IMAGE-BASED RENDERING FOR TELECONFERENCE SYSTEMS

Eddie Cooke, Peter Kauff, Oliver Schreer

Heinrich-Hertz-Institut (HHI),
Einsteinufer 37,
10587 Berlin,
Germany
{cooke, kauff, schreer}@hhi.de http://bs.hhi.de/

ABSTRACT

To obtain an image-based immersive presence in a virtual world, two important factors should be considered: system configuration and multiview representation. We present two non-adversary system configurations. The first is the well-known convergent wide-baseline set-up while the second is a unique proposal under investigation at our institute, which is based around a parallel *multiple narrow-baseline* camera set-up. In the domain of multiview representation we introduce two non-conflicting representations that can be implemented independent of the chosen system configuration, dependent on whether compression or scalability is important to the overall system. We then discuss our implementation of an image-based rendering system for an immersive teleconferencing application where three conferees meet around a shared virtual table. The system uses a wide-baseline configuration with two stereo camera pairs capturing the reference images. The system is designed to deal with hand gestures as well as the synthesis of areas occluded in one or more of the reference images but required in the derived view. We introduce the notion of a confidence map designed to indicate, for the derived image, which reference image should provide the required texture and disparity information for a surface.

Keywords: image-based rendering, immersive teleconference, multiview representation, multiple narrow-baseline, confidence map.

1. INTRODUCTION

The goal of immersive video-conferencing systems is to allow geographically displaced conferees to experience the full spectrum of manifestation they are use to in real world meetings, i.e. gestures, eye contact, parallax viewing, etc.; in a virtual environment, [Schre01a]. To achieve this goal, 3D images of the conferees are synthesised and positioned consistently around a shared virtual table as shown in Fig. 1. To generate the required realistic 3D video objects, a multiview camera set-up captures the conferees, while disparities that represent the depth of the video objects are estimated between corresponding images. This virtual 3D scene is then rendered onto a 2D display using a virtual camera whose placement coincides with the current position of the conferee's head.

On this quest to obtain an immersive presence in a virtual world two important factors have to be taken into account: system configuration and multiview representation. System configuration refers to the way in which the camera set-up is organised. Due to the large size of immersive displays and the relatively short distance to the conferee, strongly convergent set-ups are very popular for teleconference systems. In section 2 we introduce two non-adversary configurations. The first is the well-known convergent wide-baseline set-up while the second is a unique proposal, still under investigation at our institute, which is based around parallel *multiple narrow-baseline* camera set-ups. Multiview representation expresses how the reference images from the cameras are initially combined to provide the required surface texture and disparity information for the synthesis of virtual views of the conferee. In section 3 we again present two non-conflicting representations that can be implemented independent of the chosen system configuration. The choice of representation is dependent on whether compression or scalability is important to the overall system.

In section 4 we present an approach to image-based rendering for an immersive teleconferencing system that we have implemented where three conferees meet around a shared virtual table. The system uses a wide-baseline configuration with two stereo camera pairs capturing the reference images. When head-tracking information is available to indicate the required position of the virtual view then viewpoint adaptation is implemented when this is not the case a default derived view is displayed. The system is designed to deal with hand gestures as well as the synthesis of areas occluded in one or more of the reference images but required in the derived view. Here we introduce the notion of a confidence map designed to indicate, for the derived image, which reference image should provide the required texture and disparity information for a surface. Section 5 illustrates the resultant derived view of our system while section 6 deals with conclusions and the direction of our proposed future work

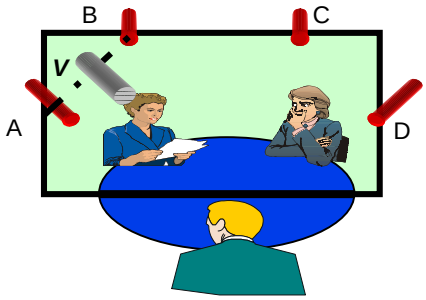
2. SYSTEM CONFIGURATION

Here we will present two diametric systems. Both systems use a multiview camera set-up and generate disparity maps, through point correspondence in images, in order to represent the depth of the video objects. The first system is the well-known wide-baseline approach, which is *static* in terms of both the number of potential users and virtual view creation abilities. The other is a new approach being developed at our institute, which we call a multiple narrow-baseline system; it is designed to be *dynamic* in terms of the number of conferees and view synthesis capabilities.

Wide-Baseline

Fig. 1 illustrates a popular wide-baseline configuration containing two convergent camera pairs, (A,B) and (C,D). The wide-baseline allows a maximum amount of information to be captured, while the convergent set-up is required to ensure enough overlap in the reference images for disparity estimation. The extent of the required convergence angle is dependent on the size of the display with respect to the distance to the conferee. The inherent problem with this approach is that we are aspiring to achieve too much through our single camera pair: maximum information and reliable disparity maps.

This leads to the development of very system specific solutions that are static in so far as when the number of participants needs to be extended then a new camera configuration as well as new algorithms for 3D analysis and synthesis are required. The generation of reliable disparity maps, which is crucial to the system, becomes difficult due to the critical nature of the camera set-up.

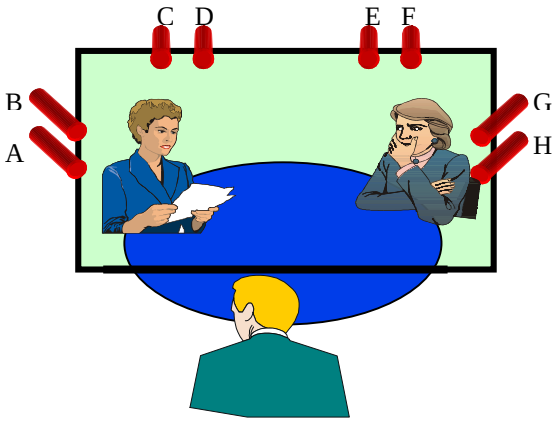


Typical wide-baseline system configuration with two stereo pairs (A,B) and (C,D).
Figure 1

Moreover, certain gestures, such as hand movements, cause severe problems for disparity estimation within the camera pair, simply due to the difference in what each reference image captures, [Pritc98]. Such problems can be reduced through the use of segmentation masks, [Schre01b], or 3D models, [Wu01], but these are far from optimal solutions.

Multiple Narrow-Baseline

The goal behind this configuration is to overcome the limitations of the wide-baseline set-up by producing a system that is modular, both in terms of system requirements as well as algorithmically, and scalable, with respect to the number of conferees. The cost of this additional flexibility is a much more algorithmically complex view synthesis procedure.



Multiple narrow-baseline system. Contains four axe parallel camera pairs.
Figure 2

The proposed system is illustrated in Fig. 2. Comparing Fig. 2 to that of Fig. 1 the camera pairs (A,B) and (C,D) can be thought of as replacing the camera pair (A,B) of Fig. 1. The camera pair set-up is based on the simplified geometry of parallel cameras. This reduces the disparity estimation complexity and therefore produces more reliable disparity maps, [Berto98]. The obvious benefit of this configuration is that critical positions, and hence the need for 3D models or segmentation masks, are less likely to arise. Once we have reliable disparity maps for each camera pair we can generate a derived view based on video object information provided by both camera pairs. The derived view is provided by *stitching* information from either camera pairs (A,B) and (C,D), or from individual cameras, i.e. B and C, together. The system is both scalable, depending on the display size and number of conferees pairs of parallel cameras can be added or removed; and modular, analysis and synthesis algorithm being dynamic enough to adapt to the current camera configuration.

It is important to note that the two system configurations are not presented here as rivals. The wide-baseline system is a practical low expense solution, which has difficulties meeting the challenges of a fully immersive teleconferencing system without the aid of 3D models or segmentation masks. In this respect algorithmic solutions to date have been very rigid and system specific. The multiple narrow-baseline approach is a high-end solution that aims to be much more dynamic and standalone, however, it is currently in its infancy with many of its challenges still to be overcome.

3. MULTIVIEW REPRESENTATION

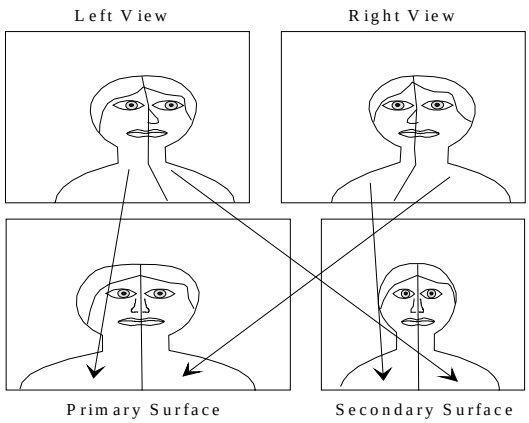
Regardless of which camera configuration is chosen for our teleconference system a multiview representation for our 3D video object is required. Here we again present two non-adversary approaches. Firstly, *Incomplete 3D*, which processes the texture surface of an object visible in several cameras and combines it such that those points available in several views are only contained once and with the highest possible resolution. Secondly, *Intermediate View*, which warps the two reference images to an intermediate derived position that is system defined. This view provides a texture that may be used as a default view and also a corresponding disparity map. In systems where viewpoint adaptation is possible the disparity map provides depth information and where it is not available the receiver is still provided with a realistic texture as an initial view.

These approaches are independent of the system configuration; the deciding factor being whether compression or scalability of the view representation is important.

Incomplete 3D (IC3D)

IC3D is a compact disparity-based multiview data representation that was developed in the context of the MPEG-4 multimedia standard, [Ohm98]. Through disparity analysis between reference images it limits the number of pixels that are required to be encoded. Each area that is visible within more than one reference image is encoded only once and with the highest possible resolution. In this sense, it eliminates redundant image contents in multiview images, directly after utilising these redundancies for a reliable search of point correspondences, but before encoding, thus resulting in higher compression. A synopsis of the theoretical background of IC3D follows; here we focus on the simplified scenario of a two-camera stereo rig with parallel camera geometry, capturing an almost convex 3D object, i.e. a conventional head-shoulder scene, see Fig. 3. However, IC3D is not in any way restricted to such a scenario, [Kauff01a].

To generate the IC3D representation, one single unwrapped texture of a video object's 3D surface is extracted from the multiple camera views. For this purpose the disparity maps between all available views must first be estimated. Through analysis of disparity gradients the areas that are best visible from particular camera positions are determined. These areas are called the areas of interest (AOI) of the individual cameras.



Generation of primary and secondary surfaces
Figure 3

In our simplified scenario the two AOI's are attained through a separation line perpendicular to the baseline of the stereo rig. Such a separation line is drawn in the two top sketches *left view* and *right view* in Fig. 3 and it roughly follows the inter-ocular

axis of the stereo cameras. Obviously, in the case of a convex object surface the two AOI's of the left and right camera are located left and right, respectively, of the separation line. Due to the parallel stereo geometry corresponding points in the reference images lie along the same scan line. Hence, in order to create the primary surface, which represents the texture of the 3D object, the AOI's are stitched along the separation line, *primary surface* in Fig. 3, through a simple horizontal shift. Due to photometric differences in the two cameras the primary surface suffers from an artefact along the separation line created during the stitching of the two AOI's. Therefore, using disparity-controlled projection, a secondary surface is generated from the image regions complementary to the AOI's, *secondary surface* in Fig. 3. This secondary surface represents an auxiliary plane, blended with the primary surface to obtain a smooth transition at the separation line.

Intermediate View

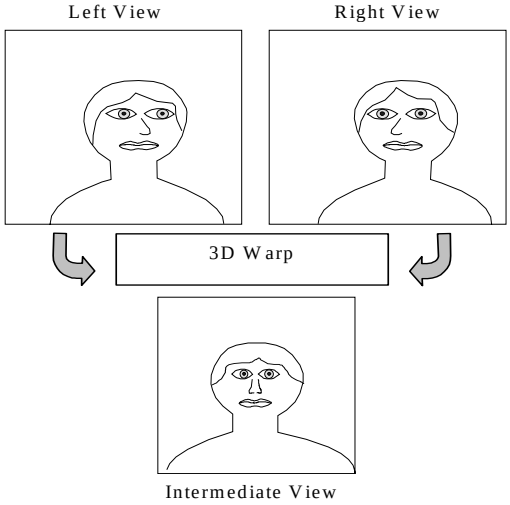
Given a pair of stereo images and corresponding disparity maps our aim is to generate a virtual intermediate view, Fig. 4. This intermediate view can be thought of as coming from a third, virtual, camera that has its own viewpoint, camera parameters, and position in the world coordinate system (WCS).

Through the use of trilinear warping, [Shash95], this virtual camera can be placed anywhere in the WCS. Trilinear warping or 3D warping makes use of the existence of a relative affine invariant λ in the general disparity equation from Eq. (1), where H and e denote the homography and epipole quantifying the different orientations and locations of two cameras in the 3D space, and $p=[x_1,y_1,l]^T$ a 2D point in the first image, corresponding to a point $p'=[x_2,y_2,l]^T$ in the second one, [Zhang96]:

$$p' \cong H \cdot p + \lambda \cdot e \qquad (1)$$

Starting from two disparity equations describing point correspondences across three camera views, a set of nine trilinear warping functions can be derived from λ. In [Avida98] approaches are discussed as to how this framework can be applied to novel view synthesis in general.

During the generation of the intermediate view the associated problems of implementing image warping arise: occlusion errors and holes. Moreover, since the intermediate view is designed not only to provide a texture map that is usable as an initial synthesis, but also to provide the basis for further 3D warping, a corresponding disparity map must also be generated. Hence the aforementioned problems filter through to the disparity map. These problems can be overcome by implementing a rendering technique that employs occlusion ordering, while holes can be filled via a hidden layer technique, [Chang99]. These two points are discussed in more depth in section 4.



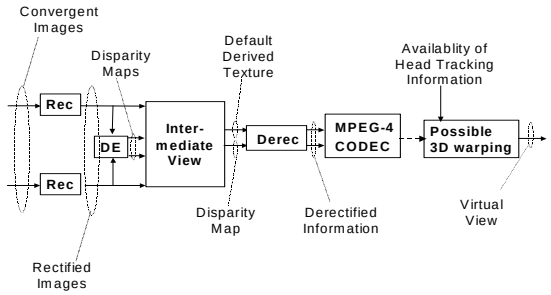
Generation of Intermediate View
Figure 4

4. IMPLEMENTATION OF IMAGE-BASED RENDERING SYSTEM

The rest of this paper focuses on the implementation of an immersive video conferencing system where three conferees meet around a shared virtual table. The system uses a wide-baseline configuration with two stereo camera pairs capturing the reference images. When head-tracking information is available to indicate the required position of the virtual view then viewpoint adaptation is implemented, nevertheless, a default derived view is defined such that an acceptable initial view synthesis of each conferee is provided in the situation where head-tracking is not implemented. Hence, the intermediate view, which provides both an acceptable initial texture and disparity map, is used for view representation.

The system is designed to deal with hand gestures as well as the synthesis of areas occluded in one or more of the reference images but required in the derived view. The resulting texture and associated disparity map are jointly encoded as arbitrarily shaped video object, using the MPEG-4 standard, where the disparity map is transmitted in a grey-scale alpha plane. Fig. 5 illustrates the system design in the form of a functional block diagram. Our first step is to rectify the convergent reference images. Rectification, which allows us to mimic a parallel stereo geometry, is required for two reasons. Firstly efficient hole-filling and occlusion interpolation across the two reference images is only

possible if cameras are parallel and share the same image plane. Secondly it allows us to implement a fast disparity estimation algorithm such as proposed in [Kauff01b]. Rectification defines two geometric 2D transforms that warp the stereo images such that epipolar lines always coincide with corresponding horizontal scan lines in the rectified images. Here we will not go into the details of rectification and the interested reader is referred to [Fusie97] for further details.



Outline of system.
Figure 5.

In our system the default derived view is positioned at the midpoint of the baseline, as indicated by camera *V* in Fig. 1, the intermediate view can however be placed anywhere in the WCS as explained in section 3. Eq. (2) and (3) describe the 3D warp required to synthesis the derived image, x_{iv}:

$$x_{iv} = x^{L} + v \cdot d^{LR}(x, y) \quad (2)$$

$$x_{iv} = x^{R} + v \cdot d^{RL}(x, y) \quad (3)$$

Since we are dealing with rectified images we need only concern ourselves with the horizontal positioning of a pixel, represented by x. Here, x^L and x^R represent the horizontal positioning of a point in the left and right reference image respectively. While $d^{LR}(x,y)$ and $d^{RL}(x,y)$ denote the associated disparity value. The coefficient ν represents a scaling factor of 0.5, the center of the baseline.

5. CONFIDENCE MAP

Eq. (2) and (3) would seem to imply that only one reference image and disparity map are required to create the intermediate view, however in practice the two disparity maps are not one-to-one correspondent. The most important task of the synthesis algorithm is to distinguish from which reference image a surface on the object should be taken. For each surface there are four possibilities:

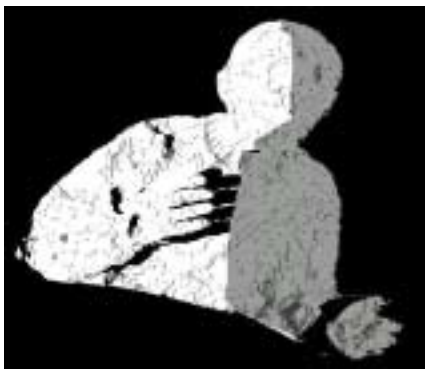
1. Visible in left and right reference image.
2. Visible only in left reference image.
3. Visible only in right reference image.
4. Neither visible in left nor right reference image.

Points 2 and 3 have a trivial solution, while for point 4 we can only estimate the surface through interpolation. Point 1, however, is not so clear-cut. We want to ensure that we use the maximum surface information available to us through the reference images yet we want to avoid the redundancy of warping identical samples. In order to aid this decision we create a confidence map. The confidence map is designed to indicate, for the derived image, which reference image should provide the required texture and disparity information for a surface. Associating confidence with surfaces in reference images, our new approach to point 1 is to choose to warp the common surface from the reference image with the highest confidence value.

Based on our camera set-up and the mainly convex nature of our object surface it is intuitive that the right side of the right reference image has a higher resolution than the corresponding side in the left image and vice versa. This is because, with the exception of maybe the hands, the shared surfaces are captured at a less oblique angle in this camera. We therefore define a separation line that is based on Eq. (4):

$$SL(i) = w \cdot SL(i-1) + (1-w) \cdot MP(i) \quad (4)$$

Where *MP(i)* is the midpoint of the rectified object at row *i*, *w* is a weighting factor between [0,1] and *SL(i-1)* is the separation line position at row *i-1*. This is implemented initially on the right reference image and through disparity-controlled projection its corresponding position in the left reference image is located.



Initial Confidence Map for Derived Image
Figure 6

This separation line is used to create our initial confidence map. Starting with the right side of the separation line in the right reference image through our initial 3D warp we obtain the right side of our confidence map. Warping the left side of the separation line in the left reference image completes the map. Fig. 6 illustrates our initial confidence map indicating which reference image provides the texture and disparity values for a surface, white for the left reference image and grey for the right. The division between the two, caused by the separation line, is clearly visible.

Occlusions & Hole-Filling

When warping surfaces in a reference image to a derived image more than one pixel may compete for a particular image location. This arises due to the overlapping of surfaces in the derived image that are at different depths in the reference image, i.e. non-convex objects. This can lead to occlusion errors, i.e. background overwriting foreground, in the derived image. In order to avoid this we implement a back-to-front occlusion-ordering warp, [McMil97]. This ensures that the traversal of the reference image during warping is such that occluded pixels are always overwritten.

As illustrated in Fig. 6, after image warping the initial confidence map contains holes. Currently we identify two different types of holes that we differentiate between based on size. A small hole indicates discrepancies in the disparities of a continuous surface in the reference image. Implementation of a median filter on the initial map has the affect of closing these smaller holes and removing outliers due to false disparities.

Hole-Filled & Filtered Confidence Map
Figure 7

However, if the hole is due to the movement of a foreground object with respect to the background, i.e. a newly disclosed surface, the hole size, which is dependent on the severity of the 3D warp, can be substantially larger. Information about this newly disclosed surface maybe found in the other reference image. Therefore, we implement a back-to-front reverse occlusion ordering traversal on this side of the separation line in the other reference image. This filling is based on the assumption that we only have two surface layers: foreground - arm and hand, and background - torso, and that the background should be used to fill the hole. In the case where more layers exist the last layer in the reverse occlusion ordering traversal will be used i.e. the back most. Fig. 7 indicates the corresponding confidence map. Any holes that remain in the confidence map are holes that are larger than the filter mask itself. To fill these a linear interpolation filter is applied along the horizontal direction.

Critical Regions

While creating derived views of teleconferees it is important to note that certain anomalies are more noticeable and hence off-putting to the user than others. In a teleconferencing system that allows a full range of body gestures a consistent and realistic synthesis of the hands is vital for user acceptance. We must avoid effects such as a hand with, for example, 2 thumbs, 6+ fingers, or standalone fingers.

The first step in the process is to identify the hands in the reference images. In contrast to the huge amount of approaches in the field of gesture recognition, we are not interested in the orientation of the hands or the motion, [Pentl97], but more in creating a closed region that coincides as much as possible with the real contour of the hand. This is achieved through a segmentation process that is separated into two subsequent steps. In the first step, a bounding box is tracked over time for each hand in order to limit the search region and to speed up the whole processing. In the second step, a region growing technique is applied starting at a seed point in the centre of the hand to extract the closed region of the hand, [Mehne97]. For more information see [Schre01b], the results are illustrated in Fig. 8.

Segmented binary mask & contour in original image.
Figure 8

Once the hands are identifiable we can implement a consistent approach to their processing. The first item to examine is the interaction of the hands with the separation line in the confidence map. We need to avoid the situation where a hand that lies on the separation line is partly taken from each reference image. In such a case we would end up with a derived image in which the hand was displaced or, in the worse case scenario, simply cut off along the separation line. Based on our initial definition of surface confidence the hand in the right side of the right reference image has a higher confidence than the same hand in the left reference image and vice versa. We extend our algorithm to guarantee that the hand is only ever taken from this reference image, Fig. 9(c). Our algorithm ensures that once this decision has been made, regardless of its current position, i.e. which side of the separation line it is on, the information is consistently provided from this cameras' reference images. This has the desired effect of providing consistency to the form of the hand over a sequence.

(a) Left Rectified Image (b) Right Rectified Image

(c) Confidence Map (d) Derived View
Figure 9

In our confidence map holes created due to lack of information in one reference image are filled using information from the other reference image. This approach reduces artefacts such as the stretching effect of textures across newly disclosed surfaces. However, for critical areas such as the hands, it can produce new effects that are even more disturbing, i.e. hands with extra thumbs or fingers, or ghost fingers. In order to avoid these new artefacts our algorithm ensures that holes in one reference image are not filled with information that has been identified as belonging to a hand in the other reference image.

However, as can be seen in Fig. 8 these hand masks and the associated texture and disparities may not have a one-to-one relationship. These false disparities associated to hands can lead to situations where, during the warping of the hand, part of it will remain *stuck* to the background. This problem can cause artefacts to appear, i.e. unconnected hand parts, which we call the *halo effect*. These errors present two distinct problems:

1. Part of the hand, *indicated in the mask as belonging to the hand*, has a disparity value associating it with the background object.

This is easily dealt with since we know, through use of the mask, that this outlier exists. The median filter, mentioned in subsection *Occlusions & Hole-Filling*, removes this problem.

2. Part of the hand, *not indicated in the mask as belonging to the hand*, has a disparity value associating it with the background object.

Here the problem is more difficult to solve. The mask does not indicate that part of the hand has been falsely associated with the background object. This problem is solved through our hole-filling algorithm. During our implementation of a back-to-front reverse occlusion ordered filling we introduce a buffer around the hole. This buffer ensures that such *halos* are overwritten with background information supplied by the other reference image, Fig. 10.

(a) (b)

Derived hand with (a) *halo effect* & (b) processed to remove *halo.*
Figure 10.

Again this filling is based on the assumption that we only have two surface layers: foreground - arm and hand, and background - torso, and that the background should be used to fill the hole.

6. EXPERIMENTAL RESULTS

Fig. 9 demonstrates the resultant intermediate view of our proposed system. Fig. 9(a) and (b) illustrate the rectified images taken from the left and right cameras respectively. Fig. 9(c) shows the corresponding confidence map for our derived intermediate view, with the separation line indicated. Fig. 9(d) is the resultant intermediate view created via this confidence map.

After derectification both texture and associated disparity map can jointly be encoded as an arbitrarily shaped video object, using the MPEG-4 standard, where the disparity map is transmitted in a grey-scale alpha plane.

7. CONCLUSIONS AND OUTLOOK

We have shown that the combination of confidence map and hand mask provides us with a means to deal with the problems introduced through hand gestures as well as the synthesis of areas occluded in one or more of the reference images but required in the derived view, Fig. 9.

Currently our main focus is on the implementation of a multiple parallel narrow-baseline system configuration. Via this set-up we aim to generate more reliable disparity maps, than in the wide-baseline system, and hence remove the need for the generation of hand masks or 3D models. However, this is only one aspect of this configuration. We are also researching ways of modularising the image analysis and synthesis processes so that they are flexible enough to handle numerous systems and conferees.

8. ACKNOWLEDGEMENTS

This study has been supported by the Ministry of Science and technology of the Federal Republic of Germany, Grant-No. AK 002. Furthermore the authors would like to thank TU Delft for the provision of test sequences.

REFERENCES

[Avida98] Avidan,S., Shashua,A.: Novel View Synthesis by Cascading Trilinear Tensors, *IEEE Trans. on Vis. and Comp. Graphics*, Oct - Dec 1998.

[Berto98] Broggi,A.: GOLD: A Parallel Real-Time Stereo Vision System for Generic Obstacle and Lane Detection, *Trans. on Image Processing*, Vol.7, No.1, January 1998.

[Chang99] Chang,C-F. et al.: LDI Tree: A Hierarchical Representation for Image-Based Rendering, *Proc. of SIGGRAPH 99*, 1999.

[Fusie97] Fusiello,A. et al.: Rectification with unconstrained stereo geometry., *Proc. of BMVC'97*, pp.400-409, 1997.

[Kauff01a] Kauff,P. et al.: Advanced Incomplete 3D Representation of Video Objects Using Trilinear Warping for Novel View Synthesis, *Proc of PCS'01*, pp.429-432, 2001.

[Kauff01b] Kauff,P. et al.: Fast Hybrid Block and Pixel Recursive Disparity Analysis for Real-Time Applications in Immersive Tele-Conferences, *Proc. of WSCG'01*, 2001.

[McMil97] McMillan,L.: An Image-Based Approach to Three-Dimensional Compuer Graphics., PhD Thesis, University of North Carolina at Chapel Hill, pp.45-49, 1997.

[Mehne97] Mehnert,A., Jackway,P.: An improved seeded region growing algorithm, *Pattern Recognition Letters,* Vol.18, pp.1065-1071, 1997.

[Ohm98] Ohm,J-R. et al.: Incomplete 3D – A new Technique for Multiview Data Representation, *Proc of IMDSP'98*, pp.311-314, 1998.

[Pentl97] Pentland,A. et al: Pfinder: Real-Time Tracking of the Human Body, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.19, pp. 780-785, July 1997.

[Pritc98] Pritchett,P., Zisserman,A.: Wide Baseline Stereo Matching, *Proc. Int. Conf. on Computer Vision* , pp. 754-760, Jan 1998.

[Schre01a] Schreer,O., Kauff,P.: An Immersive 3D Videoconferencing System Based on a Shared Virtual Table Environment, *Proc. of Int. Conf. on Media Futures*, May 2001

[Schre01b] Schreer,O. et al.: Hybrid Recursive Matching and Segmentation-Based Postprocessing in Real-Time Immersive Video Conferencing, *Vision, Modeling, & Visualization*, 2001.

[Shash95] Shashua,A.: Algebraic Functions for Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[Wu01] Wu,Y., Huang,T.S.: Hand Modeling, Analysis, and Recognition for Vision-Based Human Computer Interaction, *IEEE Signal Processing Magazine*, pp. 51-60, May 2001.

[Zhang96] Zhang,Z., Xu,G.: Epipolar Geometry in Stereo, Motion and Object Recognition, *Kluwer Academic Publisher*, 1996.