DIBR-BASED 3D VIDEOS USING NON VIDEO RATE RANGE IMAGE STREAM

Xiaoyi Jiang[1] *and Martin Lambers*[2]

[1]Department of Mathematics and Computer Science, University of Münster, Germany
Email: xjiang@math.uni-muenster.de
[2]Department of Electrical Engineering and Computer Science, University of Siegen, Germany
Email: lambers@fb12.uni-siegen.de

ABSTRACT

The fundamental assumption of 3D videos using depth-image-based rendering is the full availability of range images at video rate. In this work we alleviate this hard demand and assume that only limited resources of range images are available, i.e. corresponding range images exist for some, but not all, color images of the monoscopic video stream. We propose to synthesize the missing range images between two consecutive range images. Experiments on real videos have demonstrated very encouraging results. Especially, one 3D video was generated from a 2D video without any sensory 3D data available at all. In a quality evaluation using an autostereoscopic 3D display the test viewers have attested similar 3D video quality for our synthesis technique and rendering based on depth ground truth.

1. INTRODUCTION

Recently, an advanced concept of depth-image-based rendering (DIBR) has been proposed for 3D videos [1]. Using a single stream of monoscopic images and a second stream of range (depth) images, a high-quality stereoscopic stream for any nearby viewpoint is synthesized in such systems. Compared to the end-to-end stereoscopic video stream, this concept has a number of advantages [1]: backward compatibility with existing 2D video systems; flexibility (optimal 3D effects customized to different 3D displays and user needs; support of multiview 3D displays); efficiency (coding and transmission of the range video stream cheaper than a monoscopic video stream). The most important components of DIBR-based 3D videos are: content generation, coding, transmission, virtual view synthesis and 3D display. Our current work is devoted to content generation.

3D content generation: The fundamental assumption in DIBR is the availability of range images at video rate. This can be achieved by a real-time 3D camera [1]. The practical value of this approach, however, is still limited. Currently, only very few range cameras deliver real-time range videos and their use is typically restricted by limiting factors such as operational environment (indoor, outdoor) and ranging area (angular field of view, depth of field). In addition high-rate range cameras tend to be expensive and this hinders their use in a broad range of applications.

Alternatively, depth information can be computed from a single image [2]. Several classes of shape-from-X methods follow this goal, for instance shape-from-shading. So far, however, there is still no proof of their practical usefulness.

A third way of 3D content generation is recovery from 2D videos. Despite of the advances in the past, automatic 3D reconstruction remains a tough challenge [3]. Some approaches require an object segmentation [4] which causes additional uncertainty to the difficult recovery task.

Our approach: We assume that only limited resources of range images are available, i.e. corresponding range images exist for some, but not all, color images of the monoscopic video stream, and propose to synthesize the missing range images between two consecutive range images. This allows to: a) Ease the recording of 3D material. Instead of using expensive video-rate range sensors it is possible to use cheaper sensors that generate less range images and complete the missing range images automatically; b) Enhance existing 2D video material with 3D effects by automatically completing depth information from a few, possibly manually created, range images. Given the vast amount of existing 2D material, this is an important application.

The basic idea of our approach is to estimate motion in the monoscopic video stream and to apply this motion information for synthesizing the missing range images. The technical details are described in next section. Given a color image and its corresponding range image, a stereo pair can be synthesized by a special 3D image warping technique [1, 5]. Experimental results are reported in Section 3. Finally, we conclude the paper with some discussion.

2. SYNTHESIZING RANGE IMAGES

It is assumed that a monoscopic (color) video stream is given by n frames $F_0, \ldots, F_{n-1}$, along with depth information D_0 and D_{n-1} for the first and the last frame only. The goal is to expand the depth information from D_0 and D_{n-1} so that a complete set of range images $D_0, \ldots, D_{n-1}$ is available for the subsequent depth-image-based rendering step.

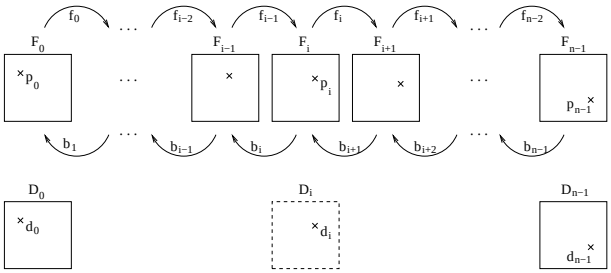

Fig. 1. Depth tracking from motion analysis

Range image synthesis by depth tracking: The basic idea is to track each point in the scene as it moves to different pixel positions from frame to frame. For a point P at position $p_i \in F_i,\ 0 < i < n-1$, the corresponding positions $p_0 \in F_0$ and $p_{n-1} \in F_{n-1}$ are then known, and therefore also the associated depths d_0 and d_{n-1}. The depth d_i required for the unknown depth map D_i can then be computed by an interpolation of d_0 and d_{n-1}.

To be able to compute the position $p_{i+1} \in F_{i+1}$ of a point P with a given position $p_i \in F_i$, it is necessary to know the *forward motion* vector $f_i(p_i)$: $p_{i+1} = p_i + f_i(p_i)$. Similarly, the position p_{i-1} in F_{i-1} of the point can be computed when the *backward motion* vector $b_i(p_i)$ is known: $p_{i-1} = p_i + b_i(p_i)$. To handle all points of the scene, the forward and backward motion vector fields $f_0, \ldots, f_{n-2}$ and $b_1, \ldots, b_{n-1}$ that contain motion vectors for each pixel position must be computed. The complete process of depth tracking from motion analysis is illustrated in Figure 1. Details will be given later in this section.

Based on the depth tracking the missing range images are synthesized in the following way. A point P with a given position p_i in frame F_i is tracked backwards to some position p_0 in frame F_0 and forwards to some position p_{n-1} in frame F_{n-1}. We distinguish between three cases: 1) Both backward and forward tracking are successful: A linear interpolation of the corresponding depth values d_0 and d_{n-1} is performed to determine d_i; 2) It can only be tracked to one of both end frames: d_i is set to the depth value at this end frame, thus assuming that it remains constant over time; 3) It can neither be tracked to F_0 nor to F_{n-1}: the depth d_i is arbitrarily set to "far". To avoid too frequent occurrences of such untrackable situations, it is necessary that the observed scene does not change too much, so that roughly the same objects are visible in F_0 and F_{n-1}, albeit at different positions. If this requirement is not fulfilled, the scene may have to be divided into sub-scenes.

Details of depth tracking: The key part of depth tracking is the per-pixel motion estimation: The more precise it is, the better the depth approximation. We have experimented with two approaches to compute motion vector fields: optical flow and block matching.

Optical flow: The per-pixel motion vector fields are in fact dense optical flow fields. We have used the local method of Lucas-Kanade (LK), the global method of Horn-Schunck (HS), and the recent combined local/global approach (CLG) [6], which has both the high robustness of local methods and the full density of global techniques.

Block matching: Block matching techniques are commonly used in feature tracking applications and in stereo correspondence search: to find a match for the pixel at position p in frame F_0 at some position q in frame F_1, a block of size $(2k+1) \times (2k+1)$ around p is examined, and the best match for this neighborhood is searched in F_1. If $\hat{q}$ is the position of the match candidate currently under consideration, then its matching costs are defined as:

$$C(p, \hat{q}) \ = \ \sum_{r=-k}^{k} \sum_{c=-k}^{k} c(p + (r, c), \hat{q} + (r, c)) \qquad (1)$$

The candidate position $\hat{q}$ with the lowest matching costs wins. The cost function c differs between various block matching variants.

One popular cost function uses the absolute difference of pixel values: $c_{\text{SAD}}(p, q) \ = \ |F_0(p) - F_1(q)|$. This expression can be easily extended to handle color images. The YUV color space is widely used in video processing applications. The Y component represents the brightness of a point, and the U and V components define its hue and saturation[1]. Thus, the following term can be used:

$$c_{\text{SAD}}(p, q) = L \cdot |Y(F_0(p)) - Y(F_1(q))| + (1 - L) \cdot \tfrac{1}{2} \cdot \\ (|U(F_0(p)) - U(F_1(q))| + |V(F_0(p)) - V(F_1(q))|)$$

Each of the components Y, U, V is expected to be in $[0, 1]$ in this equation. L is the luminance weight: It determines how much influence luminance differences should have in comparison to color differences.

To reduce the uncertainty of the method, our SAD block matching variant for depth tracking uses the following matching cost function, which is an extension of Eq. (1):

$$C_{\text{SAD}}(p, \hat{q}) = D \cdot \tfrac{\text{distance}(p, \hat{q})}{k} + \\ (1 - D) \cdot \sum_{r=-k}^{k} \sum_{c=-k}^{k} c_{\text{SAD}}(p + (r, c), \hat{q} + (r, c)) \qquad (2)$$

The additional term is a distance penalty: Larger motion vectors cause higher matching costs. The parameter D determines the balance between the distance penalty and the original matching score. The distance penalty reduces the uncertainty, for example in areas with periodic textures, where good matches are found at multiple positions.

We also considered a locally adaptive support-weight technique for block matching [7]. For space limitation the details are omitted.

[1] Though not in the direct way as for example the HSL color space does.

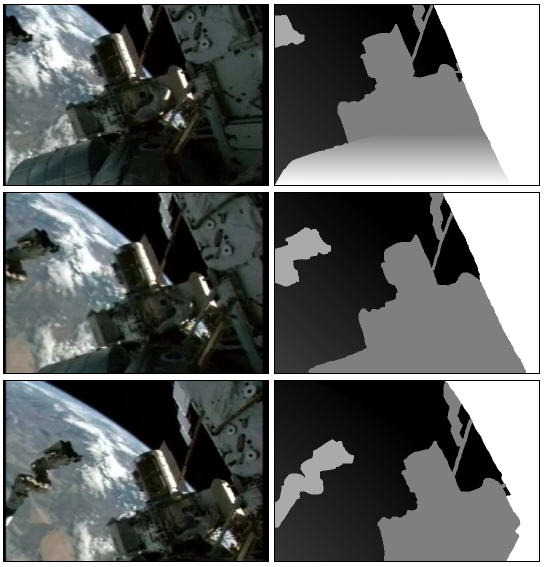

Fig. 2. *Nasa* video: first, middle, and last image with manually specified depth data

Consistency check: A postprocessing step often used for stereo correspondence search is adapted to improve the motion vector fields for depth tracking, regardless of the motion estimation method they were created with. To catch unreliable motion vectors, the motion estimation is done in both directions: from F_0 to F_1, leading to the vector field f, and from F_1 to F_0, leading to the vector field b. The reliability of a motion vector v in f will be high if the corresponding motion vector in b points back to the position of v or near to it. A threshold t determines the maximum allowed difference for vectors to be considered reliable. If the difference is greater, the vector v in f is marked as unreliable. In a second step, all unreliable vectors in f are replaced by interpolating neighboring reliable vectors. This is done in a way that ensures that vectors with a high number of reliable neighbors are replaced first, to avoid propagating errors as much as possible. The result is an improved vector field f^*. By swapping the roles of f and b, the same can be done with b, leading to an improved field b^*.

3. EXPERIMENTAL RESULTS

Three example videos were used. The videos *Interview* (10 seconds, 251 frames) and *Orbi* (5 seconds, 126 frames) are widely used in DIBR literature. They have known depth maps for each video frame, which can be used as ground truth when evaluating the computed depth data.

The third scene *Nasa* (18 seconds, 451 frames) is part of a NASA mission video[2]. It is a conventional 2D video *without* any depth data. Three simplistic, qualitative range images were created manually with minimal efforts: one for the first frame, one for the middle frame, and one for the last frame, resulting in a distance of 225 frames between two known depth maps; see Figure 2. The range images only use four different range levels: far (background), middle (the shuttle part in the center), near (the part on the left), and very near (the part on the right).

[2]`http://spaceflight.nasa.gov/gallery/video/shuttle/sts-114/qtime/114_fdh05_clip3.mov`

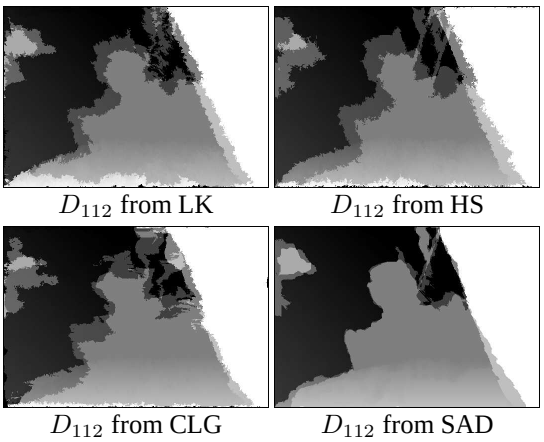

Fig. 3. *Nasa* video: depth maps D_{112} from depth tracking with different motion estimation methods

Evaluation of depth tracking: We report some results of evaluating the depth tracking quality. Figure 3 shows the depth images D_{112} that result from the different motion estimation methods when using the first depth image D_0 and the middle one D_{225} of the *Nasa* video. While all optical flow approaches have their problems with this video, the block matching method manages to follow object boundaries quite precisely across a relatively long distance of 226 frames (9 seconds).

Since the videos *Interview* and *Orbi* have depth ground truth, quantitative measures can be computed for evaluating the depth tracking quality. For the *Interview* video, for instance, the frames $F_0, \ldots, F_{250}$ (ten seconds) were extracted, along with real depth data $D_0, D_{25}, \ldots, D_{250}$ for every 25th frame. The missing depth maps were computed using depth tracking with the motion estimation methods. Looking at the synthesized depth images, the SAD block matching method seems to deliver the most accurate depth map in this case as well. The differences between ground truth and the computed depth maps can be used as an error measurement:

$$E_{\mathrm{DT}} = \frac{\sum_{i=0}^{n-1} \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} |D_i(x,y) - \mathrm{GT}_i(x,y)|}{nNM}$$

This average depth error value is in $[0, 255]$ (the smaller, the better). Table 1 shows the results for the complete series of depth maps. The depth maps with the lowest errors according to this measurement are the ones from the SAD block matching method, confirming our impressions. Similar behavior has been observed for the video *Orbi*.

Rating of 3D video quality: Several 3D video variants

	HS	LK	CLG	SAD
error E_{DT}	2.32	2.46	2.30	1.74

Table 1. Error values for all *Interview* depth maps created with the motion estimation methods

of the videos *Interview*, *Orbi*, and *Nasa* were computed from different depth data, and prepared for display on a 2018XLQ 19" 3D monitor from DTI[3], which is an autostereoscopic display, allowing 3D viewing experiences without the need of wearing any glasses. Interested readers can find a summary of (mostly commercially available) autostereoscopic displays at http://www.stereo3d.com/displays.htm.

A group of 10 test viewers was asked to rate both the image quality and the quality of the 3D effect of each variant, on a scale from 0 ("very bad" or "nonexistent") to 10 ("excellent"). A note handed out to each test viewer clarified that image quality means the absence of noise and distortion in the image, and 3D effect quality means the impression of real depth. The viewers did not have any experience with 3D videos, and they did not know anything about the nature of the videos and their variants.

The following three variants were tested:

1. *2D:* The original 2D video. Presenting this variant allows to measure the impact that depth-image-based rendering has on image quality.
2. *Ground Truth:* A 3D video based on real depth maps. This variant is expected to show the best results in terms of 3D effect quality. For the *Nasa* scene, this option was not used due to the lack of real depth data.
3. *Depth Tracking:* A 3D video based on depth data that was computed using the depth tracking method. For *Interview* and *Orbi*, every 25th depth map from ground truth was used as initial depth data. For *Nasa*, the three artificial depth maps were used.

The rating results are shown in Figure 4. Comparing the rating of the *2D* variant with the results of other variants shows that the image quality always suffers a little from depth-image-based rendering. This effect may be reducible by choosing better parameters for the rendering step, but this was not subject of the test. Most viewers noticed the absence of any 3D effect in the *2D* variant. Remarkably, the 3D videos synthesized by depth tracking was rated similar to those from the ground truth depth information with respect to both image quality and 3D effect.

4. CONCLUSIONS

In this paper we have considered a range image synthesis technique for reducing the need of full availability of a range video stream in DIBR-based 3D video creation. Our approach eases the recording of 3D material by using less expensive range sensors and enables to enhance existing 2D video material with 3D effect by limited manual overhead. Experiments on three videos have demonstrated very encouraging results. Especially, one 3D video was generated from a 2D video without any sensory 3D data available at all. In all cases the test viewers have attested similar 3D video quality for our synthesis technique and rendering based on depth ground truth.

[3] http://www.dti3d.com/

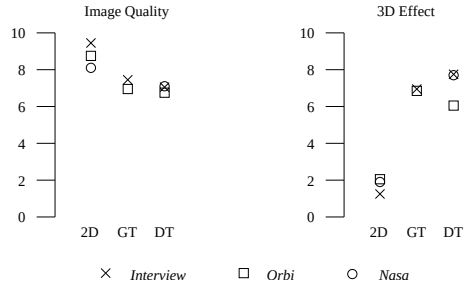

Fig. 4. Rating of image quality and 3D effect for the variants *2D*, *Ground Truth*, *Depth Tracking*

5. REFERENCES

[1] C. Fehn et al., "Key Technologies for an Advanced 3D-TV System," in *Proceedings of SPIE Three-Dimensional TV, Video and Display III*, Philadelphia, 2004, pp. 66–80.

[2] S. Battiato et al., "3D Stereoscopic Image Pairs by Depth-Map Generation," in *Proceedings of 3D Data Processing Visualization and Transmission*, 2004, pp. 124–131.

[3] M. Pollefey, "3D from Image Sequences: Calibration, Motion and Shape Recovery," in *Handbook of Mathematical Models in Computer Vision*, N. Paragios et al., Ed. 2006, pp. 389–403, Springer.

[4] K. Moustakas et al., "A Non Causal Bayesian Framework for Object Tracking and Occlusion Handling for the Synthesis of Stereoscopic Video," in *Proceedings of 3D Data Processing Visualization and Transmission*, 2004, pp. 147–154.

[5] L. Zhang and W. J. Tam, "Stereoscopic Image Generation Based on Depth Images for 3D TV," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 191–199, 2005.

[6] A. Bruhn et al., "Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[7] K.-J. Yoon and I.-S. Kweon, "Locally Adaptive Support-Weight Approach for Visual Correspondence Search," in *Proceedings of IEEE Conference on CVPR*, San Diego, CA, USA, 2005, vol. 2, pp. 924–931.