. Fuchs. Animatronic Shader Lamps Avatars. In *Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 27–33. IEEE Computer Society, 2009.

[19] LOOXIS GmbH. FaceWorx. http://www.looxis.com/en/k75.Downloads_Bits-and-Bytes-to-download.htm, February 2009.

[20] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[21] D. Nguyen and J. Canny. Multiview: spatially faithful group video conferencing. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 799–808, New York, NY, USA, 2005. ACM.

[22] D. T. Nguyen and J. Canny. Multiview: improving trust in group video conferencing through spatial faithfulness. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1465–1474, New York, NY, USA, 2007. ACM.

[23] OpenCV. The OpenCV library. http://sourceforge.net/projects/opencvlibrary/, May 2009.

[24] E. Paulos and J. Canny. Social tele-embodiment: Understanding presence. *Autonomous Robots*, 11(1):87–95, 2001.

[25] R. Raskar, G. Welch, and W.-C. Chen. Table-top spatially-augmented reality: Bringing physical models to life with projected imagery. In *IWAR '99: Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, page 64, Washington, DC, USA, 1999. IEEE Computer Society.

[26] R. Raskar, G. Welch, K.-L. Low, and D. Bandyopadhyay. Shader lamps: Animating real objects with image-based illumination. In *Eurographics Workshop on Rendering*, June 2001.

[27] O. Schreer, I. Feldmann, N. Atzpadin, P. Eisert, P. Kauff, and H. Belt. 3DPresence-A System Concept for Multi-User and Multi-Party Immersive 3D Videoconferencing. pages 1–8. CVMP 2008, Nov. 2008.

[28] Seeing Machines. faceAPI. http://www.seeingmachines.com/product/faceapi/, May 2009.

[29] H. J. Shin, J. Lee, S. Y. Shin, and M. Gleicher. Computer puppetry: An importance-based approach. *ACM Trans. Graph.*, 20(2):67–94, 2001.

[30] A. State. Exact eye contact with virtual humans. In *ICCV-HCI*, pages 138–145, 2007.

[31] S. Tachi. http://projects.tachilab.org/telesar2/, May 2009.

[32] S. Tachi, N. Kawakami, M. Inami, and Y. Zaitsu. Mutual telexistence system using retro-reflective projection technology. *International Journal of Humanoid Robotics*, 1(1):45–64, 2004.

[33] Wikipedia. Cisco telepresence. http://en.wikipedia.org/wiki/Cisco_TelePresence, April 2010.

[34] C. Woodworth, G. Golden, and R. Gitlin. An integrated multimedia terminal for teleconferencing. In *Global Telecommunications Conference, 1993, including a Communications Theory Mini-Conference. Technical Program Conference Record, IEEE in Houston. GLOBECOM '93., IEEE*, pages 399 –405 vol.1, nov-2 dec 1993.

[35] T. Yotsukura, F. Nielsen, K. Binsted, S. Morishima, and C. S. Pinhanez. Hypermask: Talking head projected onto real object. *The Visual Computer*, 18(2):111–120, April 2002.