

# MULTIPLE NARROW-BASELINE SYSTEM FOR IMMERSIVE TELECONFERENCING

Eddie Cooke, Peter Kauff, Oliver Schreer

Heinrich-Hertz-Institut,  
Image Processing Department,  
Einsteinufer 37, D-10587 Berlin, Germany  
Email: {cooke, kauff, schreer}@hhi.de

**Abstract:** *An important aim of immersive teleconferencing systems is to create realistic 3D virtual views of remote conferees. Hence, systems should be able to deal with hand gestures as well as occluded areas in reference images required in derived views. The quality of such derived views is dependent not only on the analysis and synthesis process but also the multiview camera set-up. Often the popular convergent wide-baseline stereo approach aspires to achieve too much through a single camera pair: maximum information and reliable disparity maps. We identify how this dichotomy leads to problems in the analysis and synthesis process, often leading to a restrictive system specific solution. We then define a new approach, a multiple narrow-baseline set-up, designed to overcome the limitations of the wide-baseline set-up, being modular, both in terms of system requirements as well as algorithmically, and scalable, with respect to the number of conferees.*

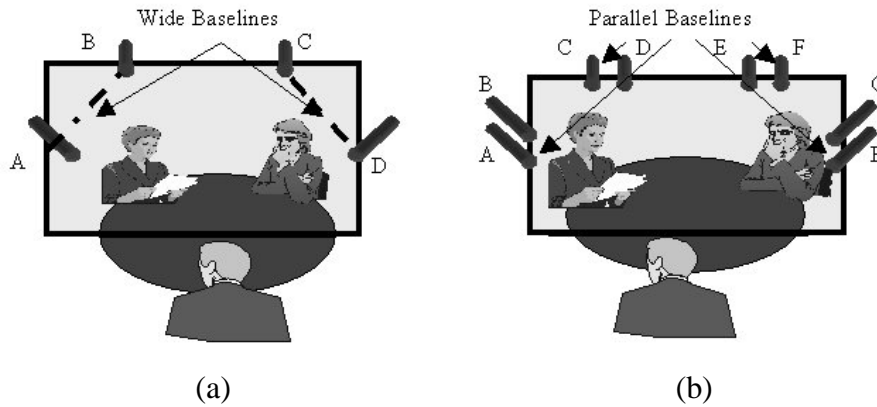
**Key words:** *multiple narrow-baseline, immersive teleconferencing, confidence map.*

## 1. INTRODUCTION

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

Hence, the level of realism of the derived view is dependent on the camera set-up and the quality of the image analysis and synthesis process. Previous research in derived view creation includes: (i) *Intermediate Viewpoint Interpolation* which produces views generated via disparity-based interpolation, [2,3,4]. (ii) *Incomplete 3D* which concentrates on maximising sampling density by combining images along a disparity based separation line, [5]. (iii) *Middle View* approach which produces a derived view and disparity map at the midpoint of the baseline through a simplified 3D warp, [6]. There are a number of shortcomings with these approaches: (i) has only been validated on head-shoulder sequences, (ii) has been extended to work with gestures but, like (iii), creates an initial derived view along the baseline and requires a hidden layer to handle occlusions that may arise due to

further 3D warps away from the baseline. In terms of teleconference system set-up (i-iii) have been developed on wide-baseline systems, which create problems for the disparity estimation process.



**Fig. 1.** (a) Typical wide-baseline system configuration with two stereo pairs (A,B) and (C,D).  
 (b) Multiple narrow-baseline system containing four narrow-baseline camera pairs.

We propose a new approach, which produces an initial derived view and disparity map outside the baseline, hence removing the need for a hidden layer, and overcomes potential disparity estimation problems by using a narrow-baseline set-up. The system is designed to be modular, both in terms of system requirements as well as algorithmically; and scalable, with respect to the number of conferees.

## 2. SYSTEM CONFIGURATION

By far the most popular teleconference system configuration is the convergent, wide-baseline, stereo camera pair, Fig 1(a). The wide-baseline is designed to maximise the amount of information captured, while the convergent set-up, whose angle is dependent on the size of the display with respect to the distance to the conferee, is required to ensure enough overlap in the reference images for disparity estimation. The inherent problem with this approach is that we are aspiring to achieve too much through a single camera pair: maximum information and reliable disparity maps. Extreme differences in a surface's orientation in the reference images, e.g. the hands, cause severe problems for the disparity estimation process, which must somehow identify them as corresponding surfaces [7]. Such problems can be reduced through the use of segmentation masks [8], or 3D models [9], but these are far from optimal solutions.

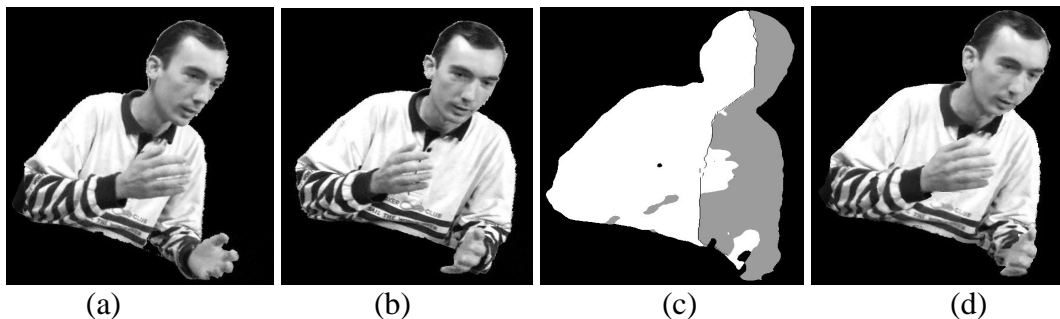
### 2.1. Multiple Narrow-Baseline System

The multiple narrow-baseline 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 algorithms being dynamic enough to adapt to the current camera configuration. An example set-up is illustrated in Fig. 1(b). The camera pairs (A,B) and (C,D) can be thought of as replacing the camera pair (A,B) of Fig. 1(a). The narrow-baseline between camera pairs ensures that the difference in surface orientation is minimised, hence more overlap between images, while the simplified geometry of parallel cameras reduces the disparity estimation complexity. The obvious benefit of this configuration is that critical

regions, and hence the need for 3D models or segmentation masks, are less likely to arise in the disparity map. Once we have reliable disparity maps for each camera pair we can generate a system defined initial derived view. Through the use of trilinear warping this virtual camera can be placed anywhere in the world coordinate system. This warping makes use of the existence of a relative affine invariant  $\lambda$  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, [10]:

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

Starting from two disparity equations describing point correspondences across three camera views, a set of nine trilinear warping functions can be derived from  $\lambda$ . In [11] approaches are discussed as to how this framework can be applied to novel view synthesis in general.



**Fig 2** (a) image from camera A of pair (AB) (b) image from camera C of pair (CD)  
(c) Confidence Map, white indicating surface from A & grey from C (d) Derived view

When warping we must distinguish from which reference image a surface in the derived view should be taken. For each object surface there are three possibilities, it is: (i) visible in both, (ii) visible only in one, or (iii) visible in neither of the reference images. While cases (ii) and (iii) are essentially trivial, (i) 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 define a confidence map, Fig. 2(c). This is designed to indicate, for the derived image, which reference image should provide the required texture and disparity information for a surface. This determination of surface confidence is based on the sampling density of a surface in the reference image. A surface visible in both images will have a higher sampling density in the image where it is less oblique. Hence, our new approach to (i) is to choose to warp the common surface from the reference image with the highest confidence value, [12]. It may happen that more than one pixel competes for the same pixel position; this arises due to overlapping of surfaces and can lead to occlusion errors. To avoid this we implement a back-to-front occlusion-ordering warp, [13]. Also, holes may occur due to the movement of a foreground surface with respect to the background. Information about these newly disclosed surfaces maybe found in other images, hence, for hole filling we implement a back-to-front reverse occlusion ordering traversal on the surface in the other reference images. This ensures that our derived view combines all the surface information the various reference images make available. Fig. 2(a,b) illustrates reference images from the example set-up of Fig. 1(b). Fig. 2(a) corresponds to the reference image from camera A in the stereo pair AB, while Fig. 2(b) is that of camera C from pair CD. The confidence map is created from a combination of the warped reference images, while Fig 2(d) illustrates the respective derived view.

### 3. CONCLUSION

We have identified the important requirements of derived views in teleconference systems and how large differences in a surface's orientation in reference images, due to the wide-baseline set-up, lead to anomalies and an often restrictive system specific solution. We defined a new multiple narrow-baseline system, which produces reference images with a greater surface overlap and hence produce more reliable disparity maps. We introduced 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. We explained how through trilinear warping we can create a default derived view and disparity map outside the baseline. Currently we are researching the extent to which we can develop our confidence map to handle more than two stereo camera pairs and what affects this has on our disparity map. Obvious application areas are image-based rendering systems requiring realistic 3D video objects i.e. Tele-learning, Tele-collaboration etc.

### REFERENCES

- [1] O. Schreer and P. Kauff, An immersive 3D videoconferencing system based on a shared virtual table environment, *Proc. of Int. Conf. on Media Futures*, 2001
- [2] E. Chen and L. Williams, View interpolation for image synthesis, *Proc. ACM SIGGRAPH'93*, 279-288
- [3] T. Werner et al., Rendering real-world objects using view interpolation, *Proc. IEEE Int. Conf. Comp. Vision*, 1995, 957-962
- [4] J.-R. Ohm and E. Izquierdo, An object-based system for stereoscopic viewpoint synthesis, *IEEE Trans. Circ. Syst. Video Tech.*, CSVT-7(5), 1997, 801-811
- [5] P. Kauff et al, Advanced Incomplete 3D Representation of video objects using trilinear warping for novel view synthesis, *Proc. PCS '01*, 2001
- [6] B.J. Lei and E.A. Hendriks, Middle view stereo representation, *International Conference on Image Processing IEEE*, 2001
- [7] P. Pritchett and A. Zisserman, Wide baseline stereo matching, *Proc. Int. Conf. on Computer Vision*, 1998, 754-760
- [8] O. Schreer et al., Hybrid recursive matching and segmentation-based postprocessing in real-time immersive video conferencing, *Vision, Modeling, and Visualization*, 2001
- [9] Y. Wu and T.S. Huang, Hand modeling, analysis, and recognition for vision-based human computer interaction, *IEEE Signal Processing Magazine*, 2001, 51-60
- [10] Z. Zhang and G. Xu, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluwer Academic Publisher, 1996
- [11] S. Avidan and A. Sashua, Novel view synthesis by cascading trilinear tensors, *IEEE Trans. on Vis. and Comp. Graphics*, 1998
- [12] E. Cooke et al., Image-based rendering for teleconference systems, *WSCG'02*, 2002
- [13] L. McMillan, *An Image-Based Approach to Three-Dimensional Computer Graphics*, PhD Thesis, University of North Carolina at Chapel Hill, 1997, 45-49