Mediating Presence in
Browser Based Videoconferencing

Jimmy Nyström, Nicklas Nyström, and Peter Parnes
Luleå University of Technology,
Department of Computer Science, Electrical and Space Engineering
{jimmys, nicnys, peter.parnes}@ltu.se

Abstract. The new technologies of the Open Web Platform (OWP) are currently redefining what is possible in a browser application, opening up a new world of possibilities for browser based videoconferencing. This paper explores new ways to combine these new web technologies with depth sensors. A prototype videoconferencing system was developed, featuring online remote collaboration exercises, collaborative viewing and manipulation of 3D models, and 3D motion parallax video. To enable intuitive interaction with the system, it was outfitted with a depth sensor based touch interface.

Shared viewing and manipulation of 3D models was found to be a useful addition to classic videoconferencing and can be used to convey information in ways not possible with shared images or video. The motion parallax 3D view developed holds promise in the field of mediated gaze, but did not quite reach the goal due to technical limitations. The system shows promise for further investigation into collaboration on interactive surfaces.

Keywords: Videoconferencing, interactive surfaces, depth sensors, motion parallax, WebRTC, WebGL

1 Introduction

The European union is striving for a more climate-friendly and less energy-consuming economy[7]. One step towards this goal could be to decrease our traveling, preferably without making people feel isolated and disconnected. This is the primary focus of the Mediating Presence project of EIT ICT Labs, wherein several prototypes have been developed for this purpose. This paper describes one of them.

Videoconferencing can be a great tool for mediating presence and has several benefits compared to physical meetings, especially when they can replace long business trips. However, videoconferencing is still not considered a fully adequate substitute for traditional meetings. There seems to be a need for new tools and improvements.

According to a survey conducted by the authors, critical areas that are in need of improvement are accessibility, the sense of presence and the ability to
cooperate. Advances in OWP and WebRTC are now making browser applications more accessible and user-friendly than ever before. By allowing the browser itself to perform high-performance tasks without the use of external plugin software, using such an application can be as easy as opening a webpage. This paper describes the development of methods to combine new web technologies and depth sensors in the field of remote online collaboration. In particular, the goal has been to find new ways to add to the users’ sense of being in the same location, or ‘non-mediation’. A prototype videoconferencing system with the following features was developed:

**Regular videoconferencing** based on WebRTC, with audio/video communication and text chat.

**Share** lets users share, view and manipulate 3D models together in real-time. Users are able to rotate and change the appearance of the models in different ways.

**Play** is a multiplayer game to be used as an example of a collaboration exercise using WebGL. The puzzle game Klotski was chosen for this purpose.

**Peek** uses a 3D camera to stream video/depth data to give users a three-dimensional view of each other. Head tracking is used to change the perspective of the 3D environment, creating a motion parallax effect.

2 Background

This section describes previous work and important concepts related to mediating presence.

2.1 Shared Mediated Spaces

There exists a number of prototypes and finished products in the area of shared mediated spaces. For example, A. Chung and D. Jung built a shared whiteboard for online collaboration, where users could see rough representations of each other on the screen[9]. They concluded that body gestures were not very helpful for faster task completion or better user experience using their system. However, this may only be true for full-body gestures conveyed through rough models. In 2012, A mediated sketching table was designed as part of a project within EIT ICT Labs[10]. The shared workspace was created by projecting live video from remote users on top of the local user’s workspace, similar to the setup in Figure 2. The work was continued in the Tangibles project, involving a class of computer science students (including the authors) at LTU in 2012. Using this sketching table, it was concluded that being able to see each other’s hands was a helpful tool for remote drawing lessons.

2.2 Depth and Motion Parallax

One way of increasing the feeling of presence is to make the video appear more realistic, for example by giving it a sense of depth. The human brain determines
depth mainly through stereopsis; the two slightly different images received by
the eyes are compared, and the disparity is used to calculate distances. However,
the brain also uses another depth cue, called motion parallax: When observing
moving objects, those that are closer seem to move faster than those that are far
away. This information is used to determine the distance to objects. Though not
nearly as effective as stereopsis[6], motion parallax produces a consistent and
reliable impression of depth[1]. A video screen that changes its viewpoint based
on the user’s line of sight to create motion parallax can be used to achieve a
sense of depth.

2.3 Video-mediated gaze awareness

Gaze is a vital part of body language. It is helpful for giving feedback and ex-
pressing feelings and attitudes[5]. In fact, it is paramount in establishing trust[2].
The ability to see where someone else is looking over a video link is called me-
diated gaze. Besides its practical uses, mediated gaze also has abstract social
value; people who use eye contact frequently are seen by others as more atten-
tive, friendly and cooperative than those who do not. In [2], the authors showed
that improved eye contact increased the participants’ experienced satisfaction
with videoconferencing as a medium for negotiation. A videoconferencing sys-
tem that does not allow for mediated gaze lacks valuable social information[3].
Most current videoconferencing systems do not allow for mediated gaze. It is
possible that this may have been a limiting factor in their social acceptance[4].
A typical videoconferencing system uses a webcam placed above or below the
screen. This means that when you look directly at someone’s eyes on the screen
on such a system, to them it will appear as if you are looking away slightly.

3 Method

An online survey was conducted to investigate people’s views on current video-
conferencing systems, and to gauge their interest in different ideas for new fea-
tures. These features were rated on a scale from not useful to very useful. The
survey was made publicly available, and was also presented to the Mediating
Presence group of EIT ICT Labs. 32 people responded to the survey1. Their an-
swers were used as a basis for deciding what features to implement in a prototype
videoconferencing system. One of the proposed features was a shared workspace,
described as follows:

In a videoconference, a second screen could be used as a complement to the
video view displaying useful information such as shared images, documents or
video. The second screen could be a smartphone, a tablet computer, another
monitor or a projector projecting an image onto the table in front of you. 67% of
respondents thought such a shared workspace would make a very useful addition
to videoconferencing.

1 The full survey and a summary of all responses can be found in the authors’ joint
The idea of being able to see other users’ hands on the shared workspace received mixed responses, with a median vote of "somewhat useful". However, two previous projects[10][8] identified this as a potentially helpful tool. Because of that, such a feature will be implemented in future versions of the system.

The system was developed to answer the following questions:

– How is practical remote collaboration affected by the following factors:
  • Video of the other person’s face
  • Audio from the other person
  • An indicator of where the other person is pointing (such as a mouse pointer)
  • Video overlay of the other person’s hands
– Are there any advantages to using shared 3D models, as opposed to regular multimedia, such as images and audio/video? If so, what are they? Can shared 3D models be a viable substitute for hands-on work with physical objects?
– Is it feasible to generate a live three-dimensional view of people engaged in a videoconference, directly in the browser?
  • Could this feature then be used to achieve mediated gaze?

Several participants claimed that current videoconferencing systems are not user-friendly enough. Therefore, user-friendliness and accessibility were added as system requirements. The technologies used in the project were chosen with this in mind. The number of software installs required to use the system were kept to a minimum through the use of OWP. Below is a description of some of the technologies that were used:

**HTML5** is the fifth and latest revision of the HyperText Markup Language, the core language of the World Wide Web. Among the new features are improved support for multimedia and new markup and APIs allowing for powerful web application development.

**WebRTC** enables web applications to incorporate audio, video and data Real-Time Communication (RTC) directly in the browser without relying on external plugin software.

**WebGL** is a JavaScript API that allows for hardware-accelerated 2D and 3D graphics rendering in the browser.

**Zigfu** is a plugin that lets the browser communicate with depth sensors, such as Microsoft’s Kinect and ASUS Xtion. It provides video and depth data streams (though in low resolution, only 160 × 120) as well as body and gesture tracking.

4 Implementation

The prototype website features collaborative 3D model viewing, a collaborative game and three-dimensional videoconference. These are all accessible in virtual rooms, each features on a separate "tab" that the user can switch between during a conference. A depth sensor is used in different ways on each tab. The tabs, called Share, Peek and Play, are described in the sections below, along with a description of the touch interface that can be used to control the system.
4.1 Share
The Share tab allows users to view 3D models by simply dragging a file and dropping it in the browser window. Once loaded, the model can be shared with other users with the click of a button. A number of control panels are available for manipulating the model, allowing users to change - among other things - its size, color and orientation. Changes to the model are propagated to everyone viewing it, creating a shared experience. 3D models can be created by scanning physical objects with the depth sensor, but this requires third party software at the moment.

4.2 Peek
The Peek tab uses depth sensors to produce a live three-dimensional video chat that gives a real sense of depth by creating a motion parallax effect in response to user movement. To produce three-dimensional video, depth and video data from the depth sensor is streamed over the WebRTC link. The location of the pixels in the depth image, together with the depth information, is mapped to corresponding \((x, y, z)\) positions in 3D space. A texture generated from the video data is used to determine the color of each point. An example of how a rendered point cloud might look can be seen in Figure 1.

![Point cloud rendering with interpolation.](image.png)

The rendered image can be rotated arbitrarily. This changes the person’s apparent line of sight, and can be used to compensate for the offset between the camera and the screen that was discussed in Section 2.3, effectively removing one of the greatest obstacles to mediated gaze. Finally, the motion parallax effect is created using the body tracking feature of the Zigfu depth sensor plugin: As the user’s head moves, the orientation of the 3D scene changes, making it appear as if the video has real depth.
4.3 Play

A multiplayer version of the sliding block puzzle game Klotski was implemented to serve as an online collaboration exercise. It is played by two participants, who must cooperate in order to solve each puzzle. The participants can communicate using audio and video, and by pointing: a sparkling dot is used to indicate the movement of the other participant’s mouse pointer. The time taken to finish each puzzle is logged, enabling quantitative measurements of user performance under different conditions. A screenshot from the game can be seen in Figure 2.

![Fig. 2. Left: Example setup of the system. Right: A screenshot from the Play tab.](image)

The game can be played using the mouse, but it is mainly intended for being played on a setup as illustrated in Figure 2, using the depth sensor to turn the table into an interactive surface. This way, blocks can be moved simply by touching and dragging them. Besides imitating the feel of a physical board game, this will allow for streaming 3D video of the users’ arms and hands as a transparent overlay on top of the game, enabling even more natural hand-based communication.

4.4 Touch Interface

The three tabs described above all make different use of the depth sensor, but they are all compatible with the touch interface designed for the collaboration exercise. To create a touch surface, an initial calibration is needed. This is done in the following steps:

1. The depth sensor is pointed at the surface that will be touch enabled. It can be virtually any screen or surface, as long as video from the computer can be projected onto it.
2. A reference image of the surface is created by linearly interpolating a number of succeeding frames of depth data.
3. A mapping between camera coordinates and browser coordinates is made by asking the user to identify five reference points that are projected onto the surface.
4. The user can fine-tune the touch detection by adjusting how far away from the surface touches should be registered, and align the video and depth images. The two images need to be aligned since the depth sensor uses two different cameras to capture them.

After the calibration, each new frame of depth data is compared to the reference image, resulting in a binary image where clusters of white pixels represent potential touches. These are detected using a flood-fill algorithm. Large clusters are tracked over time, generating JavaScript mouse events that cause the browser to react as if the mouse was clicked, dragged or released.

5  Evaluation

The system was set up at Luleå University of Technology and passersby were asked to sit down and try its different features, two at a time. A total of 10 people tried out the system. After each session, the participants were asked questions and discussed the system. What follows is a summary of their responses, and the conclusions that were drawn from them.

Shared viewing of 3D models was found to be a useful addition to videoconferencing. The interactive nature of a movable 3D object opened for discussions and actions that would not be possible with a simple image or video. For example, users rotated objects and zoomed in on specific parts of them to show something to the other person. The feature may find many use cases, especially if scanning of objects is simple. All test subjects agreed that the feature was easy to use, and successfully shared and manipulated 3D objects needing minimal instructions.

The head tracking and 3D-rendering was fast and responsive, resulting in a natural sensation of depth. Thus motion parallax 3D video streaming was proven to be feasible. Mediated gaze was not fully achieved, as the video quality was deemed too low.

According to user tests using our Klotski implementation, having video of the other user’s face while solving practical collaboration problems is not very useful. Users focused on the puzzle itself instead of the video window. On the other hand, hearing the other person’s voice was paramount to successful collaboration. A visual indicator of where the other person was pointing/touching was also deemed to be very useful for cooperation, as users relied heavily on this functionality to communicate.

6  Discussion and Future Work

The implemented prototype has several interesting features that can be accessed as easily as opening a website. Most of them, however, require further research and development before they are ready for widespread use.
8 Mediating Presence

The Peek feature is limited by the resolution of the depth sensor data; it was simply too low to convey any actual eye contact. Therefore, it was concluded that mediated gaze through 3D rendering is not feasible using the Zigfu plugin in its current form. It should be noted, however, that the technique shows promise, and mediated gaze should be possible if the browser could access the video streams with higher resolution. Updates to the plugin software and more efficient encoding of data could allow for streaming full resolution depth data.

The Play collaboration exercise is a good starting point for investing remote collaboration further. For example, the effects of having video overlay of the other user’s hands will be investigated in a future project.

Sharing of 3D models works best to mediate presence if they represent a physical object in one of the user’s location. Therefore, a simple way of creating 3D models by scanning real objects is needed for this to be truly effective. In the current implementation of the system, third party software is needed to perform the scanning, but this could also be done directly in the browser.

Overall, the prototype serves as a proof of concept that cheap, off-the-shelf hardware and a web browser can be enough to create an immersive collaborative videoconferencing experience.

References

1. B. Rogers and M. Graham, *Motion parallax as an independent cue for depth perception*, 1979
4. R. Vertegaal and Y. Ding, *Explaining Effects of Eye Gaze on Mediated Group Con- versations: Amount or Synchronization?*, Human Media Lab Queen’s University Kingston, CSCW ’02 Proceedings of the 2002 ACM conference on Computer supported cooperative work, ON K7L 3N6 Canada
7. A Roadmap for moving to a competitive low carbon economy in 2050, European Commision, Brussels, 8.3.2011