3D-TV LDV CONTENT GENERATION WITH A HYBRID TOF-MULTICAMERA RIG

A. Frick, B. Bartczack and R. Koch

Christian-Albrechts-University Kiel
Computer Science Department
Hermann-Rodewald-Str. 3, 24118 Kiel, Germany
{africk,bartczak,rk}@mip.informatik.uni-kiel.de

ABSTRACT

In this paper we introduce an approach for automatic generation of Layered Depth Video (LDV) content for 3D television (3DTV) purposes. The proposed approach relies on a capturing system consisting of two time of flight cameras (ToF) and five color cameras. The LDV content is generated for the central color camera, based on the depth images provided by the ToF cameras and texture information provided by the color cameras. Through different viewpoints of the ToF cameras and color cameras, we are able to reconstruct depth and texture information not only for foreground objects, but also in the occluded areas. The estimated information is then combined to one consistent LDV representation.

Index Terms — Time of Flight, Layered Depth Video, Depth Image Based Rendering, Bilateral Filter, High Definition

1. INTRODUCTION

Future 3D-TV systems will enhance the user experience beyond HD-TV, because they offer to perceive scene depth and allow realistic user immersion into the scene. Technically, the depth impression arises through observing the scene from slightly different view points. The viewpoint difference then causes a distance dependent shift of the scene objects which is interpreted by our brain as depth. In order to achieve 3D experience, one must provide the viewer with at least two different images of the same scene.

In standard stereo approaches, the 3DTV content can be generated directly by capturing the scene with two cameras with slightly different view points. The images are then presented separately to the left and right eye of the viewer enabling a 3D impression [1, 2]. While such approaches have their advantages, they in general lack the flexibility in display choice and are limited to only two views. In contrast to standard stereo, depth based approaches extend the color information with additional depth information representing the scene geometry [2, 3]. Using depth-image-based rendering methods (DIBR,) novel views can then be generated [4].

Depth based methods overcome the difficulties of the standard stereo by being independent from display geometry and able to generate multiple virtual views, but are in general more challenging in content creation.

To limit the complexity or simply through the lack of data, many DIBR approaches try to use only one color image and one depth image representing the foreground scene geometry. This strategy inevitably leads to annoying artifacts in disocclusion regions appearing in a virtual view, for which no texture and depth information exist.

To overcome this limitations different pre- and post processing techniques were proposed, all of which succeed to some degree, but fail when the disocclusion regions become bigger and more complex in both the revealed texture and geometry [5, 6].

Another way to handle the disocclusion regions is to use more sophisticated formats like LDV or MVD, which provide additional depth and texture information for the occluded areas [7, 8].

In this paper, we propose using the LDV format, as it has its advantages in being more compact and compatible to a standard video. The LDV format extends the standard simple one color one depth image representation by additional color and depth images called occlusion layers. In these layers, the background texture and depth information is coded, so that the disoccluded areas can be properly filled in a virtual view.

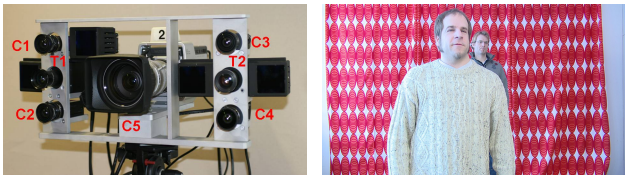
The consistancy in texture and depth between the layers is very crucial for the rendering quality, requiring very reliable and consistent depth images. However, depth estimation from color images is a difficult and error prone process especially on object boundaries and in textureless regions. Using the ToF camera yields the advantage of getting reliable depth information in these regions without complicated estimation.

In our previous work we demonstrated the use of a ToF camera for LDV generation [9]. However, the occlusion layer was generated through a simple extrapolation method, which can cause significant errors. In this paper, we use a modified camera setup with two integrated ToF cameras and demonstrate that we are able to generate a more reliable occlusion layer based largely on real depth measurements in occluded areas.

2. CAMERA SETUP AND CALIBRATION

The camera setup we use for capturing, consists of five color cameras $C1$,...,$C5$ and two ToF cameras $T1$ and $T2$ (Fig. 1 (a)) and is described thoroughly in [10]. Here, we only give a short description of the system. The middle camera $C5$ is a Sony X300 camera with a resolution of 1920×1080 pixel. The cameras $C1$,...,$C4$ are Point Grey Grasshoppers with a resolution of 1600×1200 pixel at 180 mm distance from the middle camera. Both ToF cameras are Pmd - CamCube cameras with resolution of 204×204 pixel. The ToF cameras are active range sensors that emit modulated infrared light at 20 MHz and 21 MHz frequency and measure the

This work was partially supported by the German Research Foundation (DFG), KO-2044/3-1 and the Project 3D4YOU funded under Seventh Framework Programme, Theme ICT-2007.1.5 Networked Media, Grant 215075.



(a) Camera setup (b) View of the central camera

Figure 1. (a) Camera setup: $C1$, ..., $C4$ are Point Grey Grasshopper Cameras, $C5$ is Sony X300, $T1$ and $T2$ are PMD CamCube cameras (ToF) (b) view of the central camera

time per pixel that the light needs to come back by special correlation elements. The measured time combined with light speed yields the depth information for each pixel (Fig. 2). In addition, a reflectance image measures the reflected IR amplitude.

(a) Left Pmd (b) Right Pmd

Figure 2. Left and right pmd depth images; resolution 204×204 pixel.

ToF cameras have an aspect ratio of 1:1 and are slightly rotated outwards to achieve the full coverage of the view of the central camera in 16:9 format. Compare the view of central camera (Fig. 1 (b)) with the views of both ToF cameras (Fig. 2).
Due to bad signal to noise ratio (SNR) of the reflectance image, low pixel resolution, and systematic errors in the depth measurement, the ToF cameras are difficult to calibrate with traditional calibration methods. We use therefore the method proposed in [11] to calibrate the ToF cameras together with multiple CCD cameras in a joint method, which already proved itself as very reliable in our previous work [9].

3. TOF UPSCALING

ToF cameras capture depth at the very low resolution of 204×204 pixel. To generate LDV in HD resolution of 1920×1080 pixel, ToF data must be transformed (warped) in the view of the central camera $C5$, requiring to upscale the image by factor 10.
In order to minimize warping errors and for later generation of occlusion layer, we don't go to the central camera directly but warp ToF data first to the corner cameras $C1$, ..., $C4$. So in a first step: we warp the ToF camera $T1$ to the color cameras $C1$ (Top) and $C2$ (Bottom) and the ToF camera $T2$ to the color cameras $C3$ (Top) and $C4$ (Bottom) according to the method described in [12].
Due to the smaller distance 70 mm to the top and bottom cameras in contrast to 180 mm to the central camera, this warping operation causes in general smaller errors compared to the warping to the central camera directly. However artifacts caused by the resolution change and occlusions on the object boundaries are still clearly visible and very disturbing. To correct such artifacts

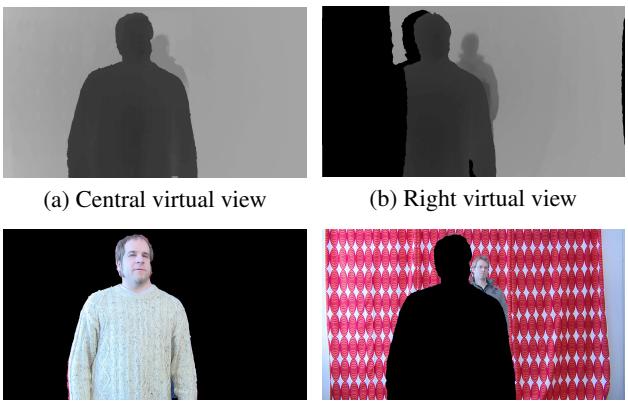
(a) Right ToF (T1) warped to $C2$. (b) Image of a mask.

Figure 3. (a) Warped Pmd depth image after color alignment at a resolution of 1600×1200 pixel (pixel without depth marked in black); (b) Image of a mask held by the person in the back (from C1)

we perform a color alignment between the warped depth map and the corresponding color image, based on bilateral filtering, as proposed in [13] and as we have already shortly described in our work in [9]. Figure 3 (a) shows warped depth image from camera $T1$ to the view of camera $C2$ after color alignment.

4. GENERATION OF LDV

After the ToF upscaling step, we have depth images Di (Fig. 3 (a)) and the corresponding color images Ii in high resolution of 1600×1200 pixel for each of the color cameras $Ci, i \in \{1, 2, 3, 4\}$.



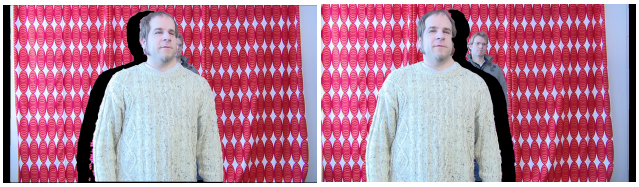
(a) Central virtual view (b) Right virtual view

(c) Foreground object (d) Background

Figure 4. (a) Depth image generated for the Central virtual view; (b) Depth image generated for the Right virtual view; (c) Foreground person segmented through depth thresholding on [a]; (d) Background segmented through depth thresholding on [a]; pixels without depth or color information are coded as black

Using the four pairs (Di, Ii) we are now able to generate (render) three virtual views, Left, Central and Right, lying on the baseline corresponging to the x - axis of the central camera. Whereby the Central view is identical to the view of the central camera $C5$ and Left and Right views are shifted by 100 mm on the x - axis to the left and right of the central view. For the right view, we use the pairs $(D1, I1)$, $(D2, I2)$ and for the left view the pairs $(D3, I3)$, $(D4, I4)$. For the central view all pairs are used.
The depth map of the Central virtual view is aftwerwards color aligned to the original image of the central camera and results in the foreground layer depth of the LDV.

Figure 4 shows depth images for Central and Right virtual views (a,b) and a segmentation of the foreground object based on depth thresholding (c,d). The depth image of the right virtual view is generated only from cameras $C1$ and $C2$ and therefore has no depth information in the left occlusion regions as in the outer left part of the image. The opposite applies to the left virtual view.
Using only the foreground layer, we are able to generate novel views, however without proper occlusion information, the disocclusions can be only handled by extrapolation. As we have mentioned in the introduction, this can succeed only if the disocclusion areas are small and contain only simple texture and geometry.
In our experimental scenario, there are mainly four depth levels: a foreground person, a mask held by the second person behind the first person, the second person behind the mask, and the background with a repetitive pattern. See figure 3 (b) to get an impression of the scene setup. Note that the mask held by the second person is completely hidden from the central camera.
Thereby we will not be able to handle disocclusions properly using only extrapolation methods described in [5, 6, 9]. Figure 5 shows an example of two virtual views generated from the foreground layer without occlusion information.



(a) Left novel view (b) Right novel view

Figure 5. Outer left and right views generated from LDV through warping; pixels without texture information are coded in black

In order to handle the disocclusions properly, additional information for occluded areas in the central camera is needed. In the LDV format such information is contained in the occlusion layer. To generate the occlusion layer, we warp the foreground depth image to the Left and Right virtual views generated previously from image pairs $(D1, I1), (D2, I2)$ for the Right view and $(D3, I3), (D4, I4)$ for the Left view.
By doing so, the disocclusion regions are revealed (see Figure 5). However, due to the viewpoint difference between $C1$,...,$C4$ and $C5$ all the necessary information is now available in the Left and Right virtual views. We simply mask all relevant depth and texture information by the disocclusion areas coded in black and transform this to the view of the central camera. This information is then combined to the occlusion layer.
Note that by rendering a virtual view from LDV, the reverse operation will be performed, so that disocclusions are filled with the correct texture information. Figure 6 shows the occlusion layer generated for the experimental scene presented before.

5. RESULTS AND DISCUSSION

For the evaluation purposes multiple virtual views, left and right from the central view, were generated. Figure (7 (a,b,c,d)) shows some results at a distance of -100, +50 and +100 mm from the central view, together with the original image of the central view.
For virtual view generation, we use a simple depth-image-based rendering method. We first transform the foreground and then the occlusion layer according to the simple scheme below.

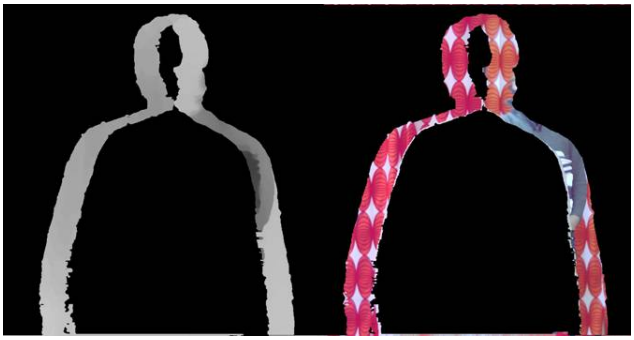


Figure 6. Generated occlusion layer for the view of the central camera; depth image left and corresponding texture right; invalid pixels are marked in black

First, the depth image of a layer is warped to the new view according to [12]. Second, the depth image is median filtered to elminate some small artifacts that may appear at object borders and the corresponding texture is then warped to ithe new view through backward mapping. After both layers are transformed, a median filter is applied to the texture on the occlusion borders to reduce texture stitching artifacts.
This rendering method is simple but as we believe sufficient to demonstrate the main contributions of our work.
As one can see in figure (7 (a,c,d)), the foreground and background textures fit quite well together on the object borders, where the disocclusions were filled with the texture from the occlusion layer. Note that the problematic region on the right side of the foreground person is also reconstructed quite well. One can clearly see a part of the mask held by the second person. This is certainly not possible with pure extrapolation methods and clearly demonstrates the benefit of our approach.
There are still some errors due to false depth estimation on the object boundaries and slight misalignment of foreground and background textures but the results clearly show that we are able to automatically generate a consistent LDV content.
One may have also noticed a difference in color between left and right occlusion textures. This is due to not proper color adaptation of different cameras used in the introduced capturing system and is only a technical problem, which we plan to handle properly in our future work.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a new camera system and demonstrated that we are able to automatically generate a reliable depth information especially in the occluded areas. Further more, we have demonstrated that the calculated occlusion layer is more reliable, than that based on pure extrapolation methods and that we are able to successfully combine the estimated foreground and background layers to one consistent LDV representation.
In our future work we want to improve the depth estimation on object borders. Our special interests lie thereby in the fusion of stereo matching methods and ToF cameras as for example proposed in [14]. We also want to improve the texture consistancy between the foreground and background layer especially in conjuction with more sofisticated LDV rendering methods.

7. REFERENCES

[1] C. Girdwood and Ph. Chiwy, "MIRAGE: An ACTS Project In Virtual Production And Stereoscopy," in *Broadcasting Convention, International*, 12-16 Sep 1996, pp. 155–160.

[2] M. Ziegler et al., "Evolution of stereoscopic and three-dimensional video," *Image Communication*, vol. 14, November 1998.

[3] A. Redert et al., "Advanced three-dimensional television system technologies," *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pp. 313–319, 2002.

[4] Liang Zhang and W.J. Tam, "Stereoscopic image generation based on depth images for 3d tv," *Broadcasting, IEEE Transactions on*, vol. 51, no. 2, pp. 191–199, June 2005.

[5] W. J. Tam et al., "Smoothing depth maps for improved steroscopic image quality," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, B. Javidi & F. Okano, Ed., Oct. 2004, vol. 5599 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 162–172.

[6] Sang-Beom Lee and Yo-Sung Ho, "Discontinuity-adaptive depth map filtering for 3d view generation," in *IMMERSCOM '09: Proceedings of the 2nd International Conference on Immersive Telecommunications*, ICST, Brussels, Belgium, Belgium, 2009, pp. 1–6, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[7] Jonathan Shade et al., "Layered depth images," in *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 1998, pp. 231–242, ACM.

[8] P. Merkle et al., "Multi-view video plus depth representation and coding," *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 1, pp. I –201–I –204, 16 2007-Oct. 19 2007.

[9] A. Frick et al., "Generation of 3d-tv ldv-content with time-of-flight camera," in *3DTV09*, May 2009, pp. 1 –4.

[10] A. Frick et al., "Realtime preview for layered depth video in 3d-tv production," to be published in SPIE 2010, "Real-Time Image and Video Processing".

[11] Ingo Schiller et al., "Calibration of a pmd camera using a planar calibration object together with a multi-camera setup," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Beijing, China, 2008, vol. Vol. XXXVII. Part B3a, pp. 297–302, XXI. ISPRS Congress.

[12] Bogumil Bartczak et al., "Integration of a time-of-flight camera into a mixed reality system for handling dynamic scenes, moving viewpoints and occlusions in real-time," in *Proceedings of the 3DPVT Workshop*, Atlanta, GA, USA, June 2008.

[13] Qingxiong Yang et al., "Spatial-depth super resolution for range images," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007.

[14] B. Bartczak and R. Koch, "Dense depth maps from low resolution time-of-flight depth and high resolution color views," in *ISVC09*, 2009, pp. II: 228–239.

(a) Most outer left view (-100 mm from cenral view)

(b) Central view

(c) Right view (+50 mm from central view)

(d) Most outer right view (+100 mm from central view)

Figure 7. Generated virtual views together with original central view