Multi-baseline Disparity Fusion for Immersive Videoconferencing

Oliver Schreer
Fraunhofer Institute for Telecommunications/ Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin, Germany
++49 30 31002-620
Oliver.Schreer@fraunhofer.hhi.de

Nicole Atzpadin
Fraunhofer Institute for Telecommunications/ Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin, Germany
++49 30 31002-404
Nicole.Atzpadin@fraunhofer.hhi.de

Ingo Feldmann
Fraunhofer Institute for Telecommunications/ Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin, Germany
++49 30 31002-290
Ingo.Feldmann@fraunhofer.hhi.de

Peter Kauff
Fraunhofer Institute for Telecommunications/ Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin, Germany
++49 30 31002-615
Peter.Kauff@fraunhofer.hhi.de

ABSTRACT

The European FP7 project 3DPresence is developing a multi-party, high-end 3D videoconferencing concept that tackles the problem of transmitting the feeling of physical presence in real-time to multiple remote locations in a transparent and natural way. 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. One of many challenges in this project is to calculate depth information for many different views in order to synthesise novel views to provide eye contact. In this paper, we present a multi-baseline disparity fusion scheme for improved real-time disparity map estimation. The advantages and disadvantages of different configurations are discussed and theoretical considerations are presented regarding disparity resolution and baseline. These observations together with experimental investigations lead to a multi-baseline configuration that allows taking advantage of small and wide baseline stereo camera as well as trifocal camera configurations.

Keywords

Immersive telepresence, 3D videoconferencing, multi-view video, disparity analysis.

1. INTRODUCTION

Recent high-end commercial solutions such as Cisco's TelePresence, 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 [1], 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.

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. 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 a disparity estimation framework, which takes advantage of different types of camera configurations within a multi-view camera system, namely small and wide baseline stereo camera systems. The outline of the paper is as follows. In section 2, the system concept is briefly presented. The multi-view approach is discussed in the next section, whereas the theoretical constraints of disparity/depth estimation are derived. A short outline of the disparity estimation algorithm is

Immerscom 2009, May 27-29, 2009, Berkley, USA.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208

given and the concept of disparity fusion is presented. In section 4, experimental results are provided in order to show the proof of concept. The paper is summarized with a conclusion.

2. SYSTEM CONCEPT

The demonstration prototype consists of a three-way videoconferencing system for two participants at each site (see Figure 1). Each of the remote participants is displayed on a separate 3D auto-stereoscopic display. A significant difference compared to currently available multi-view displays is the following: As the two local participants are looking at one remote participant from a remarkable different perspective, a novel multi-view display has been developed which provides two different viewing cones, one for each of the local participant. These viewing cones are illustrated in Figure 1 with the red and green lines. In each of the viewing cones, the local participant perceives its correct perspective of the remote participant. Each viewing cone consists of seven different views to support stereoscopic viewing and head motion parallax.

Figure 1: 3DPresence multi-party videoconferencing concept

As mentioned before, one of the objectives is to support true eye contact. This is possible by rendering novel views from a multi-view camera setup, as long as descent depth information is available. In Figure 2, the multi-view camera configuration for a single display is depicted. It consists of a horizontal and a vertical baseline stereo system forming together a trifocal camera configuration. In addition, a vertical wide baseline system is mounted by using one camera of the trifocal system. In the next section, we discuss in more detail the properties of the presented configuration.

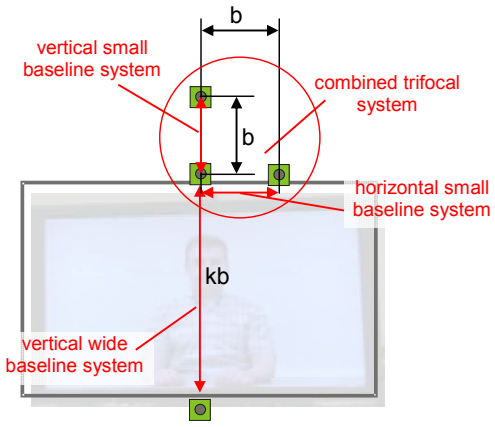

Figure 2: Multi-view camera configuration

3. MULTI-BASELINE APPROACH

The required input format for the 3D displays is double 'video-plus-depth', one video-plus-depth per shown perspective of the remote participant. The format 'video-plus-depth' is well-known from the literature. For example, it has been studied thoroughly in the European FP5 project ATTEST [2]. 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 [3]. In [4] and [5], real-time approaches are presented based on a single stereo camera system. A real-time algorithm using three cameras has been presented by [6]. The design of the multi-view camera system is influenced by the following different concepts and approaches.

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.

Furthermore, robustness can be increased in any case by using a third camera and exploiting the trilinear constraint [6]. 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 better accuracy and reliability. The invalid disparities can be inter- or extrapolated by a sophisticated post-processing step.

Following these observations, a multi-view camera setup as presented in Figure 2 will be used, which allows to take advantage of three camera sub-systems. 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

In the next sub-section, the fusion approach for these different types of camera configurations is presented in order to exploit the advantages of each.

3.1 Disparity fusion

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 system. The latter one allows estimating at higher depth resolution.

In Figure 3, the overall disparity fusion scheme is depicted. At first, the n-video streams of the camera configuration in Figure 2

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208

are segmented in order to extract the local participant from the background. The reason is that only the participant will be rendered from a novel perspective. Beside this, the computational effort decreases significantly because the participants cover only one third of the whole image region. After that, two individual stereo analysis processes will be started by using the hybrid recursive matching algorithm explained in section 3.3. Robust disparity estimation is performed on the horizontal and vertical small baseline systems including a trifocal consistency check. The resulting disparities are very robust in terms of their spatial and temporal consistency but the disparity resolution is low. Therefore, a wide baseline stereo analysis is performed for the cameras on top and bottom of the display. Due to the larger perspective distortion and occlusions, this analysis result is expected to be less accurate. Therefore the consistent and robust results after trifocal consistency check are then fed into this stereo analysis process. The robust disparities may be considered as additional candidates within the hybrid recursive matching scheme as described in section 3.3 and may also be used for post-processing.

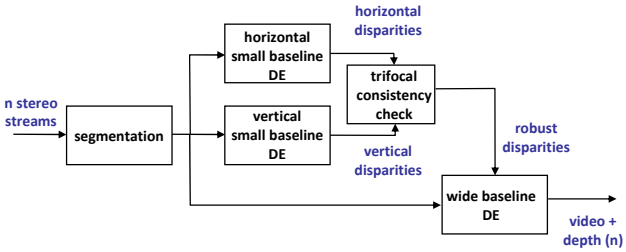

Figure 3: Disparity fusion approach

3.2 Theoretical constraints

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 [7], [8]. It can be used to determine the minimal required baseline for a given expected resolution of depth layers.

For the available camera configuration and the envisaged video conferencing scenario, the relationship between baseline, disparity resolution and depth range has been calculated. In Figure 4, the total available depth resolution was determined for the representation of the whole conferee assuming a depth range of 180mm at a distance of 120mm. It can be recognized that the number of depth layers is smaller for smaller baseline systems.

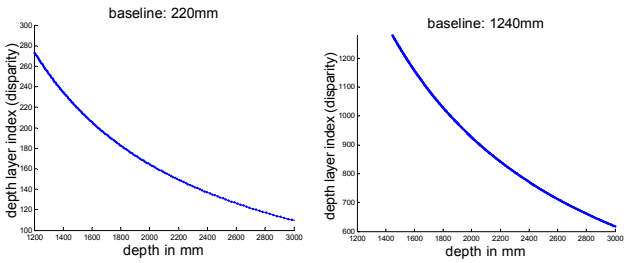

Figure 4: Depth layer for small and wide baseline for a coverage of the whole conferee's depth, range: 1.80m at a distance of 1.20m

Figure 5 shows a part of these graphs, which exemplarily represents the depth range of the conferee's nose (20mm at a distance of 220mm). It can be seen, that for a narrow baseline of 220mm just two depth layers exist for the current configuration. In contrast, a wider baseline of 124mm will result in 7 depth layers for this part of the face. In other words, the nose of the conferee will be represented just by 2 depth layers for the narrow baseline system.

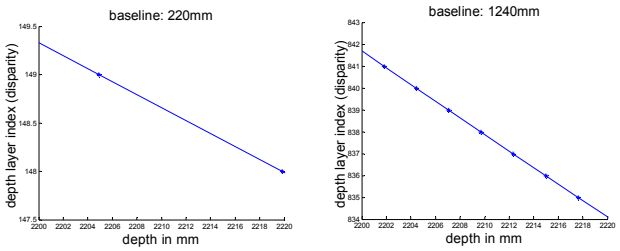

Figure 5: Depth layer for small and wide baseline for a coverage of the conferee's nose, depth range: 2cm at a distance of 2.20m

3.3 Hybrid-recursive matching

The estimation of suitable depth maps from stereo or multi-view 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 [9]. 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 6):

1. Three candidate vectors are evaluated for the current block position by recursive block matching (An additional candidate vector is defined if additional depth information from other data sources is provided.)
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.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208

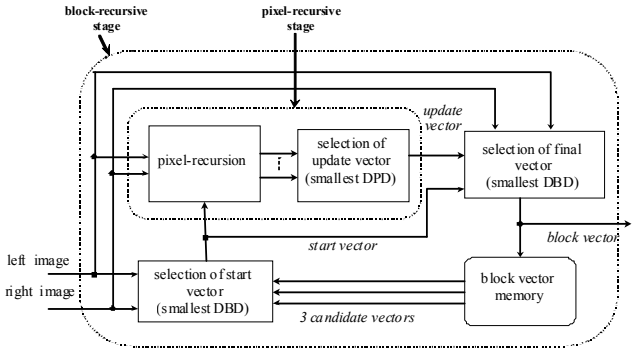

Figure 6: Outline of the HRM algorithm

The idea of the block recursion is to use information of both, the previous image and the spatial neighbourhood. This kind of recursion forces temporal and spatial consistency and it additionally reduces the local search range to a few pixel as known from many other matching algorithms. Usually, calculation of three matching scores per considered pixel is fully sufficient to achieve results comparable to a full search method. In our algorithm these three matching scores are calculated by using three candidates, which are defined by disparities from the previous and the current image. The following spatial and temporal candidates are tested for this purpose:

- A horizontal predecessor, taken from the left or right position in the actual frame.
- A vertical predecessor, taken from the bottom or top position in the actual frame.
- A temporal predecessor, taken from the same position in the previous frame.
- An additional vector is tested if external disparities from other data sources are available

If additional disparity information is provided by a short baseline system the disparities have to be adapted to the wide baseline system due to the higher depth resolution. One possibility is to define a small search range around the disparity delivered by the short baseline system and to find the best match in this search area. The other possibility is to perform pixel recursion with the external disparity as start vector. In practice the definition of a small search range and a search with a smaller search window showed the best compromise between quality and time for calculation.

The pixel-recursive stage is a low-complexity method, which calculates dense displacement fields using a simplified optical flow approach. Following the principal of the optical flow, an update vector is calculated on the basis of spatial gradients and gradients between the two frames. The gradient between the frames is approximated by the displaced pixel difference (DPD) given by corresponding points in the left and right images.

Multiple pixel-recursive processes are started at every first pixel position of the odd lines in a block around the considered pixel (see Figure 7 with a 4x4 block). Usually, in optical flow applications the disparity vector is improved iteratively at every pixel position and the iteration is finalized by a threshold criterion. However, in our approach we only apply one iteration step to each pixel position to guarantee real-time processing. As a result, each "iteration process" ends up with one incremental update vector per pixel, which is added to the initial vector to obtain the local update vector. The local update vector of the previous pixel is then taken as initial vector for the next pixel. The very first position of every pixel recursive processing path is initialised with the start vector from the block recursive stage.

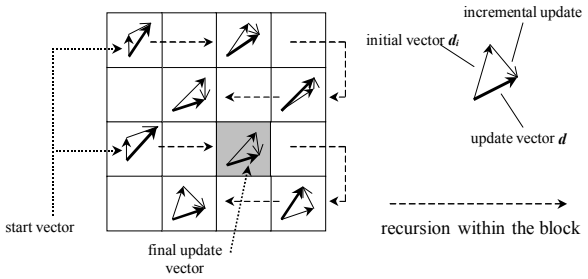

Figure 7: Outline of the pixel-recursion scheme

Finally, after processing all paths of multiple pixel-recursions, the vector with the smallest DPD among all pixel-recursive processes is taken as the final update vector. This output of pixel recursion is not necessarily the last pixel position at the end of one of the scanning paths. Often, the pixel-recursion converges over some pixels and than diverges again. Thus, the vector with the smallest DPD can be taken from an intermediate position somewhere in the middle of a scanning path (see example in Figure 7). After selecting the final update vector, its DBD is calculated and compared to the DBD of the start vector from block recursion. If the DBD of the update vector is smaller than the one of the start vector, the update vector is chosen as the final output vector, otherwise the start vector is retained.

Pixel recursion plays an important role in areas with depth discontinuities and fast motion because in this case the probability of completely wrong candidates is extremely high in the block-recursive stage. This special conditions hold for boundaries of fast moving foreground objects (e.g. fast gestures), but also for the beginning of the sequence or for scene cuts. In this sense, pixel recursion considerably improves the spatial and the temporal transient behaviour of the whole algorithm, whereas block recursion is mainly responsible for spatial and temporal smoothing of disparities in areas of homogenous depth and moderate motion.

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,

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208

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. If additional depth information from other stereo systems is available mismatches are substituted by these disparities obviously if they are not occluded in the other stereo system and above the confidence threshold.

4. EXPERIMENTAL RESULTS

The presented disparity fusion scheme has been implemented and tested in the framework of the 3DPresence demonstrator prototype. In Figure 8 and Figure 9, the original images of the trifocal short baseline camera system and the vertical wide baseline camera system are depicted. The degree of perspective distortion is remarkably larger in the latter case.

Figure 8 - Original images: trifocal short baseline

Figure 9 - Original images: wide baseline

In the following images, disparity estimation (DE) results are presented for different stages of processing. Figure 10 compares the results of a simple wide baseline system and a small baseline system. In Figure 10, left, the result of a non-guided wide baseline DE is shown. Due to large perspective differences and large occlusions the DE algorithm is not able to estimate robust disparities. Especially in the region of the arm of the right participant, the resulting disparities are erroneous. As an advantage of wide baseline systems, high depth resolution is achieved. In Figure 10, right, the result of a short baseline DE supported by trifocal constraint is shown. It can be recognized that the arms are represented much better in depth. But the higher robustness is bought by lower depth resolution which will be shown in Figure 12.

Figure 10 - Disparity images: Wide baseline (left), trifocal short baseline system (right)

In order to exploit the robustness of the short baseline system, the results are used to guide the DE of the vertical wide baseline system. The result can be seen in Figure 11. Different tests have shown that the guided HRM gives much better results than a simple second search around the disparities from the short baseline system, because the HRM yields temporal consistent disparities. Temporal consistency of disparity maps is significantly important for smooth rendering of novel views, even in this immersive video conferencing application. Due to the limited visual resolution, the improvement in depth resolution is not clearly visible in Figure 11. This becomes more visible in Figure 12 which shows the depth of the face of the right participant in higher resolution.

Figure 11 - Disparity fusion result: HRM trifocal combined with wide baseline result

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208

In section 3.2, the theoretical constraints have been explained. The disparity/depth resolution can be nicely illustrated in the following close-up views in Figure 12. In the left figure, the resulting disparities for the enhanced wide baseline system are shown, whereas on the right, the result of a small baseline system is presented. It can be recognized that the disparity/depth resolution is explicitly higher in the wide baseline case. Fine structures in depth can be recognized even in the area around the nose and eyes. The same resolution can not be achieved for the short baseline system.

Figure 12 – Depth Layer: Wide baseline refined (left), small baseline (right)

The depth maps presented in Figure 10 and 11 are used to synthesize a novel view for a virtual camera positioned in the display. Due to the fact that a 3D display is used in the 3D Presence project also a new depth map for the virtual view needs to be calculated. The synthesized view is shown in Figure 13.

Figure 13 – Synthesized view for ground truth: Wide baseline refined

5. CONCLUSION

For multi-user, multi-party 3D videoconferencing systems, spatial and temporal consistent disparity maps are required in order to synthesise novel views of the remote participants to perceive eye contact. As recognized in former approaches, pure disparity analysis on a single stereo camera configuration is not sufficient in order to provide consistent enough and robust disparities. Hence, different stereo camera configurations such as small baseline, wide baseline and trifocal camera systems have been investigated in order to find an approach which takes the advantages of all of them. Hence, a novel multi-baseline disparity fusion scheme has been presented. It firstly exploits the advantage of small baseline camera systems, which relies on spatial and temporal consistent disparities. However, the temporal consistency is mainly caused by the applied hybrid-recursive matching scheme representing the core module of all the stereo analysis modules. Secondly, a trifocal consistency check further selects the consistent disparities by cross-checking the result of the horizontal and vertical small baseline systems. Finally, this result is used as input for a wide baseline stereo system, which is able to provide the required disparity resolution and robustness of disparities.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Nguyen D., Canny J. MultiView: spatially faithful group video conferencing, In *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, pp. 799-808, Portland, Oregon, USA, April 2005.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6208
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6208