3DPRESENCE – A SYSTEM CONCEPT FOR MULTI-USER AND MULTI-PARTY IMMERSIVE 3D VIDEOCONFERENCING

O. Schreer, I. Feldmann, N. Atzpadin, P. Eisert, P. Kauff, H.J.W. Belt*

Fraunhofer Institute for Telecommunications/Heinrich-Hertz Institut, Berlin, Germany,
{Oliver.Schreer, Ingo.Feldmann, Nicole.Atzpadin, Peter.Eisert, Peter.Kauff}@hhi.fraunhofer.de
*Philips Research, Eindhoven, Netherlands, harm.belt@philips.com

Keywords: Immersive telepresence, 3D videoconferencing, multi-view video.

Abstract

Traditional set-top camera video-conferencing systems still fail to meet the 'telepresence challenge' of providing a viable alternative for physical business travel, which is nowadays characterized by unacceptable delays, costs, inconvenience, and an increasingly large ecological footprint. Even recent high-end commercial solutions, while partially removing some of these traditional shortcomings, still present the problems of not scaling easily, expensive implementations, not utilizing 3D life-sized representations of the remote participants and addressing only eye contact and gesture-based interactions in very limited ways. The European FP7 project 3DPresence will develop a multi-party, high-end 3D videoconferencing concept that will tackle the problem of transmitting the feeling of physical presence in real-time to multiple remote locations in a transparent and natural way. In this paper, we present an overall concept, which includes the geometrical design of the whole prototype demonstrator, the arrangement of the cameras and displays and the general multi-view video analysis chain. The driving force behind the design strategy is to fulfil the requirements of a novel 3D immersive videoconferencing system, including directional eye gaze and gesture awareness.

1 Introduction

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. 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 [10], which provides nearly eye contact supported by cameras mounted on top of the displays. In this approach, no 3D processing is performed. Due to the close distance of the cameras to the displayed head of the remote conferees and the far distance of the local conferees to the display, the displacement angle regarding the viewing directions can be neglected.

Figure 1: State of the art telepresence system by CISCO (top, left), the Polycom TPX system (top, right) and the HP Halo Telepresence system (bottom)

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. 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 2). The concept will be proved by developing a real-time demonstrator prototype system consisting of three 3D videoconferencing stations in different European countries.
In this paper, we are focusing on the general system concept including the geometrical design, the multi-view camera configuration, arrangement of the 3D displays and the overall algorithmic analysis chain. This paper covers general aspects of algorithmic analysis modules and cameras/hardware. The restrictions by real-time processing, coding and virtual view rendering are included in the proposed general system design.
The outline of the paper is as follows. In section 2, the shared table concept for a multi-user, multi party 3D video conferencing system is presented. The specifics of a novel multi-user 3D display and the impact on the system design is discussed in section 3. Then, the multi-view camera system is presented which will allow a robust and accurate depth estimation of the scene. The paper is summarized with a conclusion.

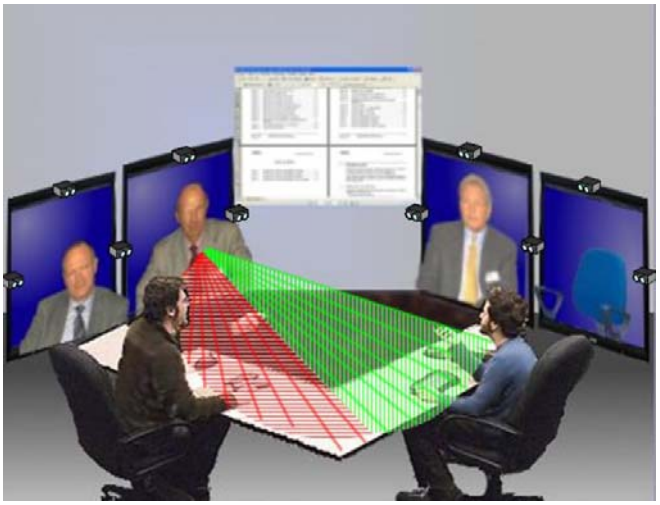

Figure 2: 3DPresence multi-party videoconferencing concept

2 The multi-party multi-user shared table concept

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 3 left. Each party has two participants and all conferees are sitting around a common virtual table. Figure 3 right illustrates the 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.
In order to obtain realistic natural views of the remote participants novel 3D displays will be realized which provide separate views from different perspectives for each of the local participants (indicated by the red and green regions in Figure 2) and in addition stereoscopic viewing of the remote conferees. This new type of 3D display will be developed by one of the industrial partners, namely Philips.

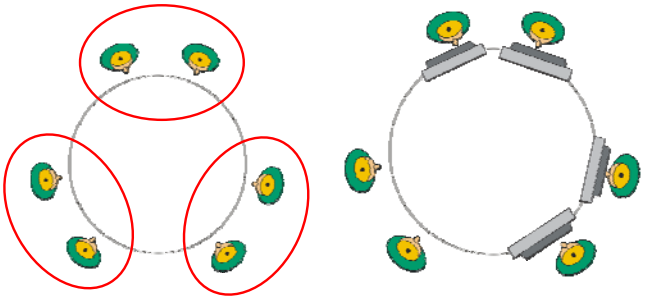

Figure 3: Real multi-party and multi-user scenario (left), virtual table environment with 2 local and 4 remote participants (right)

2.1 Life size

In 3D-Presence, due to display production limitations that we face today, the overall display size is limited to 42''. To still achieve a sense of life-size remote conferees, the video of the remote conferee needs to be scaled and the monitor must be moved closer to the local conferee, see Figure 4. 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 4. If the sizes of monitors and remote conferees are known, the exact scaling factor can be determined.

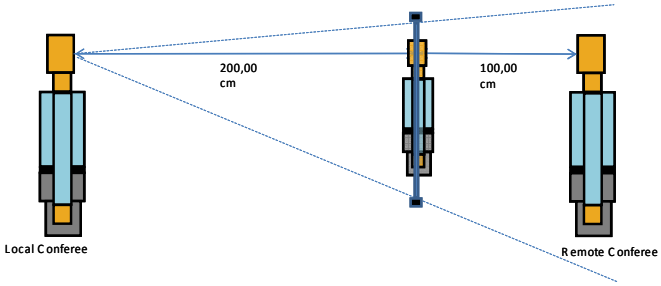

Figure 4: Constraints on display size, determination of required scaling factor in order to fit the conferee to a given display

2.2 Eye contact and gesture awareness

In general, a bi-directional eye contact between a local and a remote person is not possible using the conventional straight forward video conference approach of a camera on top of each screen (local and remote). The conflict results from the displacement angle between the camera and the eyes of the virtual person on the screen. As the local conferee is looking at the person on the screen, he/she is not gazing at the camera. Hence the remote conferee cannot perceive eye contact.

This problem will be handled by creating a virtual camera exactly at that position inside the display where the eyes of the remote conferee are rendered. This approach is known from the literature [12].
In terms of gesture awareness for a multi-user systems, it means further, that a local conferee must recognize as well the direction of the gestures of all remote users. Let's imagine a local conferee that is looking at a remote partner. If the remote partner is pointing to any other partner of the conference, then the local partner must recognize this situation as a gesture with the correct direction. Otherwise, the natural feeling of the conference will loose a lot of its value. The gesture awareness can be represented if the virtual conferees are sitting at the same position as the real participants at the real round table.
In the next section, some further constraints of the novel 3D multi-view display are discussed and a proposal for a multi-user multi-party display configuration is presented.

3 Specifics of a novel multi-view 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 will be developed by Philips, building further on previously developed design principles of 3D auto-stereoscopic displays based on slanted lenticular lenses as in [13] and [14]. In Figure 2, the conceptual idea of a display with two different viewing cones is presented. The envisaged approach is to develop a novel multi-view 3D display which provides two viewing cones with significantly different perspectives and each of them supporting multiple views.
Due to this novel display design, the viewing cones of four displays, related to the four remote conferees, must meet at the correct position of the two local conferees. This has significant impact on the overall design of the tele-presence system as the overlapping area of all viewing cones related to one local conferee must be as large as possible. The larger the area the better the local conferee can move and change its viewing position. In Figure 5, the display arrangement and the viewing cones (two per display and eight in total) are shown. The overlapping regions of the viewing cones for each local conferee are assigned by dashed lines. 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 novel 3D display which is necessary to obtain a large overlapping region.

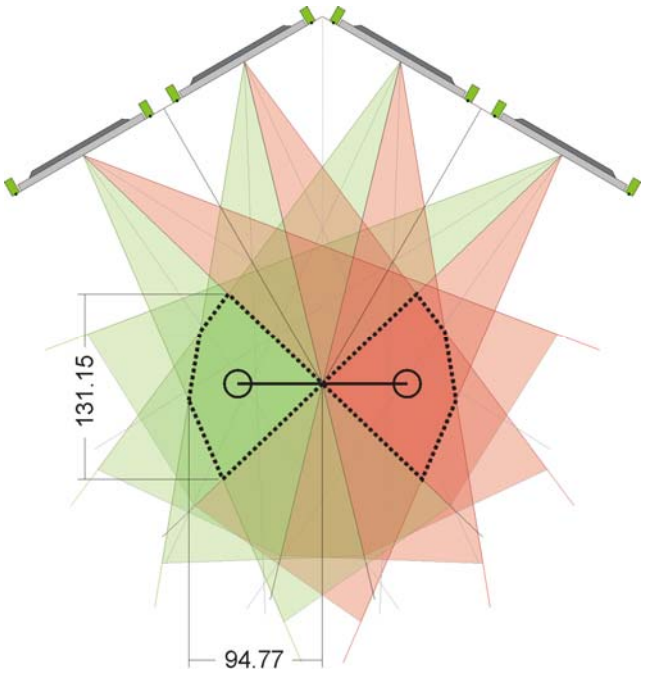

Figure 5: 8 viewing cones meeting the two local conferees

4 Multi-view camera system

The required input format for the 3D displays is double 'video-plus-depth', one video-plus-depth per shown perspective of the remote conferee. The format 'video-plus-depth' is well-known from the literature. For example, it has been studied thoroughly in the European FP5 project ATTEST [8]. This input format has been standardized by MPEG and is nowadays applied in many commercial 3D displays. This format allows computation of multiple perspectives of the scene by the rendering algorithm from a central image and a depth map. Hence, a robust and accurate depth estimation in real-time is required. Plenty of algorithms have been proposed in the past on real-time disparity analysis and a few approaches consider also view synthesis by using disparities for provision of eye contact. The European FP5 project VIRTUE tackled this issue within a complete system framework [11]. In [5] and [15], real-time approaches are presented based on a single stereo camera system. A real-time algorithm using three cameras has been presented by [9].
The design of the multi-view camera system is influenced by the following different concepts and approaches.

4.1 Disparity estimation concept

It is known from the literature that for disparity/depth estimation a dependency exists between the baseline of the camera pair, the image resolution, and the number and optimal position of related depth layers, i.e. the depth resolution [3], [7]. Considering the distance between two cameras, i.e. their baseline, the following fundamental properties can be observed. A wide baseline stereo camera system provides a high depth resolution but due to the more different perspectives the robustness of the estimation

decreases. On other hand, a small baseline stereo system provides in general disparities of better quality but the depth resolution decreases. Hence, a combination of both systems is considered. In Figure 6, the disparity estimation results of the original image on the left are presented for a small and wide baseline configuration. It can be recognized that the robustness and coverage are higher in the small baseline case (middle) but the disparity resolution is coarser. Vice versa for the wide baseline case, the robustness decreases but the depth resolution increases.

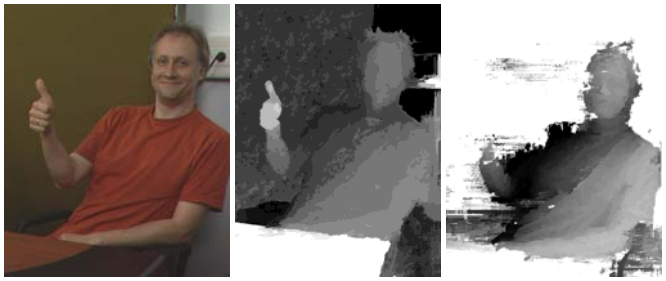

Figure 6: Visualization of depth layers for small and wide baseline systems: 22cm (middle) and 124cm (right)

Furthermore, robustness can be increased in any case by using a third camera and exploiting the trilinear constraint [9]. Due to the third view a cross check between pairs of disparities can be performed and unreliable disparities can be discarded. The number of valid disparities will decrease but the remaining disparities are of increased accuracy and reliability. Following these observations, a multi-view camera setup as presented in Figure 7 will be used, which allows to take advantage of three camera sub-systems.
Two wide baseline systems per display are mounted in horizontal and vertical direction. Between both displays a trifocal small baseline system is installed in order to provide an initial robust estimate.

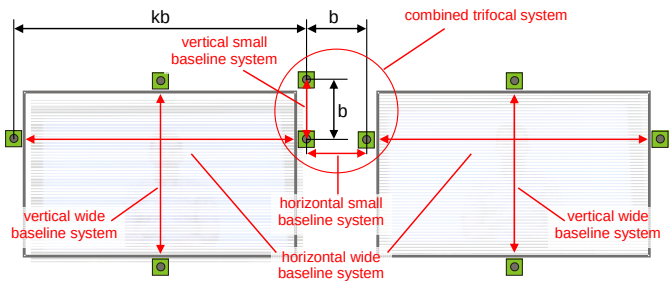

Figure 7: Camera arrangement for a single remote site

The main idea is to start with robust disparity estimation on small baseline systems of the trifocal camera sub-system. The disparities of the horizontal and vertical small baseline system can be further checked by exploiting the trifocal constraint. The resulting robust disparities will then be used as input information for the wide baseline systems. The latter one allows estimating at higher depth resolution. This is shown exemplarily in Figure 6.

The advantages of each sub-system summarize as follows:

Small baseline system: robust disparity estimation
Trifocal system: consistency check across three views
Wide baseline system: increased disparity resolution

It has to be noted that the distance between cameras is based on a multiple of the small baseline *b*. Hence, the estimated disparities of the horizontal and vertical camera pairs in the trifocal system can be compared directly without parameter adaption.
As presented in Figure 7, the trifocal camera sub-system is arranged in L-shape. The main advantage is that for the lower two cameras a disparity estimation can be performed along the horizontal scan line, whereas for the two left-hand cameras a disparity estimation in vertical direction can be applied. Due to the same distance between the horizontal cameras and the vertical cameras, horizontal and vertical disparities can be compared straight away, which increases the computational efficiency.
In this setup at least the cameras of the wide baseline system need to be mounted convergent towards the conferees. Hence, as we assume a fully calibrated multi-camera system all camera images will be rectified pair wise in order to exploit advantages of disparity estimation along the scan line.
In Figure 8, sample views are depicted for all cameras of the wide baseline and trifocal camera system.

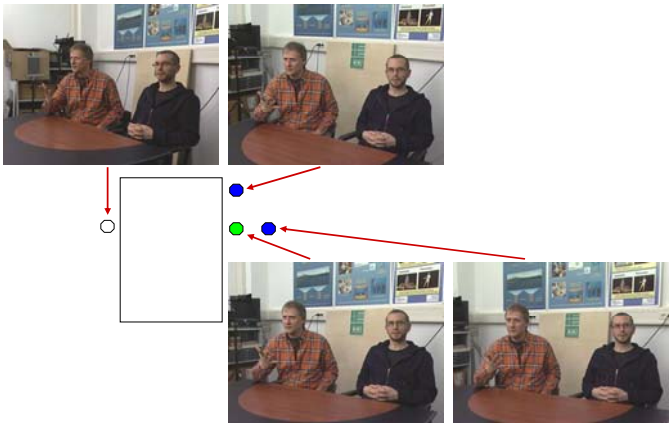

Figure 8: Sample views of the wide baseline system and the trifocal system

4.2 Hybrid-recursive matching

The estimation of suitable depth maps from stereo or multi-camera systems is certainly one of the most challenging tasks in the given context. The disparity estimation itself is based on the Hybrid-Recursive-Matching (HRM) algorithm as described in [2]. The main idea of the hybrid recursive stereo matching algorithm is to unite the advantages of block-recursive disparity matching and pixel-recursive optical flow estimation in one common scheme. The block-recursive part assumes that depth does not change significantly from one

image to the next and that depth is nearly the same in the local neighbourhood. Obviously this assumption cannot be fulfilled in all image areas - especially not in areas with high motion and at depth discontinuities. To update the results of the block-recursive stage in these areas, the pixel recursive stage calculates the optical flow by analyzing gradients and grey value differences.

In more detail, the structure of the whole algorithm can be outlined in three subsequent processing steps (see Figure 9):

1. three candidate vectors are evaluated for the current block position by recursive block matching;
2. the candidate vector with the best result is chosen as the start vector for the pixel-recursive algorithm, which yields an update vector;
3. the final vector is obtained by testing if the update vector from the pixel recursive stage is of higher quality than the start vector from the block-recursive one.

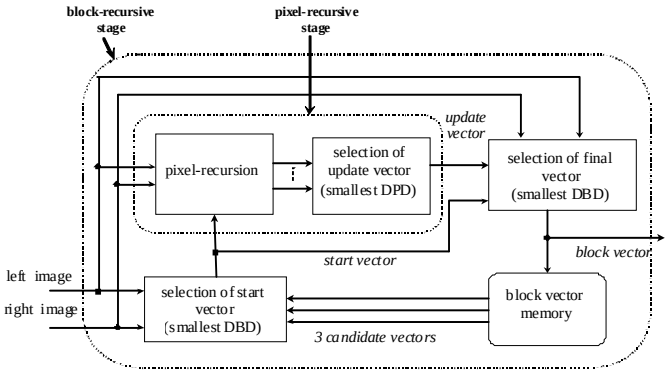

Figure 9: Outline of the HRM algorithm

Due to its recursive structure the HRM algorithm produces extremely smooth and temporally consistent "per-pixel" disparity maps. Hence, they contain highly redundant information and have almost no random noise – a property that is essential for efficient coding of depth maps. As any matching algorithm, HRM usually generates failures and mismatches in critical image areas. These mismatches are detected and corrected by sophisticated post-processing. One criterion for detecting mismatches is a confidence measure, which is directly derived from the normalized cross-correlation used by HRM. If the confidence value is below a critical threshold, the corresponding disparity is removed from the map. Furthermore, as the HRM estimates, independently from each other, two disparity maps for each rectified image pair (one from right to left and, vice versa, one from left to right), these two maps can be used to prove the consistency of the disparities. Usually, there are two reasons for the detected mismatches: ambiguities during matching (homogeneities, similarities, periodicities, etc.) or occluded areas. These two failure categories have completely different origins. Ambiguities are caused by an ill-posed matching problem; i.e., point-correspondences exist but could not be found correctly by the matcher. In contrast, point correspondences do not exist at all in occluded areas and cannot be matched on principle therefore. Thus, it is the target of further processing to distinguish between these two sources of fault. For this purpose, the missing disparity values are first reconstructed by using segmentation-driven interpolation.

In Figure 10, preliminary results of disparity estimation for the trifocal camera setup are presented. The left image shows the disparity estimation of the vertical small baseline stereo system, whereas the right image is dedicated to the horizontal small baseline stereo system. Almost the complete scene has been estimated correctly but with limited resolution in depth.

In Figure 11, the result of disparity estimation of the horizontal wide baseline stereo system is shown. The disparity map contains significant artefacts but the resolution in depth is increased.

The main challenge is now to combine the different disparity estimation results in order to achieve complete and robust disparity maps at increased resolution.

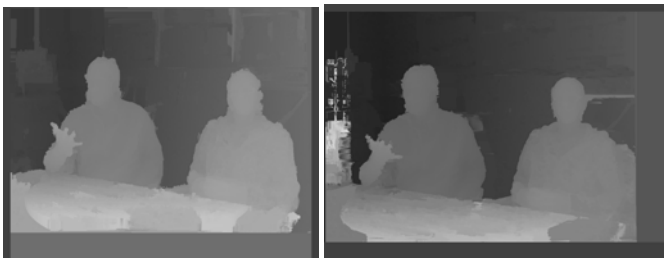

Figure 10: Disparity maps of the trifocal system: vertical baseline (left) and horizontal baseline (right)

Figure 11: Disparity map of the horizontal wide baseline system

It has to be noted that the combination of individual disparity estimation results is in progress and part of future work. We expect significant improvement by this approach whereas the way how to combine the individual results is one of the central research issues of the 3DPresence project.

4.3 Multi-view video analysis chain

The overall multi-view video analysis chain does not only consider disparity/depth estimation. As known from literature, a volumetric reconstruction based on visual hull computation can provide helpful information of the object boundaries [4]. This information is intended to be exploited as well and being combined with the disparity estimation module. The video conferencing scenario fits very well to this approach as individual foreground objects, i.e. the local conferees and a deterministic background is available.

In contrast to disparity estimation, the visual hull approach requires as many as possible significantly different

perspectives onto the scene. As shown in Figure 5, several cameras are available in the proposed demonstrator, providing views from a large range of different perspectives with a maximum of approximately 90 degrees between the two outmost cameras.
The complete multi-view video analysis chain is presented in Figure 12. Initially, foreground-background segmentation on the captured input data will be performed. The subsequent disparity estimation will benefit from this pre-processing step as all the processing can be restricted to the foreground object, which saves computational effort. By using the segmented foreground-object, a visual hull algorithm is applied in order to reconstruct the volume of the object. In addition to it, a head and hand tracking algorithm is applied to the segmented object in order to get additional helpful information on the human motion. Hand tracking and segmentation of hands will be exploited to detect overlap of hands and the body [1]. Even in these regions, disparity estimation as well as view rendering will benefit from this information. The head tracking will be used for precise view point adaptation according to the conferees head position.
The results of the volumetric reconstruction, the disparity estimation and the hand/head tracking are combined afterwards in a data fusion step.

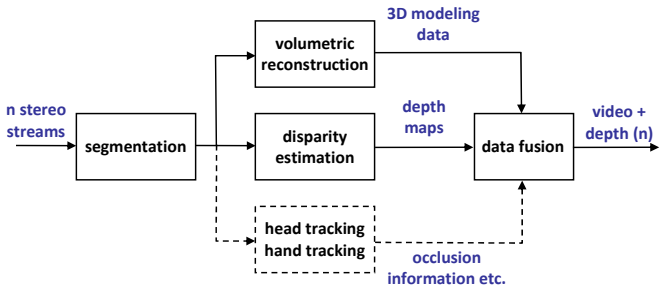

Figure 12: 3D video analysis chain

In Figure 13, three manually segmented camera views of the conferee are depicted. This segmentation information as well as the calibration information of the camera system shown in Figure 15 is used to perform a volumetric reconstruction by applying a voxel colouring reconstruction technique [6].

Figure 13: Manually segmented camera views

Due to the availability of views from the person's side, a significant increase in shape information can be exploited. Preliminary results of the reconstruction are presented in Figure 14.

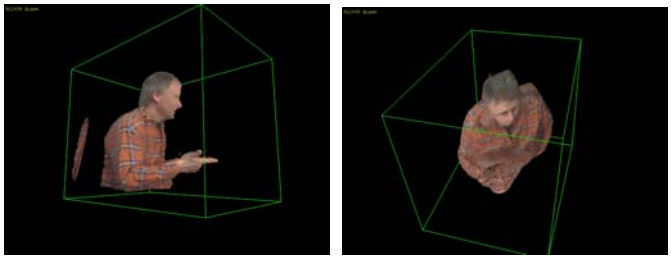

Figure 14: Preliminary results of voxel colouring reconstruction

Although also colour consistency is considered, the main source of information is the silhouette of the person. Since this information is orthogonal to the depth estimates from the small and wide baseline stereo matching, both techniques can beneficially be combined leading to a robust solution with one method complementing the other in regions that are difficult to handle for a single approach.

5 The demonstrator mock-up

Based on an initial study in the 3DPresence project a mock-up system has been installed in order to test different camera and display configurations as well as to capture test sequences for algorithm development. In Figure 15, a CAD based simulation of the mock-up is shown. It has to satisfy many constraints.

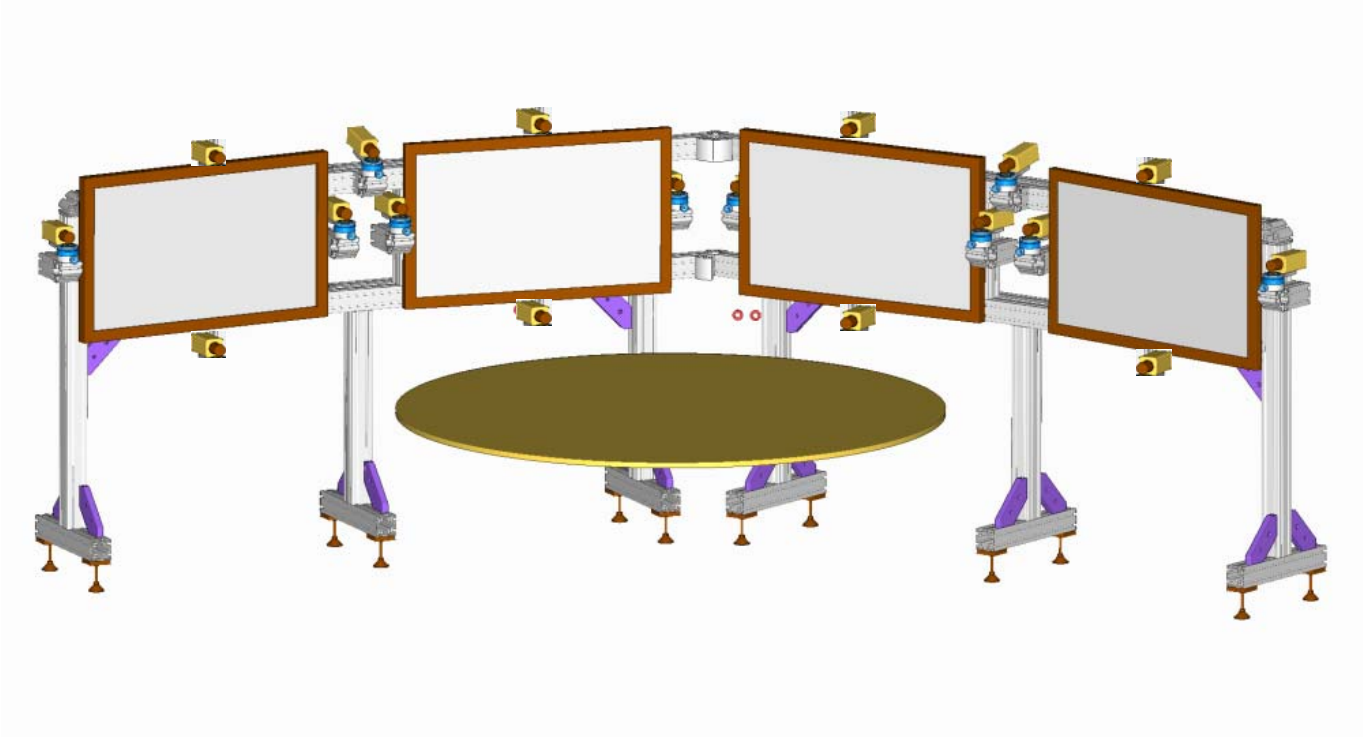

Figure 15: Drawing of the mock-up system

Based on the theoretical design, a real mock-up system has been installed fulfilling all of the aforementioned constraints. A photograph of the installation without the displays is shown in Figure 16. A very flexible, metallic framework has been designed, which allows a camera mounting at almost arbitrary positions. Due to this, it is possible to test different arrangements of stereo and trifocal sub-systems.

As the rendering quality of the virtual view needs to be evaluated, a ground truth camera is mounted exactly at the position where the local conferee is looking at the display into the eyes of the remote conferee. This ground truth allows evaluating the rendering quality in terms of producing correct eye contact, which is one of the central objectives in this project.

Figure 16: First version of the 3DPresence mock-up

6 Conclusion

We have presented a display and multi-view camera setup that strictly follows the requests and requirements of a novel multi-user, multi-party 3D videoconferencing system. The design conditions have been presented which results from the novel 3D display. Due to two independent viewing cones per display, the position and orientation of the displays required optimization in terms of maximum overlap of all viewing cones dedicated to one of the local conferees.

A multi-view camera system has been presented that consists of three sub-systems such as small baseline, wide baseline and trifocal camera system. The main idea is to take advantage of each sub-system in order to maximize the disparity estimation result. In addition to that, a completely different reconstruction method is incorporated in the overall system design namely volumetric reconstruction. As this approach benefits from most different views, advantage regarding fusion of volumetric reconstruction and disparity estimation is expected.

For all different sub-systems for disparity estimation as well as for volumetric reconstruction, preliminary results have been presented. The major research task will be to investigate different combinations of intermediate results within the multi-view analysis chain. In this context, real-time performance plays an important role. Therefore, a trade-off between quality of depth estimation and overall performance needs to be identified.

Acknowledgements

This work is part of the FP7 project ”3DPresence”, Proposal no.: FP7-215269, which is funded by the European Commission. The results presented in this paper are based on intensive collaboration between the colleagues at Philips Research Eindhoven and Fraunhofer HHI, Image Processing Department.

References

[1] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, O. Schreer, “Vision-Based Skin-Colour Segmentation of Moving Hands for Real-Time Applications”, *Proc. of 1st European Conf. on Visual Media Production (CVMP)*, London, United Kingdom, March 2004.

[2] Atzpadin, P. Kauff, O. Schreer, “Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing”, *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications*, pp. 321-334, Vol. 14, No. 3, January 2004.

[3] J.-X. Chai, X. Tong, S.C. Chan, H.-Y. Shum. ”Plenoptic Sampling”, *Proc. of SIGGRAPH 2000*, pp.307-318, New Orleans, LA, USA, July 2000.

[4] K.M. Cheung, T. Kanade, J. Bouguet, M. Holler, “A real time system for robust 3D voxel reconstruction of human motions” *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2000)*, pp.714-720, Vol. 2, June 2000.

[5] A. Criminisi, J. Shotton, A. Blake, P.H. Torr, “Gaze Manipulation for One-to-one Teleconferencing”, *Proc. of the 9th IEEE Int. Conf. on Computer Vision*, Vol.2, pp.191, Washington, DC, October 2003.

[6] P. Eisert, E. Steinbach, B. Girod,"Multi-hypothesis, Volumetric Reconstruction of 3-D Objects from Multiple Calibrated Camera Views," *Proc. Int. Conf. on Acoustics Speech and Signal Processing (ICASSP '99)*, pp. 3509-3512, Phoenix, March 1999.

[7] I. Feldmann, U. Gölz, P. Kauff, “Navigation Dependent Nonlinear Depth Scaling”, *Proc. of 23rd Int. Picture Coding Symposium*, St. Malo, France, pp. 387-390, April 2003.

[8] C. Fehn, “A 3D-TV System Based on Video Plus Depth Information”, *Proc. of 37th Asilomar Conference on Signals, Systems, and Computers*, pp.1529-1533, Pacific Grove, CA, USA, November 2003.

[9] 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.

[10] 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.

[11] O. Schreer, N. Brandenburg, S. Askar, M. Trucco, “A Virtual 3D Video-Conference System Providing Semi-Immersive Telepresence: A Real-Time Solution in Hardware and Software”, *Proc. of eBusiness and eWork* 2001, pp.184-190, Venice, Italy, October 2001.

[12] 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.

[13] C. van Berkel, Image Preparation for 3D-LCD, Proc. SPIE 1999, Vol. 3639;

[14] C. van Berkel, J.A. Clarke, Characterisation and Optimisation of 3D-LCD Module Design, Proc. SPIE 1997, Vol. 3012, p.179

[15] R. Yang, Z. Zhang, “Eye Gaze Correction with Stereovision for Video-Teleconferencing”, pp.479-494, *7th European Conf. on Computer Vision*, Copenhagen, Denmark, May 2002.