

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/268374033>

Immersive Multi-User 3D Video Communication

Article

CITATIONS

10

READS

193

7 authors, including:



[Oliver Schreer](#)

Fraunhofer Heinrich Hertz Institute HHI

124 PUBLICATIONS 1,785 CITATIONS

[SEE PROFILE](#)



[Harm Belt](#)

ASML

25 PUBLICATIONS 289 CITATIONS

[SEE PROFILE](#)



[Oscar Divorra Escoda](#)

Vonage

42 PUBLICATIONS 465 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



ACTION-TV [View project](#)



FascinatE [View project](#)

Immersive Multi-User 3D Video Communication

I. Feldmann¹, O. Schreer¹, P. Kauff¹, R. Schäfer¹, Z. Fei², H.J.W. Belt², Ò. Divorra³

¹Fraunhofer Institute for Telecommunications,
Heinrich-Hertz-Institute, Germany

²Philips Research, Netherlands

³Telefonica Research, Spain

ABSTRACT

The interest in immersive video conference systems exists now for many years from both sides, the commercialization point of view as well as from a research perspective. Nevertheless, so far the user acceptance of such systems was still very limited. This situation changed recently. Technological advances in fields like display and camera technology as well as processing hardware lead the way to a new generation of immersive tele-conference systems. On one hand, large scale and high definition displays significantly enhance the feeling of virtual presence. Nowadays, commercial solutions benefit from these facts. Besides this, research in the area of multi-user 3D display technology shows promising results. On the other hand, new fast graphics board solutions allow a high algorithmic parallelization in a consumer PC environment. In this way, real time, high quality and high resolution implementations of more sophisticated 2D and 3D acquisition algorithms, such as volume based approaches, become more and more realistic. From this point of view this paper summarizes first results and experiences of the European FP7 research project 3DPresence which aims to build a three party and multi-user 3D tele-conferencing system. The goal of this paper is to discuss general issues and problems of future generation immersive multi-user 3D video conference systems. Further on, it provides challenging first results and proposes solutions for critical questions.

INTRODUCTION

Research and development in the area of video communication from a local to one or multiple remote sides has a long tradition. Especially in the past few years the interest in generating a so called tele-presence increased rapidly. The work in this area includes topics like naturalness, feeling of physical presence, gesture awareness and eye contact etc. Recent high-end commercial solutions such as Cisco's TelePresence (see Figure 1), Polycom's TPX, and HP's HALO partially remove some of the tele-presence shortcomings of traditional systems with immersive high-quality audio and high-definition life-size video. Still, these systems do not present the remote participants in life-sized 3D, limiting the naturalness and thereby the sense of tele-presence. In addition, a fundamental problem is that eye contact is unnatural and that directional gaze awareness is missing.

An example based on a commercial system is given in Figure 2. On the left of Figure 2 the viewing direction of the remote participant towards the most right local participant appears correct. But as seen in Figure 2 right, from the position of the most right local participant, the eye contact and viewing direction is completely misleading, although the remote participant is looking directly to the local participant on its display.

Keeping eye contact is indeed one of the most relevant and challenging requirements in a tele-presence system from a non-verbal communication point of view, and while many attempts have been made, it has not yet been satisfactorily solved today.

Current state-of-the-art systems address it by mounting the camera behind a semi-transparent viewing display, but this common approach is often limited to the special case of having one single conferee at each side of the conference. Further, this approach requires a bulky optical and mechanical mounting that is only acceptable for niche market applications. A two way video conferencing system for three participants per site has been presented in [1], which provides nearly eye contact supported by cameras mounted on top of the displays.

One idea to overcome the eye contact problem is to virtually correct the captured view of the conference user. Figure 3 left illustrates this approach. The right-hand local participant is being captured by the top camera (see red arrow). In this way, when the conferees look at each other on the screen then no eye contact can be created. In order to solve this problem, the capturing camera will be virtually shifted to the eye position of the remote conferee on the screen. In this way, both participants can look each other into their eyes while being captured from the top of the displays. This approach requires a 3D acquisition chain. An early approach of this idea can be found in [2], [3].

Another problem is the gesture awareness. In order to keep the users impression as natural as possible, pointing gestures between conferees should maintain their original direction. Immersive video-conference systems usually are based on the idea of a shared virtual table environment, as illustrated in Figure 3 left and Figure 4 left. Nevertheless, the correct reproduction of gesture direction requires a geometrically correct virtual environment. Especially for multi-user and multi-party systems this is often difficult to realize for all conferees simultaneously. Following the sample in Figure 2 the misleading head pose direction illustrates that a pointing gesture would fail its directionality in the related setup too.

Finally, state of the art commercial systems are still based on 2D display technology. Nevertheless, a 3D representation of the virtual remote conferees would increase the virtual impression drastically. Nowadays, large scale 3D auto-stereoscopic display technology is available on the market. One major restriction for their application in immersive video-conference systems is the missing capability for multi-user environments. Further on, robust and high quality 3D data acquisition chains are required which are still difficult to realize, especially looking at the real-time constraints of such systems.

This paper will discuss these problems from the aspect of recent technological advances in display technology and hardware processing power. Rather than presenting technical details



Figure 1 – State of the art telepresence system by CISCO (left), and the Polycom TPX system (right)



Figure 2 – Example of correct (left) and misleading (right) eye contact of the Cisco tele-presence system

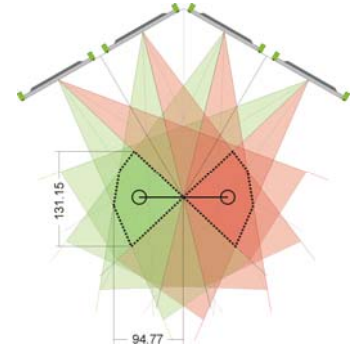
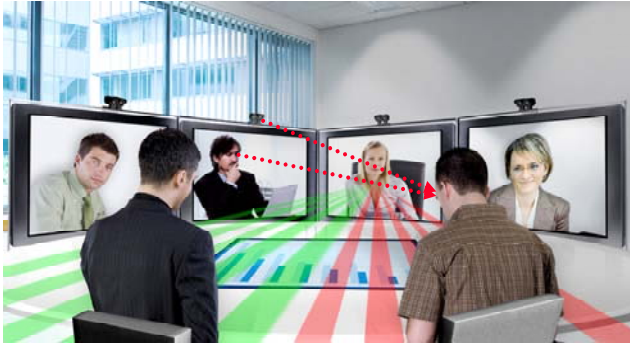


Figure 3 – Left: The 3DPresence multi-user videoconferencing concept, right: 2x4 viewing cones meeting the two local conferees

the goal of this paper is to discuss more general issues and problems of future generation immersive multi-user 3D video conference systems. The discussion is based on results and experience of the FP7 research project 3DPresence which aims to develop a multi-user multi-party tele-conference system. We will provide challenging first results and propose solutions for critical questions.

In the following, firstly, the 3DPresence multi-user and multi-party teleconferencing concept is explained. Secondly, the general challenge in the geometrical design of a shared virtual tele-conference system is highlighted. Thirdly, the multi-user auto-stereoscopic display aspect is discussed. A new multi-view and 3D auto-stereoscopic display will be introduced. Afterwards, the algorithmic challenge is discussed. It will be shown that sophisticated 3D acquisition algorithms, such as visual hull and multiple parallel stereo matching systems can be applied by still fulfilling the real-time limitations.

THE 3DPRESENCE TELE-CONFERENCE CONCEPT

The major challenge of the 3DPresence project is to maintain eye contact, gesture awareness, 3D life-sized representations of the remote participants and the feeling of physical presence in a multi-party, multi-user terminal conference system. In order to achieve these objectives, the concept of a shared virtual table is applied. All remote conferees will be rendered based on a predefined shared virtual environment (see examples in Figure 3 left and Figure 4 left). Eye contact and gesture awareness can be created by adapting virtually the 3D perspective and 3D position of all remote conferees on each of the terminal displays. Furthermore, in order to maximize the feeling of physical presence, sophisticated multi-user 3D display technologies will be developed and applied within the 3DPresence project (see Figure 3 right). The concept will be proved by developing a real-time demonstrator prototype system consisting of four 3D videoconferencing stations in Barcelona (Spain), Tel Aviv (Israel), Eindhoven (Netherlands), and Berlin (Germany).

GEOMETRICAL SYSTEM DESIGN

The geometrical design of the proposed tele-conference system is based on the idea of a shared virtual table. This virtual table is supposed to simulate a real conference situation for 3 parties and 6 participants as illustrated in Figure 4 left. Each party has two participants. Figure 4 right illustrates a possible replacement of the remote conferees by displays. Eye contact and gesture awareness will be created by virtually adapting the perspective of the view of all remote conferees to the given shared virtual environment.

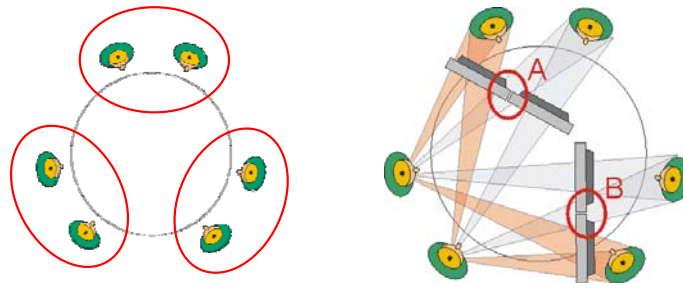


Figure 4 –left: Proposed teleconference setup for 3 parties and 6 participants in total, Constraints on display size, determination of required scaling factor in order to fit the conferee to a given display, right: sample setup using multi-view 3D displays and the viewing cone conflict for display size limitations

Nevertheless, the geometrical design of such virtual environments faces several problems. On one hand, the final position and orientation of the monitors depends on additional constraints. So, in order to create a realistic impression, the remote participants need to appear in life size. As the overall display size is limited a scaling of the remote virtual conferee can be applied as illustrated in top. To compensate this scaling factor the monitor must be moved closer to the local conferee. The condition which needs to be satisfied is that the monitor has to be located within the viewing cone of the local conferee as illustrated in Figure 5.

Another problem arises from the multi-view character of the displays. As illustrated in Figure 4 right, the positions of the displays need to be adapted to the setup of the shared virtual environment. In other words, the position of the display needs to cover totally the viewing cone from the local to the remote participant. In case of two displays each display needs to provide two virtual views. It can be seen that geometrically this constraint cannot be satisfied for all existing viewing cones (regions A, B in Figure 4 right).

Figure 6 illustrates a possible solution for this problem. As a compromise, a slight shift of the corresponding virtual conferee to the real geometrical display position will be introduced. In difference to the ideal setup this solution creates a small replacement error. The shifted conferee will appear at a slightly different position compared to the ideal setup. Further on, as illustrated in the figure, both local users will see the remote conferee at slightly different (to each other) positions. Nevertheless, as the capturing system will be shifted and adapted accordingly, the directionality of gestures as well as the eye contact will not be influenced by this solution.

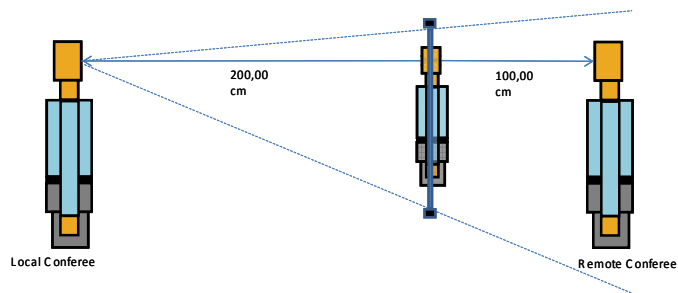


Figure 5 – Constraints on display size, determination of required scaling factor in order to fit the conferee to a given display

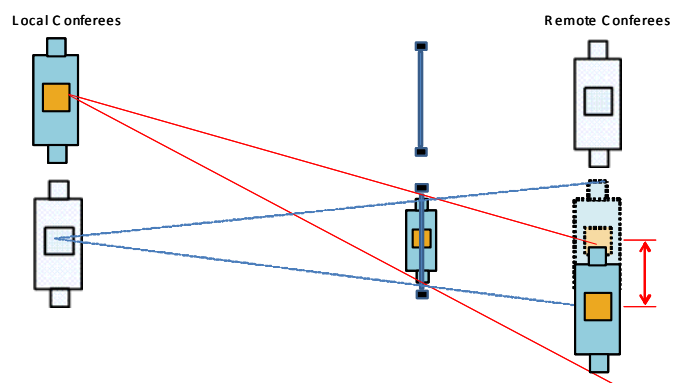


Figure 6 – Sample of by shifting the virtual remote partner to the middle of the display by introducing a small replacement error



Figure 7 – Static image on the novel 2-view 3D prototype display, as seen: left) from the position of the left local participant, right) from the position of the right local participant.

MULTI-VIEW AUTO-STEREOSCOPIC 3D DISPLAY

Current 3D displays are capable to provide multiple stereoscopic views in order to support head motion parallax. Hence, the user is able to perceive a scene in 3D and recognize different perspectives according to his head position. In contrast, the challenge in 3DPresence is to provide stereoscopic viewing for two users and head motion parallax with significantly different perspectives onto the scene. Such display type having multiple perspectives is new. It was recently developed by Philips, building further on previously developed design principles of 3D auto-stereoscopic displays based on slanted lenticular lenses as in [2] and [5].

In Figure 3 left, the conceptual idea of a display with two different viewing cones is presented. The novel multi-view display provides two viewing cones with significantly different perspectives and each of them supports multiple views for 3D. Note, that the geometrical design of these cones has significant impact on the design of the overall demonstrator geometry as the viewing cones of all four displays, related to the four remote conferees, must meet at the correct position. In Figure 3 right the display arrangement and the viewing cones (two per display and eight in total) are shown. The position of the local conferees is marked by a black circle. It can be recognized that the viewing cones have different orientations for different displays compared to the display normal. This is a special feature of the rendering algorithm inside the novel 3D display which allows for a configurable cone rotation angle. Note, that the novel display involves a panel with a resolution of 1920x1080 pixels, and a specifically designed optical sheet with slanted lenticular lenses as in [2] and [5]. The optics allow for the rendering of 15 views distributed across a cone angle as large as 46 degrees. Figure 7 depicts two simultaneously taken photographs of the novel dual view 3D display as seen from the two positions of the two local participants. Clearly, one local participant is being looked at while the other one gets a different view.

CONCEPT OF DEPTH ACQUISITION SYSTEM

In the following the general algorithmic concept of the 3D acquisition and analysis chain will be discussed. Recent technological advances in processing hardware lead to a significantly increased available processing power for consumer PC hardware solutions. Especially high performance (consumer) graphics card solutions allow an algorithmic parallelization of up to 500 or more processing pipes per card¹ for dedicated tasks. In this way, compared to state of the art 3D acquisition chains, such as [2],[3], a new generation of algorithmic processing pipelines can be designed.

¹ NVIDIA GTX 295 with two GPUs

The main idea of this paper is to use multiple dependent or independent parallel algorithmic modules. For 3DPresence, Figure 8 gives an overview of a possible processing chain. As it can be seen, three approaches have been chosen which work in parallel: The volumetric reconstruction module, the disparity estimation modules, and the hand/head tracking modules². The results of these modules are combined afterwards in a data fusion step.

The idea at this point is to develop volume based approaches in a competing sense in relation to stereo matching approaches. The reason is that volume based approaches have very contrary properties and limits compared to stereo based approaches. Especially, if the distance between the increases then the robustness for VH will increase too (within certain limits) but the robustness for stereo matching will decrease. Contrary, small camera baselines will by definition result in a high quality of reconstruction results for stereo matching but a poor quality for volume based approaches.

Another option for the final analysis chain is to use the outcome of one of the modules as an initialization of the others. Again, the idea is to use the advantages of different algorithmic groups with different levels of quality and robustness.

In 3DPresence, the disparity/depth estimation is based on the Hybrid Recursive Matcher (HRM) proposed in [2],[3]. Within 3DPresence promising results could be achieved by combining several narrow and wide baseline stereo matching systems. Figure 11 gives an example for a possible camera setup for different stereo matching pairs. For 3DPresence several combinations of this idea have been tested. The resulting algorithmic chain is shown in Figure 9. The main idea is to calculate consistent depth information in a short baseline system, whose depth resolution is then enhanced by a wide baseline system. The short baseline system is further divided in a vertical and a horizontal system. Exploiting the trilinear constraint [8] further robustness can be achieved in this way. An example is given in Figure 10. The left-hand depth map is based on a small baseline trilinear system. It has has a

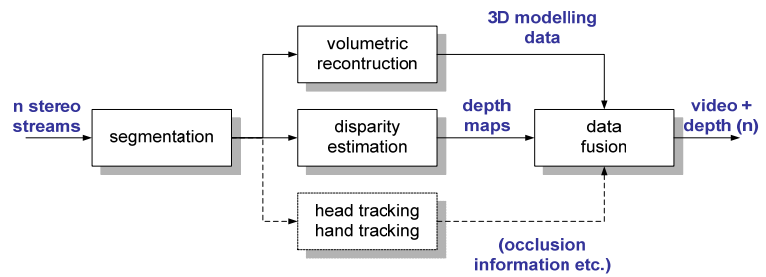


Figure 8 - Concept for the 3D acquisition and depth analysis chain

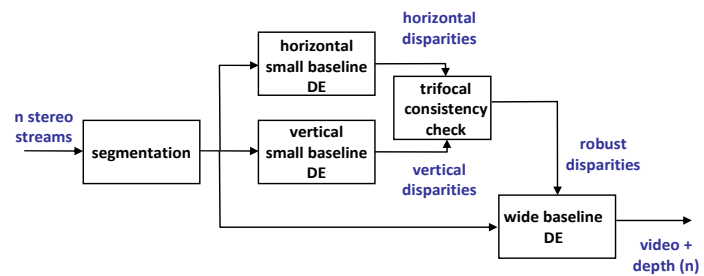


Figure 9: Trifocal short baseline system followed by wide baseline system



Figure 10 – Depth Layer: Small baseline (left), wide baseline refined (right)

² Note, that so far the hand and head tracking module has not been integrated in the 3DPresence analysis chain. It is planned to be integrated at a later step

poor depth resolution but a high robustness. This depth map could be enhanced in a second wide baseline estimation step by keeping the robustness of the small baseline system and adding a higher depth resolution based on the wide baseline module. Note that in total three separate stereo-matchers are used which are parts of a parallel and/or sequential data flow. Again, based on dedicated GPU hardware, these modules run in real-time on a consumer PC.

RESULTS BASED ON FIRST EXPERIMENTAL DEMONSTRATOR PROTOTYPE

Based on initial studies in the 3DPresence project a mock-up system has been installed in order to test different camera and display configurations. In Figure 11 top, a CAD based simulation of the mock-up is shown. A photograph of the installation without the displays is shown in Figure 11 bottom.

The processing of the volumetric approach was performed in real-time on a NVIDIA GTX295 GPU graphics card. In detail, an image based visual hull (VH) approach was implemented. A sample result for a voxel based reconstructed 3D model of a conferee as well as a related depth map representation are shown in Figure 12.

Further on, according to the general algorithmic concept presented in Figure 8 the results of HRM and VH modules were fused. Figure 13 shows the results of this process. They are compared with a pure stereo based approach (HRM) or a pure volume based approach (VH). In row A can be clearly seen, that the achieved depth resolution of the pure VH approach is much higher than that for the pure HRM. The combined version gives here an acceptable compromise. Nevertheless, the pure VH suffers from 'ghost' artefacts (see row A). Here, the HRM clearly outperforms the VH. A combined approach enhances the quality significantly. Further, an improvement in reconstruction quality can be achieved for scene details, such as the fingers in row B. The right-hand column illustrates the rendered view which was adapted to the geometry of the shared virtual table environment of the tele-conference system. Figure 14 demonstrates the virtual eye contact which was achieved by rendering the virtual camera according to the shared table setup.

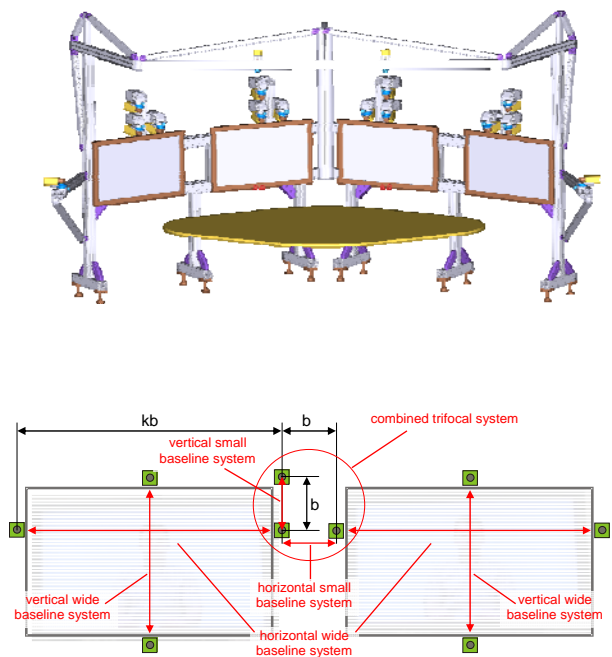


Figure 11 – top: Teleconferencing demonstrator concept including ceiling and side cameras for volumetric processing (top), experimental test mockup (bottom), possible camera arrangements (middle)



Figure 12 – Visual Hull result of a conferee, left: 3D model, middle: depth map representation, right: segmented input images

CONCLUSION

We tackled the topic of new generation immersive multi-user 3D video conference systems. Based on experiences and results of the European FP7 research project 3DPresence we have discussed general problems and solutions in the geometrical design of shared virtual table environments. Further on, we introduced first research results of a new generation of multi-view auto-stereoscopic displays and demonstrated an integration scheme for tele-conference systems. Finally, we proposed a powerful new generation real-time 3D acquisition chain which benefits from recent technological advances in the parallel processing hardware development.

ACKNOWLEDGEMENTS

This work is part of the FP7 project "3DPresence", Proposal no.: FP7-215269, which is funded by the European Commission.

REFERENCES

1. D. Nguyen, J. Canny, "MultiView: spatially faithful group video conferencing", *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, pp. 799-808, Portland, Oregon, USA, April 2005
2. O. Schreer, P. Kauff, "An Immersive 3D Video-Conferencing System Using Shared Virtual Team User Environments", *Proc. of ACM Collaborative Virtual Environments (CVE 2002)*, pp.105-112, Bonn, Germany, October 2002.
3. Atzpadin, N., Kauff, P. and Schreer, O.: Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing, *IEEE Transactions on Circuits and Systems for Video Technology*, special Issue on Immersive Telecommunications, vol. 14, no. 3, pp. 321-334, January 2004.
4. C. van Berkel, Image Preparation for 3D-LCD, *Proc. SPIE 1999*, Vol. 3639;
5. C. van Berkel, J.A. Clarke, Characterisation and Optimisation of 3D-LCD Module Design, *Proc. SPIE 1997*, Vol. 3012, p.179
6. B. Barenbrug, "3D throughout the video chain," in *Proceedings of Int. Congress of Imaging Science*, pp. 366–369, 2006.
7. W.H.A. Bruls, C. Varekamp, R. Klein Gunnewiek, B. Barenbrug, A. Bourge, "Enabling introduction of stereoscopic 3D video: compression standards, displays and content generation," in *Proc. of Int. Conf. on Image Processing*, pp. 89–92, 2007.
8. J. Mulligan, V. Isler, K. Daniilidis, "Trinocular Stereo: A Real-Time Algorithm and its Evaluation" *Int. Journal of Computer Vision*, 47, pp.51-61, April 2002.

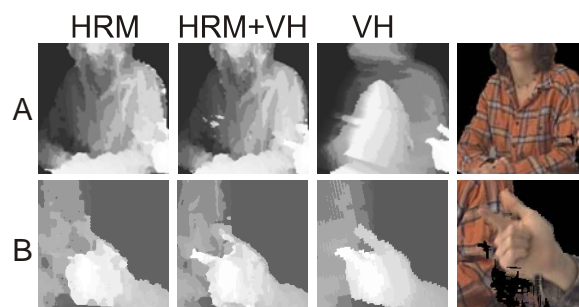


Figure 13 – Estimated depth maps based on separate and combined HRM and VH approaches for critical regions



Figure 14 – Eye contact for virtual camera based on the estimated depth map and the rendered virtual view