A FLEXIBLE 3D TV SYSTEM FOR DIFFERENT MULTI-BASELINE GEOMETRIES

Oliver Schreer, Christoph Fehn, Nicole Atzpadin, Marcus Müller, Aljoscha Smolic, Ralf Tanger, Peter Kauff

Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut
Einsteinufer 37, 10587 Berlin, Germany
Oliver.Schreer@hhi.fhg.de
http://ip.hhi.de

ABSTRACT

Interoperability, scalability and adaptability are important features for a successful introduction of future 3D TV services. Hence, new concepts must be able to adapt the multi-view geometry of the capturing system to the geometry of the 3D reproduction systems. An approach is discussed, which considers these adaptation issues based on the concept of an N x video-plus-depth data representation. The core algorithms for depth map creation on the analysis side and depth image based rendering on the reproduction side are presented.

1. INTRODUCTION

The advent of three-dimensional television (3D TV) became of relevance in the early 1990s with the introduction of the first digital TV services. Worldwide R&D activities started with the aim to develop standards, technologies and production facilities for 3D TV. Recent advances in the area of 3D display technology, image analysis and image based rendering (IBR) as well as digital image compression and transmission of video and depth data accelerated the development of concepts and first prototypes for future 3D TV services. Convincing real-time 3D TV demonstrators have been shown at SIGGRAPH´04 in Los Angeles and IFA´05 in Berlin [1][2]. In addition, the MPEG group has established a 3DAV (3D Audio/Visual) Ad-Hoc group to investigate the needs for standardization in 3D [3]. Nowadays, the introduction of 3D is considered by many people as the next logical step compared to the introduction of color TV in the 1960s.

Against this background the European IST research project ATTEST has recently investigated a new system concept for 3D TV [4]. In contrast to former proposals, which usually relied on the basic concept of an end-to-end stereoscopic video chain, this novel approach is based on an alternative data representation format known as video-plus-depth. The concept has some crucial advantages over former 3D TV proposals, such as backwards compatibility to existing DVB services, efficient compression capabilities and the possibility to adapt the 3D reproduction to different display properties, viewing conditions and user preferences. Therefore MPEG has now established a further Ad-Hoc group, which will focus on the special business case of 3D TV using *N*-video-plus-depth.

In the next section, the advanced 3D TV system concept is presented using a video-plus-depth structure. In section 3, the problem of adaptation of different multi-view geometries for the capturing and reproduction is discussed. Then, a general concept for creation of depth maps for arbitrary capturing configurations is presented. It is followed by an algorithmic solution for depth image based rendering (DIBR). A conclusions ends the paper.

2. ADVANCED 3D TV SYSTEM CONCEPT

The structure of an advanced 3D TV system is based on the ATTEST concept [4]. On the capturing side, *N* regular video streams are generated enriched with so-called depth maps providing a Z-value for each pixel. The following different inputs can be used for 3D acquisition:

- standard stereo cameras with two views or multi-baseline systems with more than two cameras for which related depth maps must be created. This is done by a depth analysis pre-processing step and an example is given in section 4.
- depth-range cameras directly providing video-plus-depth streams, based on active range cameras such as Zcam^{TM} from 3DV Systems or the NHK Axivision HDTV camera.
- post-processing tools, that allow to manually converting conventional 2D movies to the desired 3D representation format.

The *N* video-plus-depth streams are then encoded by a suitable multi-view coding profile and transmitted to the receiver. After decoding, the *N* video-plus-depth streams are converted to *M* regular video views by using DIBR techniques. This conversion process depends on the

1-4244-0367-7/06/$20.00 ©2006 IEEE

properties and the number of views of the targeted 3D display and the particular system configurations that adapts the 3D reproduction to local viewing conditions and individual user preferences. The required views for 3D reproduction depend on the display type and can be:

- standard stereo systems with glasses (fixed *M*=2 views)
- tracked auto-stereoscopic single-user systems supporting head motion parallax (variable *M*=2 views depending on head position)
- auto-stereoscopic multi-user systems with *M*>2 views, whereas the number *M* differs depending on the given display technology and is in a range of *M*=10 or even more.

3. ADAPTATION OF CAPTURING AND REPRODUCTION GEOMETRY

Following the previous section, the number *N* of source views and the number *M* of display views is different in the general case. Not only the number, but also the geometry of the capturing and display system is different under general considerations. This also holds for the special case of *M*=*N* views, because depth data are useful for modifying basic 3D parameters during rendering and, with it, for adapting the depth reproduction of the scene to the specific preferences of the user (see Section 4). Hence, an adaptation from *N* source views to *M* display views is particularly important, because of interoperability reasons for future 3D TV services. In Fig.1, some example configurations for a multi-view capturing system are depicted.

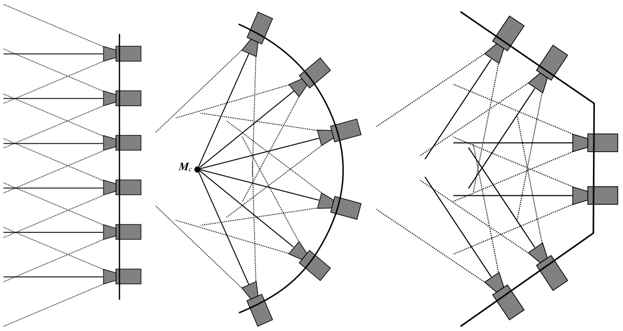

Fig 1: Parallel (left), convergent (middle) and pair wise convergent multi-view camera setup (right)

In contrast to the multi-view camera configuration for scene capture, the targeted 3D display defines the required number of *M* virtual views on the rendering side. The geometry of the virtual views can be regarded as equidistantly arranged cameras along a baseline. In Fig.2, the multi-baseline geometry related to the 3D display is shown. The geometry of this specific multi-baseline setup is known from the display parameters. In addition to that, the geometry of the virtual views, specifically the interaxial distance between two views can differ according to the user preferences. The details will be presented in section 5.

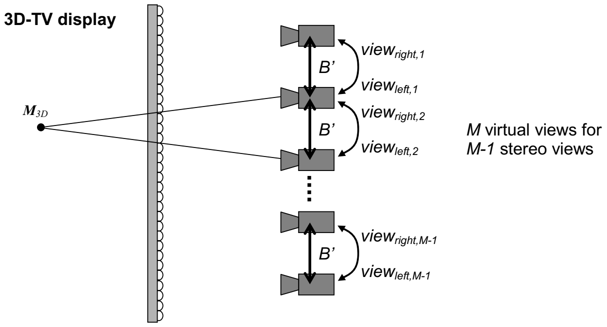

Fig.2: Required virtual views for a 3D TV display

Due to the differences in the number of views and the geometry of the capturing and the reproduction, a general intermediate data representation format like the *N*-video +depth concept becomes important. This format can be generated from arbitrary camera configurations on one hand, whereas on other hand, it contains all required information for the synthesis of all destination views.

4. GENERATION OF DEPTH MAP

If just original camera views are available on the capturing side, a reliable depth map for each original image must be generated in order to follow the *N*-video+depth concept. The estimation of suitable depth maps from stereo or multi-baseline systems is certainly one of the most challenging tasks in the given context. The following considerations briefly describe a solution, that has been used in conjunction with the above system proposal. Results are shown on the basis of a multi-view test sequence with *N*=6 views (see Fig.3) with a distance of 1.5m between the outer cameras and convergent setup. For the sake of clarity and without lack of generality, it is assumed, that the distance between adjacent cameras is given by same baseline *B*, that the focal length *F* is the same for all cameras and that all cameras converge to one common 3D point $\boldsymbol{M}_c=(X_c,Y_c,Z_c)^T$.

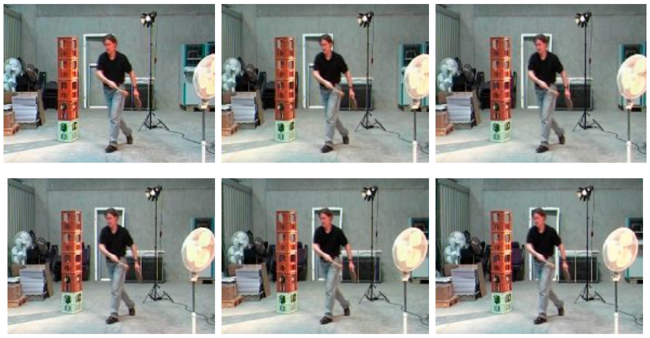

Fig.3: Test sequence with *N=6* views (from top to bottom)

The presented approach assumes a multi-view camera setup of *N* views, whereas the cameras are mounted without any geometric restrictions. This might be a parallel configuration or any convergent setup depending on the specific system requirements as presented in the previous section. Supposing that the multi-view set-up is properly calibrated the resulting camera parameters are used to firstly rectify pair wise the *N* views [5]. This allows for the search of corresponding points along scan lines and, thus, eases the subsequent processing. Then, disparity maps are estimated by applying a suitable matching algorithm to a pair of rectified images. Based on camera parameters and the resulting disparity maps, a depth map is derived for each rectified image. Finally, the depth maps are de-rectified in order to get associated depth maps for the original images.

For disparity estimation, we have chosen a fast HRM (Hybrid Recursive Matching) algorithm [6]. 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, which are detected and corrected by sophisticated post-processing. One criterion for detecting mismatches is a confidence measure, that is directly derived from the normalized cross-correlation used by HRM. As the HRM estimates independently two disparity maps for each rectified image pair (right to left, left to right), both maps are used to prove the consistency of the disparities. In case of a multi-view system with *N>2*, this consistency check can also be extended towards trifocal constraints [7].

As an example, Fig.4 (top left), shows a result of these confidence and consistency checks after disparity estimation from view 3 to 4 of the test sequence in Fig.3. The black pixels indicate areas, where the checks have detected mismatches and, hence, where disparities have been removed. Grey areas contain the original HRM disparities that survived the check. Usually, there are two reasons for the detected mismatches:

- ambiguities during matching (homogeneities, similarities, periodicities, etc.) and
- 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 occluded areas, point correspondences do not exist at all and cannot be matched in principle. Thus, the aim of further processing is to distinguish between these two sources of fault. For this purpose the missing disparity values are first reconstructed by using segmentation-driven interpolation exploiting two different techniques: color clustering (Fig.4, top right) and change detection (Fig.4, bottom left).

Following the assumption that disparities are spatially and temporally consistent in color segments and stationary backgrounds, different interpolation operators are used in dependence on segment size, motion content and already existing depth information. This interpolation process again results in a dense "per-pixel" disparity map. Then, a second consistency check is applied to this refined disparity map. Supposing that the previous interpolation step was successful and has corrected all mismatches caused by ambiguities, this second consistency check will mainly detect the remaining occluded areas (Fig.4, bottom right).

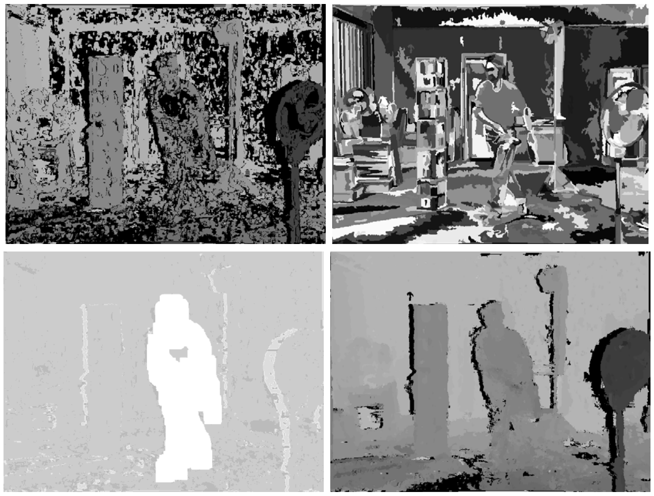

Fig.4: Results of 1st consistency check, color clustering, change detection and 2nd consistency check (left to right, top to bottom).

After post-processing a depth map *Z(u,v)* is re-calculated from disparity map *D(u,v)* by using Eq. (1) and it is then de-rectified such that it again fits to the original camera image.

$$Z(u,v) = \frac{r \cdot F \cdot B}{\left| D(u,v) + D_0(M_C) \right|} \tag{1}$$

Here, *r* is a scaling factor caused by de-rectification and D_0 denotes a disparity offset taking into account that a sensor shift occurs during rectification. For a multi-baseline system it is useful to derive D_0 consistently across all views from the common convergence point $\boldsymbol{M}_c$. Note, that *r, F, B*, D_0 and $\boldsymbol{M}_c$ are constants, which are defined by calibration and rectification. Finally, the missing depth data in the occluded areas are reconstructed. If *N*>2, they are taken from the opposite image pair providing the complementary data in occluded areas. Note, that depth data in non-occluded areas are consistent due to prior testing of the trifocal constraint. Fig.5, shows the final result of a depth map for camera 3 merged from disparity analysis results in camera pairs (2,3) and (3,4).

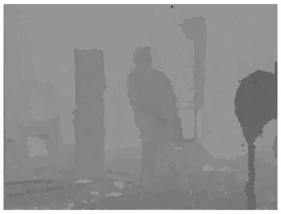

Fig.5: Resulting depth map

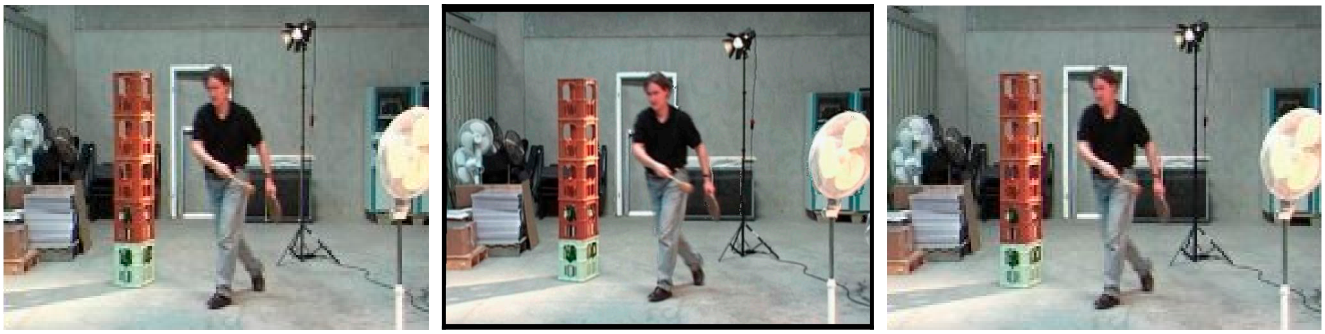

Fig.6: Result of depth image based rendering (middle), based on left and right original image (left and right)

5. RENDERING FOR 3D DISPLAY

For rendering a particular virtual view, the closest camera view and the related depth map are selected. The warping of the original view is again based on a rectified arrangement. The rectified view and the associated depth map are used to calculate a parallax map *P(u,v)*, which contains the correspondence between pixels in the rectified and the virtual view (see Eq.(2)). This parallax map shifts a color sample at pixel *(u,v)* in the rectified camera image along the scan line to position *(u',v')=(u+P(u,v),v)* in the virtual view.

$$P(u,v) = P_0 + \frac{F \cdot X_V(B', X_0)}{r' \cdot Z(u,v)}) \qquad (2)$$

Eq. (2) contains several parameters, which can be used for controlling the final depth impression at the 3D display. X_v describes the distance from the virtual view to the real camera view. It mainly depends on the locations where the *M* virtual views have been placed relatively to the *N* camera views and can be derived from a reference point X_0 and the interaxial distance *B'* between two adjacent virtual views. Note that, *B'* also defines the spatial density of virtual views along the multi-baseline system. It can therefore be used to scale the depth impression. Furthermore, the parallax offset P_0 can be used to shift the 3D scene relatively to the screen surface; i.e. behind or in front of the display. Finally, $\boldsymbol{M}_c$ and, with it, D_0 influence the appearance of head motion parallax viewing; i.e., how objects in different depth layers move relatively to each other while moving the head. It must be noted, that D_0 is only relevant for tracked system where the reference point X_0 and, with it, X_v varies in dependence of the user's head movement. In 3D displays with stationary views (X_v =const) P_0 and D_0 have the same effect and can be merged into a single parameter therefore. Self-occlusions are handled by following the occlusion-compatible ordering and holes areas are filled with samples from a complementary camera view. Therefore, a congruent warping is applied to another nearby camera view and the warped samples are inserted into the excluded areas of the virtual view. Suitable blending techniques are used to reduce artifacts. The final view is then de-rectified to the original orientation of the virtual view. Fig.6, middle, shows the final DIBR result, where the *N*=6 original views (Fig. 3) have been converted to *M*=45 virtual views. The depicted view is located in the middle between camera 2 and 3. The novel view is just based on the closest left and right original view. The complete rendering is running in real-time on a standard PC based on original views and the related depth.

6. CONCLUSIONS

Due to the high diversity in 3D display technology, interoperability and scalability will be one of the challenges in the design of suitable 3D TV services. In this context, the paper has proposed a flexible solution, which is based on a multiple video-plus-depth structure. This data representation can be used to convert *N* transmitted video-plus-depth streams to the *M* views of a given 3D display. Related DIBR techniques have been evaluated on the basis of a tracked auto-stereoscopic 3D display. The results are promising and indicate that the conversion from *N* to *M* views can be achieved with sufficiently good quality.

7. ACKNOWLEDGEMENT

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

8. REFERENCES

[1] W. Matusik, and H. Pfister, "3D-TV: A Scalable System for Real-Time Acquisition, Transmission and Autostereoscopic Display of Dynamic Scenes", *Proc. of ACM SIGGRAPH*, pp. 814-824, LA, USA. Aug. 2004.

[2] 3D Image Processing, "Three-Dimensional Real-Time HDTV: An Overview", White Paper, http://www.3d-ip.com, 3D Television, 2005.

[3] A. Smolic, and D. McCutchen, "3DAV Exploration of Video-Based Rendering Technology in MPEG", *IEEE Trans. CSVT,* 14(3), 348-356, March 2004.

[4] C. Fehn, "Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV", *Proc. of Stereoscopic Displays and Virtual Reality Systems XI*, pp. 93-104, San Jose, CA, USA.

[5] A. Fusiello, E. Trucco, A. Verri, "Rectification with Unconstrained Stereo Geometry", *BMVC*, pp.400-409, Essex, UK, Sept. 1997.

[6] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing", *IEEE Trans. CSVT,* 14(3), pp. 321-334, March 2004.

[7] A. Shashua, "Trilinear Tensor: The Fundamental Construct of Multiple-View Geometry and its Applications", *Workshop on Algebraic Frames For The Perception Action Cycle,* Kiel, Germany, September 1997.