Natural Interaction with Video Environments Using Gestures and a Mirror Image Avatar

Christian Kray, Dennis Wilhelm, Thore Fechner, Morin Ostkmap

Natural interaction with video environments using gestures and a mirror image avatar

Christian Kray, Dennis Wilhelm, Thore Fechner, and Morin Ostkmap

Institute for Geoinformatics, Westfälische Wilhelms-Universität Münster
c.kray@uni-muenster.de, d.wilhelm@uni-muenster.de
t.fechner@uni-muenster.de, m.ostkmap@uni-muenster.de

Abstract. Video environments are a promising option for a variety of applications such as training, gaming, entertainment, remote collaboration, or user studies. Being able to interact with these environments enables further applications and extends existing application scenarios. In this paper, we propose a novel interaction technique that combines natural gestures with mirror images of the user to allow for immersive interaction with video environments. The technique enables movement inside the 3D space depicted by the video as well as the placement and manipulation of virtual objects within the 3D space. We describe a potential application scenario, where interactive public displays are placed inside a scene by one user and then experienced by another user. We also briefly report on a user study evaluating the gesture set we defined for controlling movement within the video.

Keywords: gestural interaction, mirror image, avatar, video

1 Motivation

There are many scenarios, where a realistic visual simulation of a real environment is needed: training, gaming, entertainment, or remote collaboration are some examples. In addition, the design and evaluation of ideas and (context-dependent) systems can strongly benefit from such simulations. The design process of ubiquitous systems such as public display networks can also benefit from realistic simulations of the intended deployment area: instead of implementing a fully functional system or deploying screens in the real world, designers can create mock-ups using simulations and thus gather feedback from users early on in the development process. Immersive video environments are one option to convincingly simulate the real world. While they are easy to create and very realistic, they do not include semantic or geometric information. Movement in the depicted 3D space and interaction with objects shown in the footage is thus not realized easily.

In this paper, we propose an approach to overcome these issues. The approach also enables the injection of new content into the video footage and the subsequent experiencing of this content. Our system combines gestural interaction with a mirror image of the user that serves as an avatar within the video

environment. It thereby enables the intuitive selection of 3D locations shown in video environments as well as the placement of virtual objects inside the 3D space depicted by the video footage.

2 Related Work

There are different approaches to realize visually convincing simulations. One is to use virtual environments (VEs), i.e., computer generated scenarios based on a detailed 3D model. Another approach is to use photographs or video footage to generate an immersive experience. Synthetic 3D models allow for fine grained details and interaction, while the actual modeling requires a lot of work. While parts of the process can be automated, e.g., using 3D scanners [6], the overall effort is still considerable. Conversely, photographs or video footage provide a realistic (audio-)visual simulation and can be captured quite easily and effortlessly but interaction with the shown objects is limited.

Different means of interaction are available to move within the simulation, e.g., treadmills, and to manipulate objects shown in the simulation (such as gloves or voice commands). Gestures can also be used for this purpose. Särkelä et al. [9] analyzed different gestures and notions to let users navigate in a VE. Vogel and Balakrishnan [12] define a design space for freehand pointing and clicking interaction. Nancel et al. [8] investigated how to implement mid-air pan-and-zoom gestures on wall-sized displays. Benko and Wilson [3] analyzed multi-point mid-air gestures for omnidirectional immersive environments.

A key issue in creating immersive experiences is the lack of haptic feedback. Vogel and Balakrishnan suggest to compensate for this by using additional visual and auditory cues [12]. In general, humans perceive visual stimuli more pronounced than auditory or tactile ones [10]. One possible approach to provide additional visual feedback is to employ a mirror metaphor. Mirrors are commonly used, and most users are thus familiar with how to operate them. This metaphor has been used in a variety of contexts [1, 5, 7, 10, 11]. Uses include raising awareness of health issues, motivating behavior change, or simply using a mirror image to focus attention during interaction.

Similar to the approach presented by Ahn et al. [1], the approach proposed in this paper uses the user's mirror image as a video avatar. The avatar can be used to navigate within the VE and to manipulate virtual objects. In contrast to previous work, our system does not merely substitute a cursor with a mirror image. Our approach is focussed on providing an immersive experience by using intuitive gestures in combination with the video avatar. It supports both the creation of augmented scenes, where virtual objects are inserted into video footage and the exploration of such scenes. Gotardo and Price [4] aimed at a similar workflow and developed a system that is comparable to the one presented here. However, their approach is based on a much more complex hardware setup. Moreover, the user interface designed by Gotardo and Price is based on a heads-up-display (HUD) rather than on gestures to let users select actions (e.g., selecting or scaling objects). A purely gestural interface may be experienced as a

more natural way to interact with an immersive video environment (IVE), both by designers and participants. In addition, the system proposed here allows for a quick and easy creation of scenes from video footage and does not require custom and expensive hardware.

3 Approach

The main goals of our approach were to enable intuitive interaction with video environments and to immerse people watching video footage as much as possible into the real world scene being shown. Our aim was thus to enable users to perform various actions in the simulated environment while providing them with a strong feeling of presence. The basic idea underlying our proposed interaction technique is to create a realtime mirror image of the user and overlay it over the video footage. Using a simple, layer-based depth model and a small set of gestures, users can place their mirror avatar inside the depicted real-world scene and interact with the environment via the avatar (e.g., to place virtual objects). In the following paragraphs, we describe all essential components of the approach in detail and provide an overview of our prototypical implementation.

3.1 Mirror image avatar

A live mirror image of the user constitutes the focal point of interaction, see Figure 2(a)–(d). The background behind the user, which is also captured by the camera pointed at the user, is eliminated in real-time, e.g., using standard difference image or chroma-key techniques. The cut-out mirror image serves as an avatar or proxy for the user: its location in 3D space defines where interaction can take place.

In particular, the avatar defines the depth layer that is currently selected. Since both the user and the system's depth camera know how tall the user is, the actual size of the avatar as depicted on screen defines its depth position inside the video footage. The size of the avatar is used as a depth cue to inform the user about the layers and objects that can be selected. For example, by placing the avatar on the third layer three, the user can interact with objects located on that layer and can inject virtual objects on that layer.

3.2 Gestures

Avatar control. In order to move the avatar within the 3D space defined by the video footage and the layer model, users can perform a number of gestures. When one foot is placed in front of the other, users can move their avatar along the X-, Y-, and Z-axis. To control the movement in X- and Y-direction, users use a one-handed gesture. By extending one arm in one of the cardinal directions, a user can specify in which direction the avatar should move, see Figure 1 (left). For example, extending the arm to the left moves the avatar in that direction.

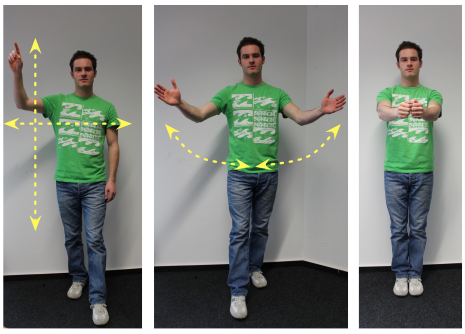

Fig. 1. Key gestures used to control the avatar, interact with the video scene, and to manipulate virtual objects: moving gestures (left), scaling gesture (middle), switching gesture (right).

Movement continues while the arm is extended. Movement stops when users bring their arm close to their bodies, or when they perform a different gesture.

A two-handed gesture controls movement along the Z-axis (depth). Putting both hands closely together in front of the body shrinks the avatar. This corresponds to moving it deeper into the image, i.e., it increases its distance from the camera. By spreading both arms, users can increase the size of the avatar and thus decrease its distance to the camera, see Figure 1 (middle). The scaling stops when users either return their arms to a relaxed position, or when they perform another gesture. The size of the avatar determines which depth layer is selected: the depth information specified for each layer and the actual height of the user enables the system to compute the best match, i.e., to find the layer which corresponds best to the current size of the avatar.

In order to evaluate these gestures, we carried out a comparison study that contrasted it with an alternative set of gestures, where movement was controlled by walking in place while orienting one's body in the target direction. Twenty participants (ten male, ten female, average age 22.35 years, SD: 1.927) were recruited via word of mouth and other means from around the university. They were asked to navigate several immersive video scenes in the IVE with the help of the mirror image avatar using two different sets of movement gestures. All participants had to use both methods, but the order of exposure was randomized. Due to space constraints we cannot report on the entirety of the results here but can only summarise the key findings. Participants were largely successful in navigating the avatar to the target locations using either method. We frequently observed people stopping when their avatar reached a street and avoiding "collisions with cars," which indicates a high degree of immersion. The gesture set depicted in Figure 1 was rated more favorably (SUS score of 74.25, SD: 12.76) than the comparison set (SUS score of 64.63, SD: 23.13). 70% of the participants also preferred the static gesture shown in Figure 1 overall.

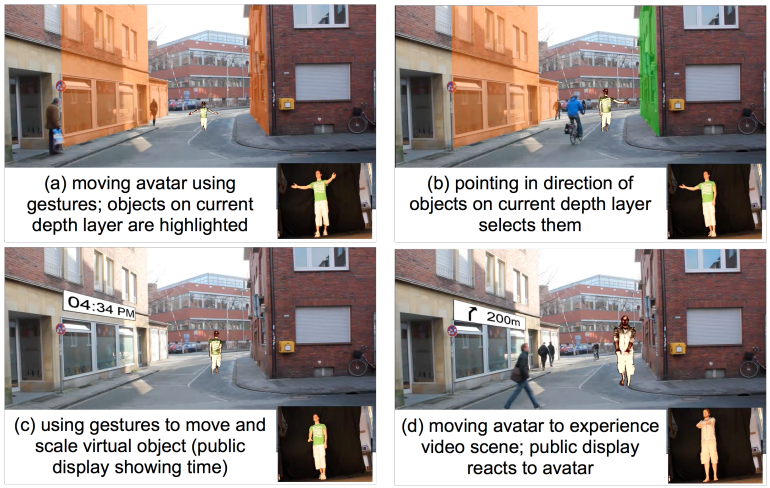

Fig. 2. Example scenario: (a) moving and scaling the avatar, (b) selecting scene element, (c) placing virtual object, (d) experiencing virtual object (e.g., public display).

Object manipulation. In addition to moving the mirror image avatar, we defined a set of gestures to select objects depicted in the video footage (e.g., buildings or signs), to inject virtual objects (e.g., public displays, new buildings, or audio sources) as well to move and scale those objects. In addition, gestures to select content to inject into the scene or other content-related activities can be defined. For example, in our prototypical implementation we included gestures to select an item to inject from a list of options.

In order to select an object in the video footage, users first need to place their avatar on the corresponding layer. Users then select an object by simply pointing in the direction of the object. The gestures to place an object in 3D space are the same ones as those used for moving the mirror image avatar. Users can switch between moving their avatar and moving objects by putting their hand together, extending their arms in front of them and then maintaining this pose for a short time, see Figure 1 (right). Visual feedback, i.e., a progress bar, indicates the switch from one set of actions to another.

Experiencing augmented video scenes. Once video footage has been augmented with a number of virtual objects, people can experience the new "scenario" in the following way: by moving their mirror image avatar through the 3D space defined by the video footage and the layer model (as described above), they can interact with the objects augmenting the video scene. The layer model provides means to, for example, measure the distance to a public display and realize proxemic interaction [2]. The next section describes an example scenario, which demonstrates how users can experience augmented video scenes.

4 Example Scenario

We created a prototypical implementation of our approach using predominately web-based technologies and two cameras (a webcam and a depth camera), which runs in real-time inside a standard browser on a desktop PC. In order to demonstrate the use of our approach, we propose the following example scenario, see Figure 2. Designers want to create a public display system that consists of a series of screens distributed throughout the city. These screens are meant to react to passersby and provide them with personalized information. Instead of installing real displays at the planned deployment sites, short video clips of those sites can be recorded and be used to prototype the system.

The first step is to construct the layer model: for each scene, the designers need to define a number of layers that are located at different distances (depth levels) from the plane defined by the camera. In addition, they can mark up a number of objects shown in the video footage and link them to a specific layer. Layers and objects define how people can later interact with the video footage. Once layers (and optionally objects) are specified, people can interact with the footage. Designers can now use the gestures shown in Figure 1 to move their avatar around the video scene. As they move their avatar, previously defined objects are highlighted, see Figure 2 (a)—indicating, which objects can be "reached" from the current position of the avatar. When designers have positioned their avatar at the desired location, they can point at objects, which are associated with the layer the avatar is located on, see Figure 2 (b). Additionally, the designers can inject virtual objects into the footage. In the example scenario, the designers insert, move, and scale a mockup of a public display, see Figure 2 (c). The virtual object is associated with the layer on which the avatar is located.

The resulting augmented video footage can then be explored by other people, e.g. stakeholders or the people who commissioned the public display system. Figure 2 (d) shows a stakeholder who is exploring the scene that the designers created previously. Using the gestures shown in Figure 1, the stakeholder moves they avatar around the video scene. The public display reacts to the avatar, and when its within its activation area, the display content changes and shows directions targeted at the user. The stakeholder can thus experience the design in a realistic way and explore, for example, whether the activation area of a public display or its content are fit for the intended deployment location.

5 Discussion

The initial prototypical implementation suffers from a number of limitations. It currently only supports a small set of objects that can be placed inside the video scene, and at the moment, their orientation in space cannot be changed. In addition, only one user can interact with the system at the same time. Furthermore, movement of the avatar is not restricted so that users can place it in physically impossible positions (e.g., floating above ground), which could break immersion. Finally, both the avatar and virtual objects are simply overlaid over the video

footage: moving objects such as cars that intersect with these simply disappear behind them regardless of where they are supposed to be in the 3D space defined by the video. This is another aspect that can negatively affect immersion.

Most of these limitations can be addressed by improving the current implementation. A more generic import mechanism for assets to be injected could use a different set of gestures (or a mobile device) to allow for the use of arbitrary virtual objects. A more sophisticated and robust gesture recognition system would allow for rotation gestures and enable multi-user interaction. A more sophisticated layer model could also specify permissible locations of the avatar on each layer to prevent avatars from being moved to physically impossible locations. Realizing physically correct occlusions involving the avatar would require a deeper analysis of the video and a more sophisticated spatial model.

Finally, several limitations relate to the gestures we used. While the gestures we defined were learned quickly by the participants and positively received in our user study, further studies are required to identify the most immersive/intuitive set of gestures. So far, we only assessed a subset of all the gestures, i.e., movement control. In addition, we did not test whether the use of devices, e.g., mobile phones, to carry out certain actions, such as injecting virtual objects, would be more immersive/intuitive, neither on their own nor in combination with gestures. Further studies on these aspects are desirable as well.

Generally speaking, mirror image avatars could also be used with photographs or true 3D environments, i.e., virtual worlds. We chose to use video footage as we expected the real-time motion of the avatar to blend in more naturally with the movement naturally occurring on video footage, and would thus create a strong sense of presence and immersion. Using true 3D environments would allow for correct occlusions but constructing realistic virtual worlds requires a lot of effort. Compared to a desktop scenario, where a user would place objects and experience augmented video scenes, we argue that the gesture-based approach combined with a large screen provides a more realistic and immersive experience. Initial informal feedback from people seeing the system in action as well as observations from the initial user study on the movement gestures seem to confirm this, but we intend to carry out a series of user studies to investigate these aspects in more detail. Clearly, further studies are also needed to identify the most suitable gesture sets for different tasks.

6 Conclusion

In this paper, we proposed a novel approach to interact with video environments in an immersive and intuitive way. Using their avatar, users can move inside the footage in three dimensions, and place virtual objects inside the scene depicted on the video. Knowledge about the height of the user and the layer model enable the system to place the video avatar in three dimensions. The system can be used for various applications, for example, the prototyping and evaluation of ubiquitous and situated systems. We presented an example scenario, where designers first

placed an interactive public display in a video environment and a stakeholder then explored the resulting scenario.

The initial prototype, though limited, used web technologies to illustrate the feasibility of the approach. A first study provided initial evidence for a high degree of immersion and the usability of the proposed approach. Future work will focus on carrying out a series of further user studies. With the latter, we plan to explore properties of mirror image avatars as a means of interaction in simulated environments, to compare this approach to alternatives, e.g., 3D controllers, and to investigate the integration of mobile devices with the system, e.g., as a secondary controller for content selection.

References

1. Ahn, S., Lee, T.S., Kim, I.J., Kwon, Y.M., Kim, H.G.: Large display interaction using video avatar and hand gesture recognition. In: A. Campilho, M. Kamel (eds.) Image Analysis and Recognition, *LNCS*, vol. 3211, pp. 261–268. Springer (2004)
2. Ballendat, T., Marquardt, N., Greenberg, S.: Proxemic interaction: designing for a proximity and orientation-aware environment. In: Proc. ITS '10, pp. 121–130. ACM (2010)
3. Benko, H., Wilson, A.D.: Multi-point interactions with immersive omnidirectional visualizations in a dome. In: Proc. ITS '10, pp. 19–28. ACM (2010)
4. Gotardo, P.F.U., Price, A.: Integrated space: authoring in an immersive environment with 3d body tracking. In: SIGGRAPH '10 Posters. ACM (2010)
5. Iwabuchi, E., Nakagawa, M., Siio, I.: Smart makeup mirror: Computer-augmented mirror to aid makeup application. In: J.A. Jacko (ed.) Human-Computer Interaction. Interacting in Various Application Domains, *LNCS*, vol. 5613, pp. 495–503. Springer (2009)
6. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proc. UIST '11, pp. 559–568. ACM (2011)
7. Morikawa, O., Maesako, T.: Hypermirror: toward pleasant-to-use video mediated communication system. In: Proc. CSCW '98, pp. 149–158. ACM (1998)
8. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., Mackay, W.: Mid-air pan-and-zoom on wall-sized displays. In: Proc. CHI '11, pp. 177–186. ACM (2011)
9. Särkelä, H., Takatalo, J., May, P., Laakso, M., Nyman, G.: The movement patterns and the experiential components of virtual environments. Int. J. Hum.-Comput. Stud. **67**(9), 787–799 (2009)
10. Andrés del Valle, A.C., Opalach, A.: The persuasive mirror: Computerized persuasion for healthy living. In: Proc. HCI International '05 (2005)
11. Vera, L., Gimeno, J., Coma, I., Fernández, M.: Augmented mirror: Interactive augmented reality system based on kinect. In: P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, M. Winckler (eds.) Human-Computer Interaction – INTERACT 2011, *LNCS*, vol. 6949, pp. 483–486. Springer (2011)
12. Vogel, D., Balakrishnan, R.: Distant freehand pointing and clicking on very large, high resolution displays. In: Proc. UIST '05, pp. 33–42. ACM (2005)