Integrating Communication with Interaction:
Computer Vision Challenges for Interactive and Intelligent Environments

Jeremy R. Cooperstock
Centre for Intelligent Machines
McGill University
Montreal, QC, H3A 2A7
Canada

Abstract

Interactive, Intelligent Environments involve a convergence of various research themes, including high-fidelity visualization, communication, gestural expression, and virtualized reality systems. Recent advances in real-time acquisition, transmission, and rendering of multimodal data (e.g. audio, video, haptic) allow for the synthesis of significantly improved perceptual representations of a virtual or real (e.g. remote) environment than were previously possible. Furthermore, increased computational power permits the synthesis of a rich responsive media space that responds to a large number of participants engaged in a complex, expressive activity. Unfortunately, current systems tend to concentrate almost exclusively on one aspect or the other, supporting the representation and interaction with a virtual world, or supporting distributed human communication, but never both. The ideal interactive intelligent environment is one that permits effective distributed human-human communication among large numbers of participants at multiple locations, simultaneously with data visualization capabilities and interaction with dynamic, synthetic objects. A significant challenge for the next generation of such environments is to develop the necessary physical infrastructures and software architectures that combine these capabilities appropriately.

1: Introduction

Intelligent, interactive environments (IIE) typically employ video, sound, and possibly kinetics, in their reaction to live human activity. The human-computer interaction paradigm is that of local participants whose interaction is augmented by rich computational media. In contrast, Shared Reality [1] deals more with the issues of human-human interaction. It employs techniques that connect physically separated rooms so that geographically distributed people can exchange high-fidelity media with minimal delay: high-resolution video, multichannel audio, and vibrosensory (e.g. floor vibration) data. While each such space offers its respective strengths, we argue that the ideal IIE depends on supporting not only human communication with computer-generated artifacts, but fundamentally, on computer-mediated interaction between people. To achieve this goal relies on the combination of these approaches, fusing the sensory interpretation of responsive media with the synthesis and communication technologies of shared environments for distributed participants.

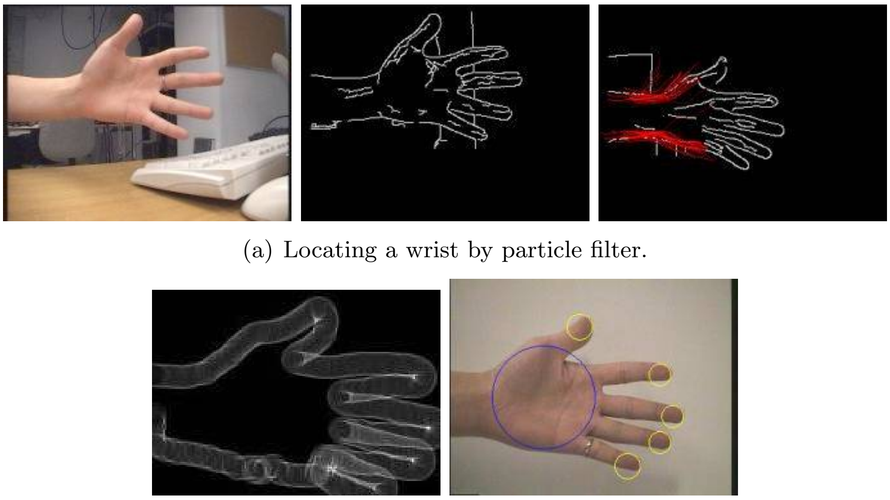
The remainder of this papers surveys a variety of technologies related to both human-computer and human-human interaction in the context of IIE, then describes a number of challenges and promising directions for ongoing research.

2: Interaction

2.1: Human-Computer Interaction

Attempts to enrich the capabilities of human-computer interaction build on a wide variety of technologies, including tracking, stereoscopic display, and input devices (e.g. dataglove) with greater expressivity than the limited computer mouse. The use of video projection to create an immersive space for the visualization and manipulation of a virtual world was popularized by the CAVE Automatic Virtual Environment [2]. This allows for large scale 2D and 3D visualization in environments where a limited number of physical objects and other humans can occupy the same space [3].

Position tracking permits the calculation of an appropriate perspective projection based on the viewer's position. Gesture recognition, in particular focusing on the user's hands, is of particular importance to effective interaction. Computer vision approaches to this problem, as shown in Figure 1, have been considered as a means of specifying control parameters or for the manipulation of virtual objects in the environment without the encumbrance of a worn dataglove.

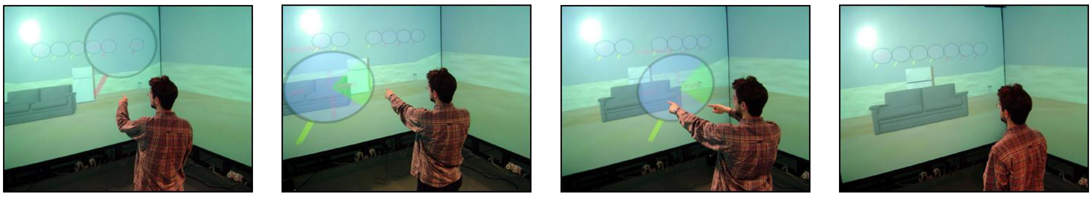


(a) Locating a wrist by particle filter.

(b) Fingertip identification using the circular Hough transform.

Figure 1. Location and extraction of a user's hand and fingertip for gesture analysis. (a) Edge detection segments the hand from a known (static) background and a particle filter identifies the most probable location and orientation of the wrist. (b) Various techniques, such as the circular Hough transform can be used to identify fingertip positions, although this requires a segmented view of the hand from the background, such as obtained using a skin-colour classifier.

Even without the resolution required to track a user's hand gesture in a relatively large environment, arm tracking can be employed reasonably to permit effective bi-manual manipulation and control over the virtual environment, using tools designed appropriately to the application. An example is illustrated in Figure 2 in which a special control tool, known as the *pieglass* [4], is used to provide a large vocabulary of actions that can be applied to objects in the environment.



(a) Pointing (b) Selection (c) Invocation (d) Deselection

Figure 2. (a) Holding the arm partially extended moves virtual cursors on screen, (b) fully extending the non-preferred hand selects a control tool, (c) fully extending the preferred hand invokes an action, specified by the wedge of the selected tool, and (d) lowering the arms releases the control tool.

2.2: Human-Human Interaction

With respect to human-human interaction, relevant technologies include video compression, segmentation, view synthesis, and network communication architectures. For the distributed case, multi-party interaction typically presents inordinate difficulties [5], both of managing and displaying multiple video streams in a socially effective manner, as well as coordinating the potentially disruptive sound sources in order to avoid feedback and general disruption. N-way interaction, with $n > 2$ frequently breaks down, even for co-present individuals working within a common synthetic environment.

Conventional efforts approach the problem of multiparty interaction from an AccessGrid model [6], for example, the Virtual Auditorium [7], which arrays the various users in discrete window coordinates on a large display. This arrangement fails to reproduce any of the positional cues associated with normal physical co-presence and makes no effort to support interaction with the environment. Another direction is to represent the participants as avatars, fully embedded in a synthetic 3D environment, as in the various CAVE Research Network (CAVERN) projects.

Virtualized reality techniques, which use video segmentation and compositing to blend the video of participants into a synthetic background, take this a step further. The hypothesis is that by providing a live video representation of each individual, the physically distributed participants gain a strong sense of co-presence, that is, the feeling of being together in a common, shared physical environment. Typically, these techniques involve the combination of information from multiple cameras to perform view synthesis, that is, generating a novel view from an arbitrary (virtual) camera position based on interpolation or volumetric modeling techniques [8].

Similarly, video mosaicing, that is, stitching together a number of independent views to produce an output image of higher resolution or greater spatial extent, can be used for improved visual quality. However, state of the art techniques for the latter problem are unable to produce correct results from translated cameras viewing non-coplanar scenes (cf. [9]). Such cases typically result in duplication of background texture due to the depth-varying amounts of overlap between adjacent input images.

As an example of the virtualized reality approach, the Shared Reality environment [1] contains a number of screens, cameras, projectors, microphones, speakers, and a high-fidelity vibro-mechanical system for sensing and/or actuating movement on a platform. Active computer processing of input sources and synthesis of the output stream is applied in a networked, multi-room environment (see Figure 2.2). However, despite convincing insertion of segmented video of remote participants into the scene, the background environment remains highly limited in its complexity and affordances for interaction.

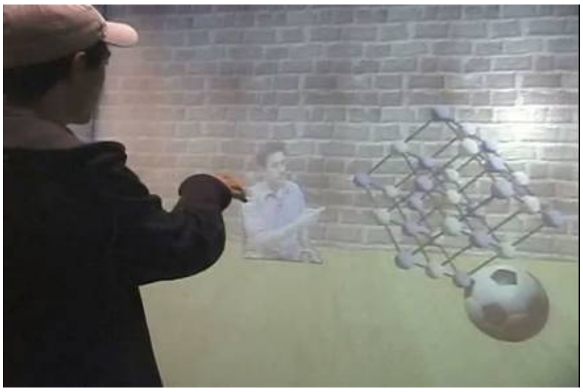


Figure 3. The prototype Shared Reality environment, with a remote participant rendered as if part of the local (synthetic) environment and matched with spatialized audio to reinforce the illusion of co-presence.

Other recent efforts to employ a teleimmersive video-based rendering of distributed users includes the Office of the Future project stemming from the TeleCubicles initiative [3], TELEPORT [10], Immersive 3D Videoconferencing [11], Hewlett Packard's Coliseum project [12, 13], and the Tele-immersive Environments for EVErybody project [14]. These have achieved impressive results, typically in the context of a desktop environment. However, in order to achieve their goals, these systems tend to impose physical constraints such as the need for special headgear or glasses, restrictions on user mobility based on the arrangement of cameras, limited use of other sensory modalities, or limited ability to interact equally effectively with synthetic, responsive media elements driven by the underlying applications.

Even when these issues are addressed, a major challenge to effective distributed interaction persists. This relates to viewing angles of the participants relative to the camera positions and orientations. For a single desired viewpoint, artificial view synthesis can simulate the properties of a virtual camera by taking advantage of information obtained from actual neighbouring cameras. However, for more than one participant, the divergence between the viewer's line of sight, intersecting the projection surface, and the optical axis of the camera responsible for that view, results in diminished gaze awareness and an inability for participants to communicate effectively through deictic (pointing) gestures. In our experience, this was most evident during a distributed jazz session (see Figure 2.2) with performers at McGill and Stanford University, as the musicians had great difficulty providing each other visual cues, for example, to coordinate turn-taking. Put simply, there is no single camera position that can provide a correct view for multiple participants, unless they are all located along the same optical axis.

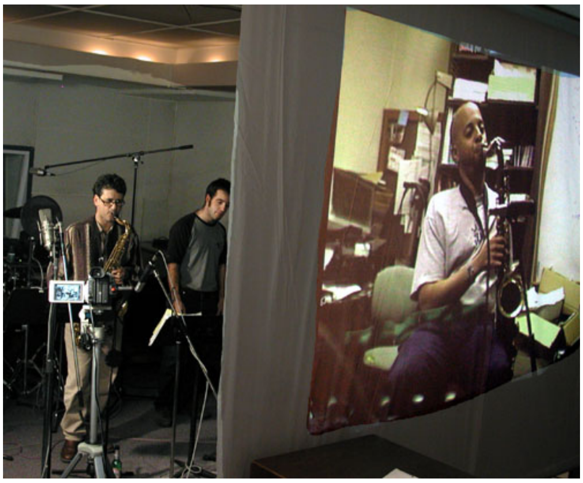


Figure 4. Musicians at McGill and Stanford universities playing together with low-latency network communication support. While the participants appear life-size on screen, camera-viewer disparity lead to confusion in establishing important communication cues such as eye-contact and subtle pointing gestures.

3: Challenges for Computer Vision Technologies in IIE

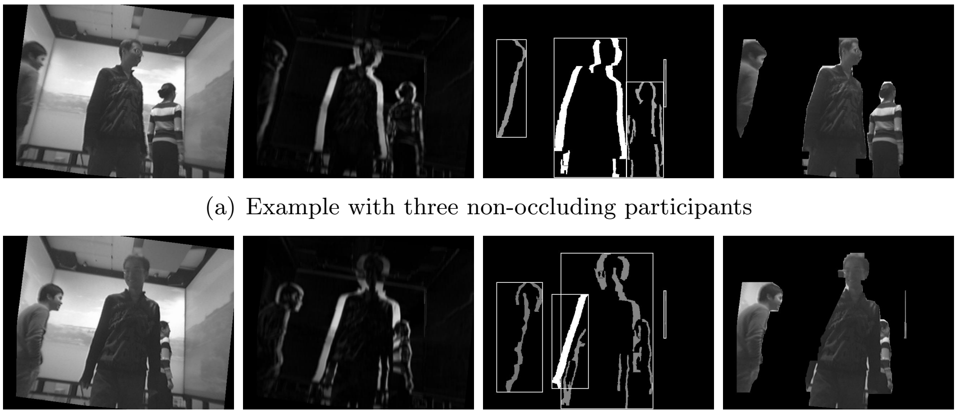
An obvious goal for research in IIE is to increase the *intelligence* of the environments with respect to their competence and capability of responding to human activity. Moreover, the environments should be robust to the number and location of users, ideally supporting highly coupled, complex group activity in a distributed setting. Such interaction may involve computer-generated (virtual) objects or media that respond in real-time to aggregate, as well as individual, human gesture. These goals impose onerous computational demands if seen in the context of tracking and gesture recognition. No less important, though, is the ability to support interaction, both human-human and human-computer, in a manner where the technology does not dominate the experience, that is, like Ubiquitous Computing [15], with the computer effectively disappearing from the users' awareness. In this light, we survey a number of important challenges related to communication and interaction in IIE.

3.1: Video Segmentation

Video segmentation, whether for analysis or rendering purposes, is a frequent component in IIE problems. It is a prerequisite for the video compositing described above, in order to blend remote participants into a shared, virtual environment, as well as an important first step in tracking or gesture analysis.

We have obtained significant improvements in this area through the exploitation of *differential disparity* information obtained from an indoor environment with known room geometry. First, during a calibration phase, we obtain accurate background depth maps by manual measurement (a fairly simple procedure for rectangular room geometries) of the wall surfaces relative to the cameras. While automated methods would be preferred, these often suffer from uncertainty when dealing with textureless regions. Furthermore, the computation of the background depth map can be performed once, off-line, for a given room configuration.

Then, during operation, we compare stereo pairs using the disparity information obtained from the depth map in the first step. This avoids any expensive search along epipolar lines for stereo matches, as we are interested only in whether the current depth agrees with the background map. The *differential disparity map* is then formed as the set of any pixel locations for which a correspondence is not found based on the values indicated by the background disparity map. In practice, textureless foreground regions tend to satisfy the correspondence, but the outlines of these regions invariably do not. Thus, we obtain reasonably clear contours of all foreground objects, not included in the original depth map, from which we may infer depth based on contour width [20]. Most importantly, these results are applicable to dynamic environments, where video may be projected on the wall surfaces, and the algorithm can execute at video frame rates. While still imperfect, this enables significantly improved video tracking and background segmentation operations to be performed, as illustrated in Figure 5.



(a) Example with three non-occluding participants

(b) Example with one person occluded by another

Figure 5. Segmentation results using the differential disparity map and contour grouping. From left to right, the first image is the video input, and the second illustrates the differential disparity map, in which those pixels that do not satisfy the background hypothesis are marked white. The third image contains contours that are grouped by thickness, representing common depth. Note that the technique succeeds in isolating individuals (as shown by the enclosing bounding boxes) despite occlusion. Finally, the fourth image illustrates our segmentation results.

Noting the potential for real-time object tracking with a certain accuracy in depth measurements, we believe that significant performance improvements are possible for such tasks as gesture analysis and front-projection shadow removal, discussed in the following section.

3.2: Shadow Removal

Front projection is a popular alternative to the space-consuming requirements of rear-projection. Unfortunately, occluders, in particular, human participants moving around the environment, create shadows on the display, thereby interfering with seamless communication. Shadow detection in front-projection environments has been considered as a means of eliminating shadows created by occlusion [16] while avoiding the space requirements of rear-projection systems. Traditionally, this process begins with an occlusion-detection step, based on a pixel-by-pixel comparison of a camera view of the projected display with a geometric- and colour-transformed version of the projection frame buffer, in order to identify those pixels whose colour differ from the expected values. Sukthankar et al developed a planar homography approach [17] that has been used by other researchers for occluder light suppression [18] and shadow removal [19]. This approach avoids the need for colour transformation by generating the predicted camera images off-line, but cannot scale to operation on live video data.

Current techniques have achieved modest results, typically operating near video rates but still suffer from ghosting and luminosity equalization, as seen in Figure 6.

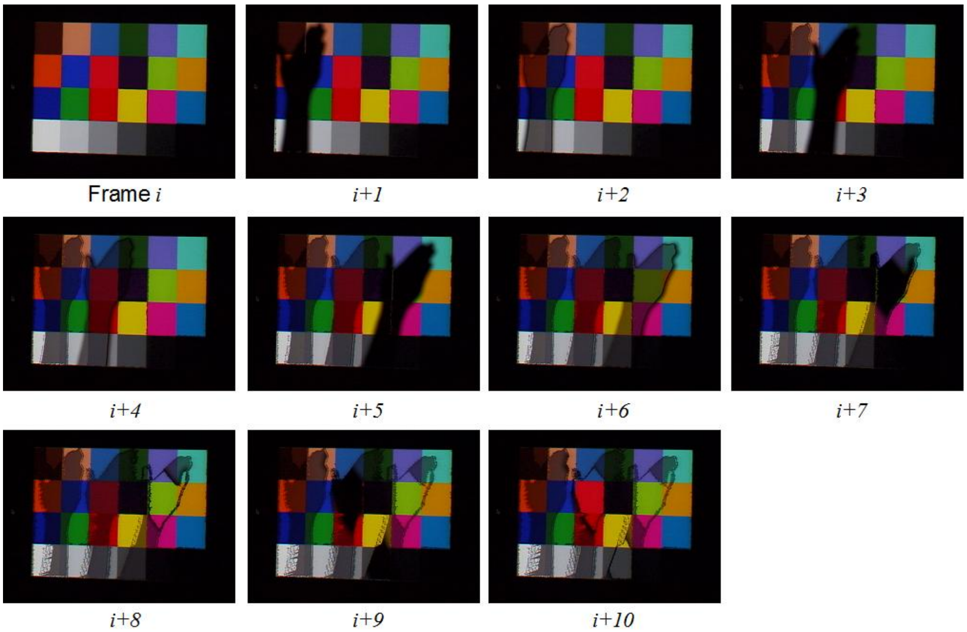


Figure 6. Shadow removal results for a sequence of captured camera frames.

As a possible improvement, rather than the traditional pixel-by-pixel colour-comparison between transformed images, it would be superior to employ direct knowledge of the occluder position in 3D and then, using appropriate camera-projector homography, directly estimate the corresponding shadow region it will produce, allowing for fill-in by another projector [21]. The video segmentation method described above represents a possibly suitable approach to determining the occluder position despite the presence of a dynamic (projected) background.

3.3: Responsive Media Synthesis

Other research has achieved environmental responsiveness through real-time synthesis of rich video and audio textures in reaction to human gesture and movement data, acquired from multimodal sensing. An illustrative example, shown in Figure 3.3 is the extraction of motion parameters from a video of a user's hand, which are then used to affect a physics-based fluid model (based on Navier-Stokes equations) such as smoke.

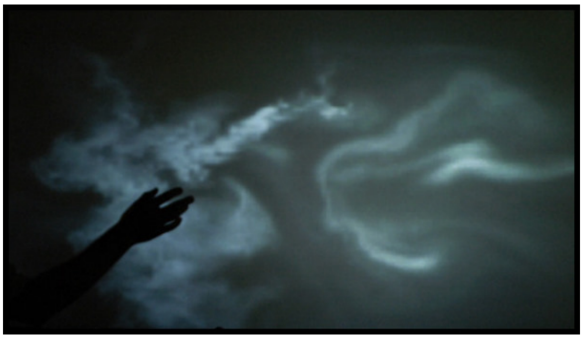


Figure 7. Parameterizing video textures through gesture using physics-based fluid modeling, implemented in Jitter, to produce realistic, real-time smoke and water video effects.

With respect to large-scale local interaction, examples exist wherein users are engaged in multimodal, n-way conversations, some with hundreds or on the order of thousands of participants in a single space [22, 23, 24]. A typical scenario consists of humans in the same physical space, interacting live with one another, augmented by computational ambient media that responds to their joint activity of gesture and movement. Although deployed exclusively in an artistic context, and thus, less constrained by the imperative for repeatable, direct responsiveness to individual actions, these experimental installations demonstrated the robustness and the engagement that such hybrid computational and physical media play spaces could generate.

3.4: Perceptual Immersion and Rendering Technologies

As the number and physical separation of users increases, in particular, in the distributed, networked context, a primary objective should be that of maintaining or improving the perceived richness and responsiveness of the environment, despite the associated increase in complexity. This must be achieved without imposition of physical or behavioural constraints or encumbrances on users, intrusion of technology, or an impairment of sensory modalities or communicative affordances (speech, gesture, movement) in relation to the freedom permitted by non-computer-mediated interaction. In other words, computer-generated media, whether of manipulable synthetic objects or representations of remote participants, should look, sound, feel, and *behave* like the equivalents in the real world. An obvious problem that must be addressed is the consistency and validity of representation as perceived by multiple users, whether physically co-present or distributed. This relates strongly to problems of graphical and auditory rendering.

Achieving this objective rests on two parallel strategies. The first originates from the Shared Reality concept of a common, shared space, in which remote participants and synthetic objects are video-composited into the environment, according to a defined room layout and their respective physical positions. In other words, the environment attempts to preserve verisimilitude with respect to the appearance and spatial relationships of the real world, thereby permitting (in theory) the conventions of eye contact, gaze awareness, pointing gestures, and sound source localization, between local and remote participants to function as one normally expects. Although no commercial systems support them adequately, these social cues are imperative for effective interaction.

The second builds on advances in audio and visual rendering capabilities, such that the environment appears perceptually correct regardless of the user's position within the space. Specifically, emerging technologies of autostereoscopic (3D) video and wavefield synthesis appear to hold promise for bringing the computer-generated world much closer to that of physical reality. Autostereoscopic displays offer the benefit not only of 3D imagery, but more importantly, multiple simultaneous views through the same display medium. This permits participants in each location to experience a display rendered as appropriate for their viewing angle, thereby enabling the social cues as described above between remote individuals. Similarly, wavefield synthesis expands the listening *sweet spot*, where audio localization is effective, from a highly constrained location to most of, if not the entire room. Together, these present a possible solution to some of the teleimmersive communication limitations as described earlier.

3.5: Latency and Jitter Bounds

Humans are highly sensitive to delay and variation in delay (jitter), whether in an action response by synthetic objects or in communication with each other. For maximum effectiveness, computation performed for such tasks as sensor data fusion, tracking, gesture recognition, output synthesis, and network transmission, must complete within the latency limit for the users to feel that the response is effectively concurrent with their actions as well as those of the remote participants. This bound is highly task and context dependent, ranging from values as low as 10ms for certain forms of musical interaction to a commonly accepted range of 100-300ms for interactive simulations. While the issue of real-time response has been tackled to a certain degree in telepresence settings and human-computer interaction, the problem becomes significantly more complex as the number of simultaneous users varies, since this may influence the computational resources required for processing.

4: Conclusions

While by no means comprehensive, the challenges we have described in the preceding sections constitute an important set of research topics that will have to be addressed in the future in order for IIE to achieve their full potential. As researchers, we are fortunate to have such a rich and stimulating field of scientific challenges ahead. look forward to observing progress in these areas in the coming years.

Acknowledgments

The author wishes to thank CANARIE Inc., Valorisation Recherche Québec, and the Natural Sciences and Engineering Research Council of Canada, whose generous funding has supported the research described in this paper. The work illustrated here has been carried out by numerous graduate students including Siu Chi Chan, Mike Wozniewski, Francois Rioux, Frank Rudzicz, and Sun Wei. Special thanks are also extended to Sha Xin Wei for numerous contributions to this text. Most importantly, kudos are due to Christopher

Jaynes and Robert Collins for their outstanding efforts of organizing and hosting the IIE workshop that led to a wealth of discussion on a number of the topics covered here.

References

[1] Jeremy R. Cooperstock. Interacting in Shared Reality. In *Proc. HCI International, Conference on Human-Computer Interaction*, July 2005.

[2] Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. DeFanti. Surround-screen Projection-based Virtual Reality: The Design and Implementation of the CAVE. In *SIGGRAPH '93: Annual conference on Computer graphics and interactive techniques*, pages 135–142. ACM, 1993.

[3] 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. In *SIGGRAPH '98*. ACM, 1998.

[4] Yves Boussemart, Francois Rioux, Frank Rudzicz, Mike Wozniewski, and Jeremy R. Cooperstock. A framework for collaborative 3d visualization and manipulation in an immersive space using an untethered bimanual gestural interface. In *Virtual Reality Systems and Techniques*, November 10–12 2004.

[5] William A. S. Buxton, Abigail J. Sellen, and Michael C. She asby. Interfaces for multiparty videoconferences. In K.E. Finn, Abigail J. Sellen, and S.B. Wilbur, editors, *Video-Mediated Communication*, pages 385–400. Laurence Erlbaum Associates, Hillsdale, N.J., 1997.

[6] Lisa Childers, Terry Disz, Robert Olson, Michael E. Papka, Rick Stevens, and Tushar Udeshi. Access grid: Immersive group-to-group collaborative visualization. *Immersive Projection Technology*, 2000.

[7] Milton Chen. Design of a virtual auditorium. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 19–28, New York, NY, USA, 2001. ACM Press.

[8] Greg Slabaugh, Ron Schafer, and Mat Hans. Image-based photo hulls. In *1st International Symposium on 3D Processing, Visualization, and Transmission*, pages 704–708, 2002.

[9] Matthew Brown and David G. Lowe. Recognising panoramas. In *International Conference on Computer Vision (ICCV)*, pages 1218–25, October 2003.

[10] Simon J. Gibbs, Constantin Arapis, and Christian J. Breiteneder. TELEPORT – towards immersive copresence. *Multimedia Systems*, 7(3):214–221, 1999.

[11] Peter Kauff and Oliver Schreer. An immersive 3d video-conferencing system using shared virtual team user environments. In *CVE '02: Proceedings of the 4th international conference on Collaborative virtual environments*, pages 105–112, New York, NY, USA, 2002. ACM Press.

[12] H. Harlyn Baker, Donald Tanguay, Irwn Sobel, Dan Gelb, Michael E. Goss, W. Bruce Culbertson, and Thomas Malzbender. The coliseum immersive teleconferencing system. In *International Workshop on Immersive Telepresence*, December 2002.

[13] H.H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M.E. Goss, W.B. Culbertson, and T. Malzbender. Understanding performance in Coliseum, an immersive videoconferencing system. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2004.

[14] Zhenyu Yang, Yi Cui, Bin Yu, Jin Liang, Klara Nshrstedt, Sang-Hack Jung, and Ruzena Bajscy. TEEVE: The next generation architecture for tele-immersive environment. In *Proc., 7th IEEE International Symposium on Multimedia (ISM'05)*, pages 112–119, 2005.

[15] Mark Weiser. Some computer science issues in ubiquitous computing. *Communications of the ACM*, 36(7):75–83, 1993.

[16] C. Jaynes, S. Webb, R. Steele, and R.M. Steele. Camera-based detection and removal of shadows from interactive multiprojector displays. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):290–301, May/June 2004.

[17] R. Sukthankar, T.-J. Cham, and G. Sukthankar. Dynamic shadow elimination for multi-projector displays. *Computer Vision and Pattern Recognition*, 2001.

[18] T.-J. Cham, J. Rehg, R. Sukthankar, and G. Sukthankar. Shadow elimination and occluder light suppression for multi-projector displays. *Computer Vision and Pattern Recognition*, 2003.

[19] M. Flagg, J. Summet, R. Somani, J.M. Rehg, R. Sukthankar, and T.-J. Cham. Shadow elimination and occluder light suppression for switched multi-projector displays. *ICCV*, 2003.

[20] Wei Sun and Jeremy R. Cooperstock. Disparity from countour for multiple object segmentation. Submitted for review to Special Issue on Computer Vision for Human-Computer Interaction, 2005.

[21] M.N. Hilario and J.R. Cooperstock. Occlusion detection for front-projected interactive displays. *2nd International Conference on Pervasive Computing*, April 21–23 2004.

[22] M. Kuzmanovic and N. Gaffney. Human-scale systems in responsive environments. *IEEE Multimedia*, 12(1), 2005.

[23] X.W. Sha, Y. Visell, and B. MacIntyre. Media choreography using continuous state dynamics on a simplicial complex. Technical report, Georgia Tech University, 2004.

[24] X.W. Sha, C. Salter, L. Farabo, and M. Kuzmanovic et al. T-garden. In *SIGGRAPH Art Gallery Publication*, July 2002.