第 21 回日本バーチャルリアリティ学会大会論文集(2016 年 9 月)

Study on Telexistence LXXXV
Layered Presence: Expanding Visual Presence using Simultaneously Operated Telexistence Avatars

MHD Yamen SARAIJI [1)], Charith Lasantha FERNANDO [1)], Kouta MINAMIZAWA [1)], Susumu TACHI [2)]

1) 慶應義塾大学大学院メディアデザイン研究科 (〒223-8521 神奈川県横浜市港北区日吉 4-1-1, yamen@kmd.keio.ac.jp)
2) 東京大学 高齢社会総合研究機構 (〒113-8656 東京都文京区本郷 7-3-1)

Abstract: This paper contains an overview for a novel technique to expand visual perception in Telexistence systems from one location to multiple locations. Layered Perception (LP) uses eye gaze perceptual awareness blending to combine visual feedback from several Telexistence robots into a single location, and each robot represents a layer of presence. By estimating the visual saliency of these layers and combining them into a single visual space based on user's eye gaze motion, we can expand user's visual awareness to these multiple locations simultaneously. Here we describe the design of the proposed system as well as several applications using it.

Keywords: Teleimmersion, Telexistence, Augmented Reality.

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.

Telexistence base systems [1] provide the user full or partial representation of his body while maintaining the visual and auditory mapping with user's body, achieving an intuitive interaction in the remote site. However, the user body is restricted to a single location at a single time.

In order to achieve the experience of being presented at multiple locations simultaneously and to be capable to have full awareness of these locations, the following requirements should be addressed:

- Real-time simultaneous representation of body, visuals, and auditory feedback in multiple locations.
- Natural mechanism of blending the visuals from the multiple sources.
- Intuitive mechanism for switching the visual perception between the locations, while maintaining the awareness of the other locations.

Figure 1 Layered Presence system overview.

The proposed LP system addresses the previous points by using Telexistence robots that are synchronized with user's motion, and provides real-time visual and auditory feedback to the user. Robots are represented as layers of awareness, and these layers are blended and presented to the user's displays. User's eye gaze is tracked and used to identify the target layer to be highlighted among the other layers, in which the layer use is looking at becomes focused while the other layers are defocused using an artificial depth of field effect. 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: Vision Augmentation, Eye Gaze Applications and Multi-Space Image Fusing.

2.1 Vision Augmentation

Several approaches were proposed to address blending two different locations simultaneously at one physical or virtual location. Lindlbauer et al. [2] 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. Fan et al. [3] proposed video based image blending using see-through HMD that provides the user the awareness of both locations behind and front of him. Both approaches focus on expanding visual perception locally.

2.2 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[4][5]. Previous researches showed the effectiveness of using the eye gaze input for selection applications compared with pointing input using mouse [6]. The proposed system adopts eye gaze input modality to achieve natural and selective navigation within the different remote locations.

2.3 Multi-Space Image Fusing

In physical environments, when multiple objects 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 [7], and were also used in a multi-user applications [8] to selectively focus on the person of interest . In the proposed method, we use artificial depth of field to combine the layers of presence, in which the layers focused will be brought to the foreground and the user can perceive it clearly, and layers residing in user's peripheral vision are blurred out according to their priority according to user's gaze.

3. System Design

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

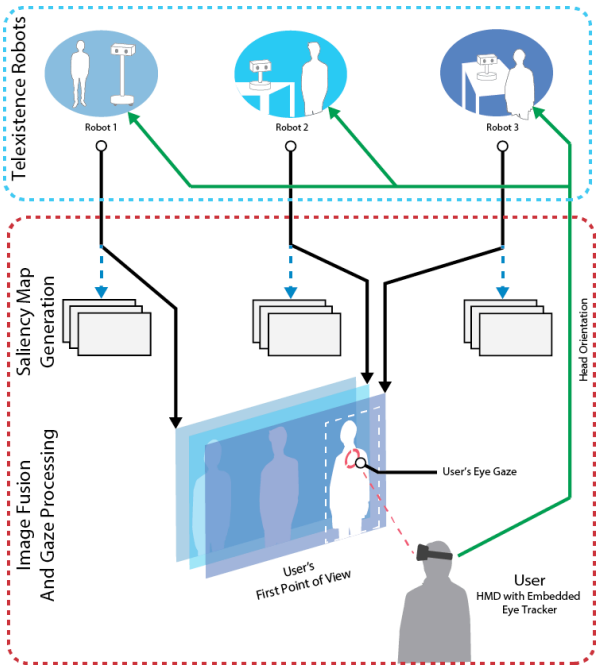

Figure 2 LP system flow and interaction between user and remote telexistence robots.

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:

- Exocentric viewpoint type, in which an external display is used to project the layers.
- 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 LP 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. In this setup, 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.

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 participant's area in video frames. Figure 3 (B) shows the final saliency map for a single layer.

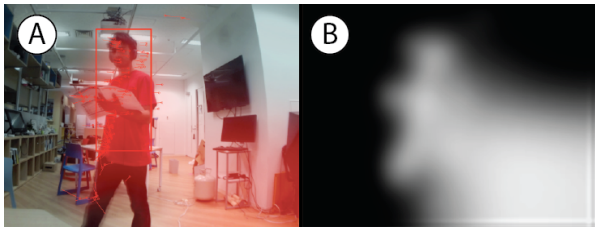

Figure 3 Saliency maps generation.

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 LP is the user should maintain clear visuals 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.

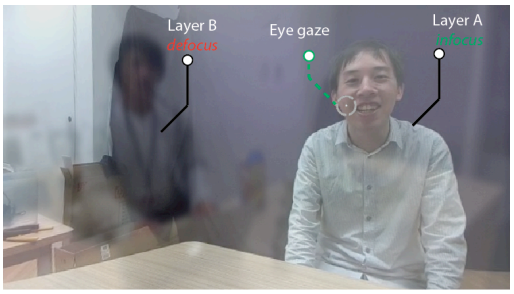

Figure 4 Two different locations fused together with an artificial depth of field .

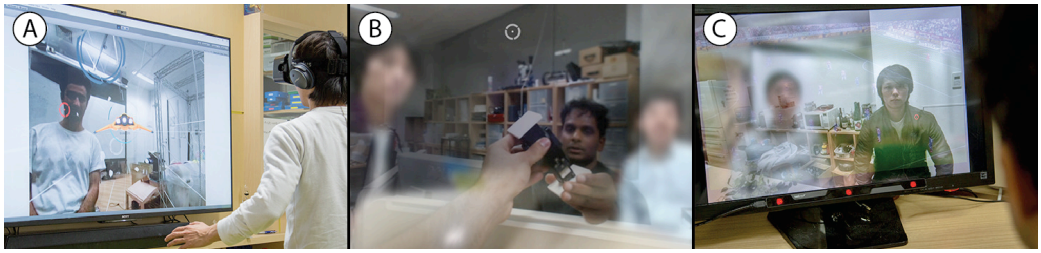

Figure 5 Saliency maps generation.

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 define two modes of viewpoint (Egocentric & Exocentric), in which are based on user's point of view and how the user perceives the remote environment or the layers.

4.1 Egocentric Applications

The user in this type of application is immersed with the task he is doing, for example when the user is being engaged in a virtual reality game while wearing HMD. By using the concept of LP, the game it self is also considered as a layer that can be fused with other layers with eye gaze enabled. By using this method, 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 LP with virtual game. Also by using a video see-through HMD, its possible to consider the local location as a layer, and thus it can be seamlessly fused with a remote location as shown in Figure 5 (B).

4.2 Exocentric Applications

In this category of applications, the user perceives the layers from an external display with an external eye tracker mounted into the display. 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.

5. Conclusion

This paper presented Layered Presence, 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 user's eye gaze motion. This paper lists two distinguished types of applications using the proposed method: Egocentric and Exocentric types of applications. By using the concept of layering, its possible to achieve an intuitive and seamless sense of presence in multiple physical or virtual locations simultaneously.

Acknowledgement

This research is supported by the JST-ACCEL Embodied Media Project.

References

[1] Tachi, Susumu. "Telexistence." *Virtual Realities*. Springer International Publishing, 2015. 229-259.

[2] Lindlbauer, David, et al. "A collaborative see-through display supporting on-demand privacy." *ACM SIGGRAPH 2014 Emerging Technologies*. ACM, 2014.

[3] Fan, Kevin, et al. "SpiderVision: extending the human field of view for augmented awareness." *Proceedings of the 5th Augmented Human International Conference*. ACM, 2014.

[4] Hutchinson, Thomas E., et al. "Human-computer interaction using eye-gaze input." *IEEE Transactions on systems, man, and cybernetics* 19.6 (1989): 1527-1534.

[5] Morimoto, Carlos H., and Marcio RM Mimica. "Eye gaze tracking techniques for interactive applications." *Computer Vision and Image Understanding* 98.1 (2005): 4-24.

[6] Sibert, Linda E., and Robert JK Jacob. "Evaluation of eye gaze interaction."*Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000.

[7] Kosara, Robert. Semantic Depth of Field-Using Blur for Focus+ Context Visualization. Diss. Kosara, 2004.

[8] Yao, Lining, et al. "FocalSpace: multimodal activity tracking, synthetic blur and adaptive presentation for video conferencing." *Proceedings of the 1st symposium on Spatial user interaction*. ACM, 2013.