# SIGGRAPH: G: Layered Telepresence: Simultaneous Multi Presence Experience using Eye Gaze based Perceptual Awareness Blending

MHD Yamen Saraiji,* Charith Lasantha Fernando, Kouta Minamizawa, Susumu Tachi

Keio University Graduate School of Media Design, Japan
Institute of Gerontology, The University of Tokyo, Japan

## Abstract

This paper proposes a novel technique to expand visual perception in Telexistence systems from a single location to multiple locations. Layered Telepresence (LT) expands perceptual awareness of human modalities by using the concept of layered perception. Telexistence based robots are used as the representation of the user, and are considered as perceptual layers containing the visual, auditory, and haptic feedback information. From each layer, saliency information is extracted and used as an indication of where the interesting regions are located at, these information are necessary for attention localization. Eye gaze is used as an input modality for weighting the layers based on the saliency map corresponding to each layer. Based on these weighted information, the layers are blended together into the same visual space and are presented to the user in a single visual space. Using this method of eye gaze based perceptual blending, we can expand user's visual awareness to multiple locations simultaneously as if the user is being ubiquitous. Here we describe the design of the proposed system as well as several applications using it.

**Keywords:** Simultaneous Multi-Presence, Perceptual Awareness Blending, Eye Gaze, Peripheral Vision, Depth of field

**Concepts:** •**Human-centered computing** → **Interaction techniques; Mixed / augmented reality;** •**Applied computing** → *Telecommunications;*

## 1 Introduction

Several off-the-shelf services provide tele-conferencing support for multiple users (such as Skype, Google Hangouts, etc) simultaneously through Picture-in-picture (PiP) grid of all users. However, in these services the user is always engaged with one participant during the session. Also, to switch between locations, the user either has to manually select which person to talk to, or the system automatically frames the active participant using Voice Activity Detection (VAD) engine. This type of manual or automatic switching reduces the engagement of the user due to the non-intuitive mechanism of switching between the remote locations. Also, this type of systems limits the perceptual awareness of the activity into a single location, that is the frame of which the user is engaged in at a specific time.

Depending on the system being used, the level of immersion also affects the experience of presence and activity engagement. Telexistence based systems [Tachi 2010] provide the user full or partial representation of his body while maintaining the visual and auditory mapping with users body, achieving an intuitive interaction in the remote site. This type of systems are highly efficient for teleoperation tasks. However, the user body is restricted to a single location at a single time. In other words, purely Telexistence systems define one-to-one relationship between the user and his representation. Same applies for teleoperation and remote control systems.

*e-mail: yamen@kmd.keio.ac.jp


**Figure 1:** *User experiencing multi-presence at two different locations using Layered Telepresence (LT).*

### Motivation

*"..he saw it as the hive-queen saw it, through many different eyes."*
*Ender's Game - Orson Scott Card 1985*

This research focuses on the emerging area of using digital and network media to expand the capability of human perceptions, and more specifically, the visual awareness and sense of presence. By expanding sense of presence we mean leveraging perceptual awareness from a singular location to being co-located at multiple locations simultaneously, achieving a true ubiquitous presence. This type of ubiquitous representation, and in order to be capable to have full awareness of being in these different locations, the following requirements should be addressed:

1. Real-time simultaneous representation of body, visuals, and auditory feedback in multiple locations.
2. Natural mechanism of presenting the visuals from the multiple sources to human user.
3. Intuitive mechanism for switching the visual perception between the locations, while maintaining the awareness of the other locations.

The proposed LT system addresses the previous points by using multiple Telexistence robots that are synchronized with users motion, and provides real-time visual and auditory feedback to the user. Robots are represented as layers of awareness, in which the information represented by each layer can be visual, auditory, or haptic feedback. These layers are blended based on the saliency found in them, and presented to the users feedback displays. Users

eye gaze is used as the main interaction modality for layers information presentation. Eye gaze is tracked and used to identify the target layer to be highlighted among the other layers based on the saliency information dominance. The layer user is looking at becomes focused while the other layers are defocused using an artificial depth of field effect, that maintains the optical flow in the peripheral vision of the user. Figure 1 shows the proposed system in action in which the user perceives two different locations simultaneously.

## 2 Related Work

Our presented system draws from the following area of research: Presence Augmentation & Alternation, Information Layering, and Eye Gaze Applications.

### Presence Augmentation & Alternation

[Lindlbauer et al. 2014] have proposed to use a physical see-through LCD display to mix the environment behind the screen and the contents of the screen achieving seamless blending between both contents. Although this approach of blending what is behind the physical object is still constraint to the same location, but technically it provides the sense of awareness of what is being displayed in the screen and what is behind it using direct interaction to control display's transparency. [Fan et al. 2014] proposed video based image blending using two-way see-through HMD that provides the user the awareness of both locations behind and front of him by blending the visuals of both views. Both approaches focus on expanding visual perception locally. Although these studies explores new types of interactions for switching between two different locations or mixing them, but also they have a role in increasing or altering the space which visual feedback is drawing input from, affecting sense of presence and enhancing it.
Presence augmentation was extended also by other researchers to include virtual presence in a different time. Researchers in [Suzuki et al. 2012] used pre-recorded 360 visuals and blends it with user's see-through HMD by phasing the visuals in and out through a specific scenario resulting the user experiencing past as being present. It is interesting to see the research direction in this area that addresses human perception in terms of presence, and the type of applications which lie beyond telecommunication and teleoperation.

### Information Layering

For presence related applications, taking human perceptual modalities into account is necessary to provide multiple telepresence avatar sources and mixing their feedback accordingly to achieve a seamless experience of co-presence at these multiple locations. Several researches in media and data presentation addressed how to combine multiple information in an intuitive manner, and the results can be expanded for telepresence applications and multi-presence.
In physical environments, when multiple objects are arranged at different depth distance from the perspective point, a well known phenomenon occurs in our visual perception: Depth of field, or image blurriness for objects out of focal plane. Previous works used this phenomenon to visualize data [Kosara 2004], and were also used in a multi-user applications [Yao et al. 2013] to selectively focus on the person of interest during teleconference applications to focus. [Cohen et al. 2007] proposed to use an architecture of sources and sinks to direct the auditory feedback from multiple sources into a single sink. Although this architecture does not define a general way of mixing or layering information from these multiple sources, but it proposes a framework for multi-sensory feedback systems.

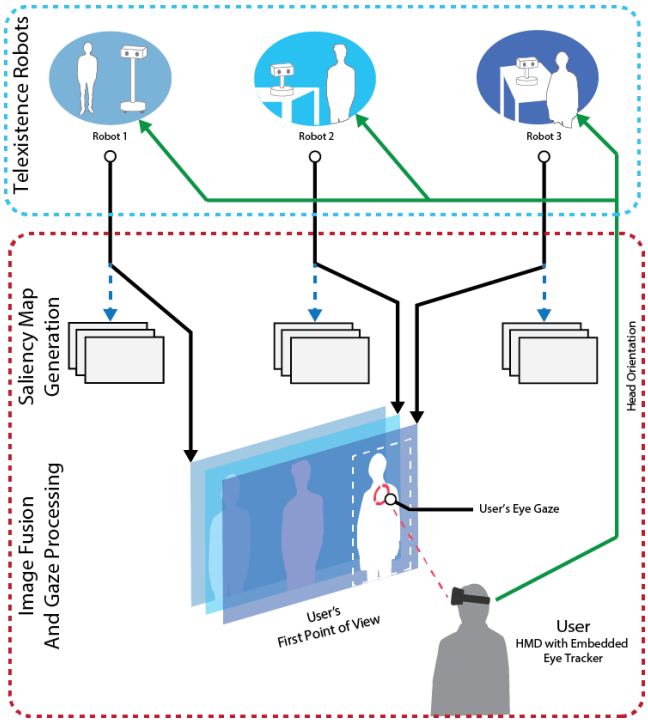


**Figure 2:** *LT system flow and interaction between user and remote telexistence robots.*

### Eye Gaze Applications

In this work, eye gaze is an important input to assist the system to bring to focus the layer of interest. Eye gaze applications had been an interesting area of research in various fields related to affective computing, computer human interaction, and others[Hutchinson et al. 1989; Morimoto and Mimica 2005].
Previous researches showed the effectiveness of using the eye gaze input for selection applications compared with pointing input using mouse [Sibert and Jacob 2000]. Smooth pursuit for eye gaze motion [Barnes 2012] was also used in several applications for target selection [Esteves et al. 2015], security related applications [Cymek et al. 2014], and general text entry [Lutz et al. 2015]. The proposed system adopts eye gaze input modality to achieve natural and selective navigation within the different remote locations.

In LT, an artificial depth of field was used to combine the layers of presence, in which the layers focused by user's gaze position will be brought to the foreground so the user can perceive it clearly, and layers residing in users peripheral vision are blurred out according to their priority according to users gaze.

## 3 System Description

The developed system is divided into a Master-Slaves Telexistence systems. The master side is the operating side where the user is located, and it contains a set of tracking tools that are used to capture user's head movement and eye gaze. Robots located at the remote sides are the same design and are connected with the user over LAN network. Figure 2 shows an overview of the system.

### 3.1 Robot Side

In this system, a custom three degrees of freedom Telexistence robot head was designed. HD stereo cameras and binaural microphones are used to enable stereo visual and binaural auditory communication to the user from robot side. In this design, we used for the pan, tilt, and roll joints servo model (HerkuleX DRS-0201/DRS-0101). The joints are driven by micro controller (Arduino ProMicro) that is connected over serial port to an embedded PC (Intel NUC D54250WYK1). For the cameras used, a low latency capture cameras were selected for this design (See3CAM CU130) that outputs sufficient frame-rate for the used head mounted display (format YUYV 640x480@60 FPS), and equipped with a 90 wide field of view lens. The image stream captured by the cameras from each robot is compressed using H264 format (using library GStreamer v1.6.1) with bit rate 3000 bps, and streamed to the user side over UDP connection. The measured end-to-end latency of the image stream using 802.11 wireless LAN is around 100-130ms for dual stream image streams. This low latency performance is required in order to reduce VR sickness for the users while using HMD.

### 3.2 User Side

User side (or master side) operates the remote robots motion using head rotation. Two types of setup were used for the user:

1. Exocentric viewpoint type, in which an external display is used to project the layers.
2. Egocentric viewpoint type that uses HMD to immerse the user with the layers.

For exocentric type that uses an external monitor to fuse the layers, we used eye gaze axis (X and Y) to drive two angles of the robot (Pan and Tilt respectively). And for Egocentric mode, we used a commercial HMD (Oculus DK2) that embeds gyroscope sensor and provides three axis angles. For eye gaze tracking, we used an off-the-shelf eye gaze sensor (Tobii eyex) that provides X-Y eye gaze coordinates in the screen space. Eye gaze is used as an input to LT system to determine which spot the user is looking at, and based on this input, the system controls the focus of the layers to the corresponding layer user is looking at. To determine the candidate layer that should be in focus, visual saliency maps were generated for each layer. User side software was developed under Unity3D environment, a custom video and audio streaming plugin was developed that ports image and audio data to Unity. The plugin uses GStreamer library to handle media streaming and decoding. Most of image processing parts were handled using EmguCV library (OpenCV .NET wrapper library) under Unity3D. The system tested on desktop PC setup with the following specifications: processing unity is Intel Core i7@3.40GHz, memory 16 GB, and graphics processing unit model NVidia GeForce GTX980. All the reported results regarding the performance, frame-rate, and latency were done using the previous setup. To maintain interactive VR experience and avoid any motion sickness, the system runs at an average of 60FPS regardless of the number of layers used due to the multi-threaded design of the system.

### 3.3 Saliency Map Generation

Saliency maps in this method are responsible to represent the presence of remote participants as a weight map generated for each captured frame while taking to consideration the temporal factor of the frames. The process of generating the saliency maps is done by a combination of two image analysis and features extraction methods. First the layers are processed for human presence, the procedure is done by applying Haar cascades classifier on each captured frame and the results are rectangles set representing the detected faces regions. These rectangles are expanded proportionally to their size to cover user body size (manually tuned, Width factor: 200%, Height factor: 400%). In practice however, using facial detection only fails to provide continuous tracking of presented people for several reasons such as partial occlusion of the face, lighting conditions, and resolution of the captured images. Also relying on facial detection only limits the visual saliency maps to capture the information of moving objects in scene. We addressed this limitation by adding a second layer of tracking using optical flow detection in the layers. Lucas-Kanade method was used to track scene features points for changing, when local changes occur in the layers, motion vectors are recorded for later registration in the corresponding saliency map of the layer. Figure 3(A) shows the detected motion vectors of a remote person.


**Figure 3:** *Saliency maps generation procedure (A) Human presence and motion vectors detection algorithm, (B) the resultant saliency weighted map.*

Next, the saliency maps are filled with weight of 1 for the pixels corresponding to the registered feature points and facial regions for each layer. To avoid the presence of hard edges along the detected regions, a Gaussian blur filter is applied to the saliency maps. To assist feature tracking consistency over time, a temporal process is applied to the calculated maps, a window of 500ms of previously calculated frames is used to calculate the weighted sum of the final saliency map. Using this procedure, the saliency maps maintained higher consistency in both tracking and representation of participants area in video frames. Figure 3(B) shows the final saliency map for a single layer.

### 3.4 Layer Fusing

Fusing the layers, or mixing them, is considered the final step of this method to deliver the layers to the user view area. One of the main considerations in LT is the user should maintain clear visuals

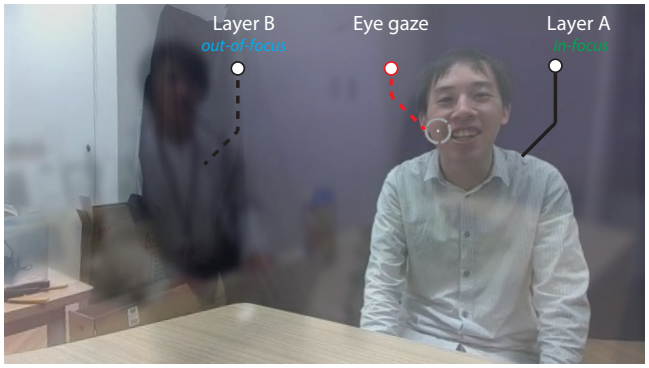


**Figure 4:** *Two different locations fused together with an artificial depth of field.*

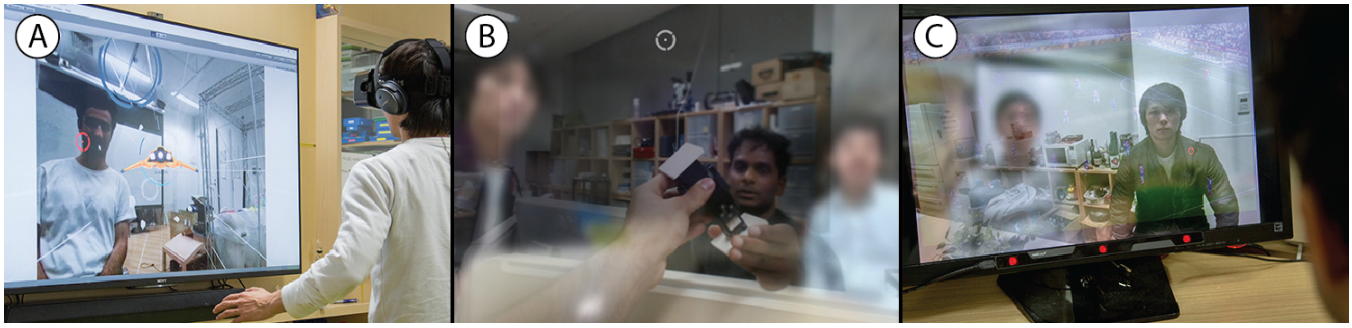


**Figure 5:** *LT System Applications: (A) Two layers system, a user engaged in VR game while talking with remote participant (B) Three layers system, a video-see through layer combined with two remote locations (C) Three layers system, a user watching soccer match while talking to two different locations*

and auditory feedback from the location he is engaged at, while being aware of the other locations simultaneously. Firstly, the layers are fused together based on the weight of each layer which is assigned based on user's eye gaze. Each layer's weight is calculated by sampling saliency map corresponding to each layer using eye gaze coordinates.
In preliminary experiments of layer fusing, we used a basic alpha blending to all layers based on the calculated weight of each layer, however we found that its difficult for the users to clearly distinguish the visuals of the layers due to visual overlapping between all locations. To address this issue, we considered using a similar phenomenon seen in image reflection over window glass. Basically this phenomenon of reflection and transparency of window glass allows to see two different locations simultaneously as well as the ability to focus at different depths that would result blurriness of objects in the background of both locations (depth of field). Using this phenomenon, a pseudo model was defined that defines a focal value for each layer, that is basically driven from the calculated layer weight, is used to control the amount of shallowness of layers out of focus. Figure 4 shows the final results of fusing two layers using the proposed method.

## 4 Applications

This method can be expanded to more than just Telexistence related applications. By using the concept of layered perception, its possible to generalize the layers to be media or even interactive applications, and apply the same procedure in combining them into a single space. Also, we categorize the set of applications into two different modes based on the viewpoint (Egocentric & Exocentric), in which they reflect how the user perceives the remote environment or the layers.

### Egocentric Applications

The user in this type of application is immersed with the task he is doing, thus perceiving the contents from the first point of view. For example when the user is being engaged in a virtual reality game while wearing HMD. By using the concept of LT, the game it self is also considered as a layer that can be combined with other layers while using eye gaze for interaction. The user is capable to be simultaneously engaged with the activity he is doing as well as with remote discussion. Figure 5(A) shows an example of LT with a virtual game, the user maintains the optical flow of the game contents and remote location while wearing HMD. Another example of this category is to use a video see-through HMD to represent the local location as a layer of presence, thus it can be seamlessly combined with remote locations. Figure 5(B) shows a user interacting with three different locations simultaneously while maintain the interaction with the local space his body is presented at.

### Exocentric Applications

In this category of applications, the user perceives the layers from an external display with an non-wearable eye tracking device. The contents of the display (application, visual media, etc.) can thus be considered as a layer of presence, while remote robots are blended with contents of the display. Figure 5(C) shows an example of interacting with a software while being engaged in a simultaneous meeting in two different locations. This type of applications is useful at increasing user multi-tasking, and as a substitute to window-based and PiP applications.

## 5 Evaluation & Discussion

A preliminary evaluation was conducted on LT system and its effectiveness using eye gaze and the simulated depth of field effect for multi-layer applications listed in section 4. In egocentric applications, users first calibrated their eye-gaze after wearing a video-see through HMD. Afterwards, the system connected to two different robots located in separate rooms, and a person is presented at each location with mutual visual and auditory feedback. Several users reported the sensation of being both persons were presented at the same location initially. They also reported a phantom sensation due to seeing their own body being layered with the other two locations. Interestingly, several users tried to reach the other participants with their hands before realizing that the others were actually layered and are not presented physically at the same location. One important factor in determining the location of the person was the color of the background which the remote participant is located at. When the color matches both locations the user was unable to distinguish the difference immediately, in contrast with different backgrounds situation.
For exocentric type of applications, a similar eye-gaze calibration step was performed on a 27" screen size. Followed by presenting the users with two different locations. The users were more likely to distinguish the difference immediately than in egocentric LT type. This can be regarded to the less immersion effect than in HMD type, in other words screen size and head motion are important to provide the sense of multi-presence.

## 6 Limitations and Future Work

The proposed interaction expands the simultaneous telepresence interaction from one location to multiple locations, however the number of layers that can be combined and presented are limited due to perceptual and technical limitations. Perceptual limitations can be summarized by the limitation of our visual system to process overlapping visuals effectively, for example when two persons are presented at the same location. Fortunately the visuals are stereoscopic (remote images coming from left and right eye of the telepresence robots), thus its possible to use eye convergence [Ryana et al. 2012] as an interaction modality to focus on a specific layer. Though eye convergence can help to selectively choose layers at different depths, but the problem of background visual noise will still be presented. It can be reduced by using efficient saliency and filtering post-processing on the layers to reduce the visual noise.
The other limitation is technical limitation of multi-layer presentation, and this limitation can be related to the processing time required for each layer and its effect on the VR experience when using current generation of VR headsets. Currently, most headsets requires at least 90FPS capable softwares to provide seamless experience and to avoid drop frames, thus the processing time for each frame should be at most $\frac{1000}{90} = 11.\overline{111}$ms in order to maintain the 90FPS constraint. With the current optimizations done for image processing, each layer costs approx 2ms for saliency map generation, and 3ms for combining the layers and producing the final image. Thus the maximum layers for this particular implementation can be estimated by the following equation:

$$Layers_{max} = \frac{11.\overline{111} - 3}{2} \simeq \frac{8}{2} = 4Layers \qquad (1)$$

Further optimizations would lead to an increase in the number of layers that can be used.

The concept of LT is to expand the bodily presence including haptic feedback. This study did not investigate how to layer the haptic feedback from multiple sources, but would be interesting topic to study. Furthermore, the mutual presence and remote interactions is necessary in telepresence, so further studies to be conducted to reproduce user's behavior at the remote locations for the non-active avatars.

## 7 Conclusion

This paper presented Layered Telepresence, a novel system for operating and blending multiple Telexistence based locations simultaneously using eye gaze perceptual awareness. The proposed system treats each robot as a layer of presence which can be blended together with other perceptual layers based on users eye gaze motion. This paper lists two categories of applications using the proposed method: Egocentric and Exocentric applications depending on the perceived layers point of view. By using the concept of layering, its possible to achieve an intuitive and seamless sense of presence in multiple physical or virtual locations simultaneously, and as a step to achieve ubiquitous sense of presence.

## Acknowledgement

This research is supported by the JST-ACCEL Embodied Media Project (JPMJAC1404)

## References

BARNES, G. R. 2012. Rapid learning of pursuit target motion trajectories revealed by responses to randomized transient sinusoids. *Journal of Eye Movement Research 5*, 3.

COHEN, M., BOLHASSAN, N. A., AND FERNANDO, O. N. N. 2007. A multiuser multiperspective stereographic qtvr browser complemented by java3d visualizer and emulator. *Presence: Teleoperators and Virtual Environments 16*, 4, 414–438.

CYMEK, D. H., VENJAKOB, A. C., RUFF, S., LUTZ, O. H.-M., HOFMANN, S., AND ROETTING, M. 2014. Entering pin codes by smooth pursuit eye movements. *Journal of Eye Movement Research 7*, 4.

ESTEVES, A., VELLOSO, E., BULLING, A., AND GELLERSEN, H. 2015. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, ACM, 457–466.

FAN, K., HUBER, J., NANAYAKKARA, S., AND INAMI, M. 2014. Spidervision: extending the human field of view for augmented awareness. In *Proceedings of the 5th Augmented Human International Conference*, ACM, 49.

HUTCHINSON, T. E., WHITE, K. P., MARTIN, W. N., REICHERT, K. C., AND FREY, L. A. 1989. Human-computer interaction using eye-gaze input. *IEEE Transactions on systems, man, and cybernetics 19*, 6, 1527–1534.

KOSARA, R. 2004. *Semantic Depth of Field-Using Blur for Focus+ Context Visualization*. PhD thesis, Kosara.

LINDLBAUER, D., AOKI, T., HÖCHTL, A., UEMA, Y., HALLER, M., INAMI, M., AND MÜLLER, J. 2014. A collaborative see-through display supporting on-demand privacy. In *ACM SIGGRAPH 2014 Emerging Technologies*, ACM, 1.

LUTZ, O. H.-M., VENJAKOB, A. C., AND RUFF, S. 2015. Smoovs: Towards calibration-free text entry by gaze using smooth pursuit movements. *Journal of Eye Movement Research 8*, 1.

MORIMOTO, C. H., AND MIMICA, M. R. 2005. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding 98*, 1, 4–24.

RYANA, L., MACKENZIEA, K. J., AND WATTA, S. J. 2012. Multiple-focal-planes 3d displays: A practical solution to the vergence-accommodation conflict? In *3D Imaging (IC3D), 2012 International Conference on*, IEEE, 1–6.

SIBERT, L. E., AND JACOB, R. J. 2000. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, 281–288.

SUZUKI, K., WAKISAKA, S., AND FUJII, N. 2012. Substitutional reality system: a novel experimental platform for experiencing alternative reality. *Scientific reports 2*, 459.

TACHI, S. 2010. *Telexistence*. World Scientific.

YAO, L., DEVINCENZI, A., PEREIRA, A., AND ISHII, H. 2013. Focalspace: multimodal activity tracking, synthetic blur and adaptive presentation for video conferencing. In *Proceedings of the 1st symposium on Spatial user interaction*, ACM, 73–76.