GENERATION OF MULTI–VIEW VIDEO PLUS DEPTH CONTENT USING MIXED NARROW AND WIDE BASELINE SETUP

Frederik Zilly, Christian Riechert, Marcus Müller, Peter Kauff
Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany
frederik.zilly@hhi.fraunhofer.de

ABSTRACT

Content production for stereoscopic 3D-TV displays has become mature in the past years. The content is usually shot using two cameras as the glasses-based target devices require two views as input. Besides stereoscopic 3D-TVs, huge progress has also been achieved in the improvement of the image quality of glasses-free auto-stereoscopic displays and light-field displays. Concerning the latter two display families, the content production workflow is less elaborated and more complex, as the number of required views differs considerably and is likely to increase in the near future. As a co-existence of all 3D display families can be expected for the next years, one aims to establish an efficient content production workflow which yields to high quality content for all 3D-TV displays. Against this background, we present a content production workflow which uses as acquisition device a four camera rig involving a central narrow baseline and a wide baseline with two satellite cameras and propose a multi-view video plus depth generation workflow optimized for the four-camera setup.

Index Terms — 3D-TV, depth image based rendering, DIBR, depth estimation, multi-view video plus depth, MVD

1. INTRODUCTION

In the past years, a variety of new 3D displays has entered the market, and the recent introduction of auto-stereoscopic displays with a horizontal image resolution of 4k illustrates the vitality of the market and related advancements in display engineering and manufacturing. However, as the acceptance of new displays is related to the availability of suitable content, it is crucial for the successful introduction of auto-stereoscopic displays that an efficient content production workflow is established. However, these displays require a varying number of views. Thus, a suitable approach is to generate an arbitrary number of views by using Depth Image Based Rendering (DIBR) [1,5,6,7]. One or more depth maps along with the respective videos are required, i.e. the data needs to be stored in the multi-video plus depth (MVD) format which is also suitable for coding and transmission of 3D-TV content. However, DIBR introduces rendering artifacts to the content which can mostly be neglected in the case of auto-stereoscopic displays as these displays perform by design an inherent subsampling and/or low-pass filtering of the content. When shown on high resolution stereoscopic 3D displays, these artifacts become exposed and are likely to impair the image quality and 3D sensation. However, as stereoscopic 3D displays play an important role in the 3D-TV market of the next years, an efficient content creation workflow must not ignore them. The best possible image quality needs to be offered to all display types.

Against this background we propose a multi-camera rig and an associated production workflow involving a linear camera array which, unlike in other approaches (e.g. [4]), is not formed by cameras in an equidistant setup, but which has two cameras in the center forming a narrow baseline and two additional satellite cameras forming a wide baseline.

2. MULTI-BASELINE CAMERA SETUP

Figure 1 gives an overview about the positioning of the four cameras on the multi-camera rig. As typical for linear camera arrays, all four optical centers are aligned on a common baseline. This simplifies the production workflow as all cameras can be rectified jointly and all subsequent processing steps such as depth estimation and DIBR can be conducted line-wise.

To ensure the proper alignment and positioning of the cameras, we use a dedicated camera assistance system [10]. The two cameras in the center (cameras 2 and 3) form a narrow stereo baseline. They are used to capture content suitable for native displaying on stereoscopic displays, i.e. without the need to involve DIBR. In addition, depth estimation is simplified and produces denser depth maps with narrow baselines compared to a wide baseline setup. Cameras 1 and 4 form the wide baseline. This allows us to interpolate a high number of views without requiring artifact-intense view extrapolation.

To create appealing content for stereoscopic 3D displays, a set of 3D productions [13] rules needs to be obeyed. A major concern is to keep the shown parallax within a comfortable
viewing range [12]. To achieve this goal, the interaxial distance, i.e. the separation of the cameras’ optical axes needs to be around 3-7 cm. As standard camera bodies and lenses exceed these distances considerably, beam-splitter rigs are used within a standard stereoscopic 3D production workflow [11]. Consequently, we use a beam-splitter to mount the two cameras used for the narrow baseline. Two additional satellite cameras are mounted outside the mirror box. Figure 2 shows a picture of the multi-camera rig. The two entry pupils of the inner camera pair can be seen inside the mirror box. The satellite cameras are attached onto adjustable mounting plates.

3. MVD4 CONTENT PRODUCTION

We aim to produce multi-video plus depth content using four videos and depth maps, i.e. MVD4 content. We require the same quality for all four depth maps although we have non-equidistant stereo baselines. A priori the wide baseline stereo pairs will produce sparser disparity maps (with higher depth resolution), while the narrow baseline stereo pair will produce denser depth maps. Consequently, we need an approach to transfer the denser depth maps from the center pair to the satellite cameras. This transfer can be achieved by applying a DIBR algorithm to the depth maps. However, the raw depth maps are not suitable for DIBR. Consequently, we propose a stratified approach for the MVD4 generation which takes full advantage of the multi-camera rig configuration.

3.1 Initial Disparity Estimation

Given the four camera images (see Figure 10) which have been rectified onto a common baseline [3], three pairs of disparity maps are estimated using the disparity estimator HRM [8] along with a left-right consistency check. For the two inner cameras, we get two disparity maps, one estimated with the narrow baseline neighbor and one with the wide baseline neighbor, e.g. for camera 2, we get the disparity maps $\text{Disp}_{21}$ and $\text{Disp}_{23}$ (see Figure 3 and Figure 4).

The left-right consistency check is a suitable method to eliminate wrong disparity values which correspond to pixels which are occluded in one of the two cameras. Fortunately, for the inner stereo pair, each camera has a left and a right neighbor. Therefore, there is a portion of pixels which might be occluded in the left neighbor, but visible in the right neighbor. When merging the disparity maps, we can fill at least some of the holes which occurred due to occlusions.

The disparities within these maps should be proportional according to the camera baselines. From the calibration data we can normalize all six disparity maps into depth maps.

Figure 4 shows the two disparity maps belonging to camera 2. As expected, the disparity map $\text{Disp}_{21}$ which has been estimated using cameras 2 and 1, i.e. the wide baseline, is sparser than the disparity map $\text{Disp}_{23}$ estimated using cameras 2 and 3, i.e. the narrow baseline.

3.2 Merging of the Inner Disparity Maps

After normalization, the disparity maps can be merged, e.g. $\text{Disp}_{21}$ with $\text{Disp}_{23}$ and $\text{Disp}_{32}$ with $\text{Disp}_{34}$ (see also Figure 3).

Figure 5 illustrates the merging process of the normalized disparity maps. The result (Figure 5, right) is denser than the two input disparity maps. Ideally, the occlusions from the left and the right object borders are filled. Disparities which are inconsistent along the two disparity maps are removed.

3.3 Filtering of the Inner Disparity Maps

We apply now a cross-bilateral median filter [2,8] to the merged disparity maps. The filter is able to fill holes in the disparity map which occurred due to occluded parts of the image. However, as we have merged the disparity maps beforehand, a denser disparity map is being filtered which adds to the quality of the filtered result. Figure 6 shows the resulting filtered disparity map $\text{Disp}_{2}$ (see also see Figure 3).
As can be seen, after filtering, no holes are left in the disparity map, i.e., it is pixel dense, and the alignment with the object borders has been greatly improved.

3.4 DIBR of the Disparity Maps to Outer Views

Once we have the filtered disparity maps, we want to render them to the satellite positions. The reasons for doing this are twofold. On one hand, the initial disparity maps for the outer views are quite sparse compared to the inner views as the parallax is higher and less reliable matches can be found. On the other hand, we can apply a trifocal consistency check as each disparity map from the inner views incorporates information from both central views.

3.5 Merging of the Outer Disparity Maps

Once the disparity maps $Disp_{2}$ and $Disp_{3}$ have been rendered to the positions of the satellite cameras, we can merge the rendered disparity maps $Disp_{2} \rightarrow 1$ and $Disp_{3} \rightarrow 4$ with the initial disparity maps $Disp_{12}$ and $Disp_{43}$. As a result, we get nearly dense disparity maps for the satellite cameras. Again, inconsistent disparities are discarded.

3.6 Filtering of the Outer Disparity Maps

In a last step, the merged disparity maps need to be filtered using a cross bilateral median filter [2,8] in order to get pixel dense disparity maps.

As can be seen, after filtering, no holes are left in the disparity map, i.e. it is pixel dense, and the alignment with the object borders has been greatly improved.

4. RESULTS

We have combined the original video files from Figure 10 with the resulting disparity maps from Figure 11 to form a MVD4 frame. We got a set of four dense disparity maps which show a good alignment with the object boundaries within the four camera images. However, no useful disparity information could be extracted for some parts of the satellite cameras. The left part of $Disp_{1}$ and the right part of $Disp_{4}$ contain wrong disparities. This is due to the fact, that these parts of the scene are seen by only one camera. Consequently, these regions should be excluded from DIBR processing.

To evaluate the suitability of the data for DIBR, we generated nine views along the whole wide baseline as shown in Figure 12. The leftmost and rightmost views coincide with the original view from cameras 1 and 4. All remaining views are synthesized using an approach described in [6]. Apparently the quality of the renderings is quite convincing. Although the rendering has been performed along a wide baseline, the results are still consistent and suitable at least for an auto-stereoscopic display. For stereoscopic displays, the unmodified original views can be used, which obviously do not show any artifacts even at full HD resolution.

5. CONCLUSIONS

We have proposed a display agnostic workflow for the generation of MVD4 content. It has shown to be suitable to create high quality content for stereoscopic and auto-stereoscopic display and is adapted to a multi-camera rig with mixed narrow and wide baseline. The proposed approach for MVD4 generation is stratified and can perform in real-time, given a real-time implementation of the involved modules shown in Figure 3. A real-time demonstrator based on the proposed MVD4 generation workflow was recently shown at Siggraph 2012 [14].

6. ACKNOWLEDGEMENT

This work has been supported by the Muscade Integrated Project [9], funded under the European Commission ICT 7th Framework Programme (MIultimedia SCAlable 3D for Europe, grant agreement n°247010).
7. REFERENCES


Figure 10. Test sequence Musicians 2 shot during the 2nd Muscade [9] test shooting. Original views of the cameras 1 to 4 from left to right. A small parallax is visible for the inner stereo pair, while a considerable parallax can be observed between the center and the satellite cameras.

Figure 11. Final depth maps after all processing steps. In combination with the video images from Figure 10, they form an MVD4 frame.

Figure 12. Virtual views rendered along the wide baseline. From leftmost view in the top column coincides with the original camera 1, while the rightmost view in the bottom column coincides with the original camera 4. All other views are synthetized.