Ulrich Fecker, Audrey Guenegues, Ingo Scholz, Andr'e Kaup

Depth Map Compression for Unstructured Lumigraph Rendering

appeared in:
Proceedings Visual Communications and Image Processing 2006
San Jose, USA
pp. 60770B-1 – 60770B-8

Depth map compression for unstructured lumigraph rendering

Ulrich Fecker[a], Audrey Guenegues[a], Ingo Scholz[b] and André Kaup[a]

[a]University of Erlangen-Nuremberg
Chair of Multimedia Communications and Signal Processing
Cauerstraße 7, 91058 Erlangen, Germany

[b]University of Erlangen-Nuremberg
Chair for Pattern Recognition
Martensstraße 3, 91058 Erlangen, Germany

ABSTRACT

Image-based rendering techniques require capturing an object or scene from many viewpoints. Often, depth maps are used in addition to the original images for rendering new views for novel viewpoints not coinciding with one of the original camera positions. Due to the high amount of data, efficient compression is necessary. When the data is stored or transmitted, it is not only desirable to compress the image data but also the depth information. In this paper, the case of sequences recorded with a single, hand-held camera is investigated. These sequences are used for unstructured lumigraph rendering. In addition, the case of multi-view video sequences is analyzed. For both cases, depth maps are compressed using H.264/AVC, and the achievable data rates are studied. For the case of unstructured lumigraph rendering, the effect of depth map compression on the quality of rendered images is analyzed in a second step.

Keywords: Image-based rendering, light field, lumigraph, depth map, image compression, video coding

1. INTRODUCTION

Image-based rendering is an approach where an object or scene is recorded from various positions. From the recorded images, new views can be generated for camera positions not coinciding with the recording positions. Ideally, the user can move around freely and watch the scene from any desired viewpoint and viewing angle. Examples for image-based rendering techniques are the light field[1] and the lumigraph[2] approaches. Due to the high amount of data, it is reasonable to store or transmit the image data in a compressed form. This has e. g. been studied in Ref. 3 and Ref. 4.

Many view interpolation techniques used for image-based rendering rely on depth maps which store the distance between the camera position and the surface of the object for each pixel or block of each recorded image in the dataset. Here, we assume that depth maps have been generated containing a depth value for each pixel in the image. Therefore, the depth maps can be depicted as gray-level images, and an additional depth image is assigned to each image in the originally recorded sequence.

When a light field or lumigraph is to be stored or transmitted, it is often desirable to additionally store or transmit the depth information. By doing so, the receiver can use the depth maps without having the computational effort of generating them. In addition, the depth information is more precise in this case, as it is then based on the original, uncompressed images. If no compression is applied to the depth maps, this leads to a severe increase in the storage requirements or the data rate on the transmission channel. Therefore, it is desirable to compress not only the original light field images, but also the depth information.

E-mail: (Send correspondence to U.F.)
U.F.: fecker@lnt.de
A.G.: gueneg@lnt.de
I.S.: scholz@informatik.uni-erlangen.de
A.K.: kaup@lnt.de

Here, we especially analyze the case of unstructured lumigraph rendering.[5] For the coding experiments, test sequences were used which were recorded using a single, hand-held camera. They show static objects. While recording, the camera was moved so that images showing the objects from various positions were obtained. Due to the manual movement of the camera, the recording positions are not aligned in a regular pattern but scattered along the irregular path on which the camera has been moved.

From the recorded and calibrated image sequences, depth maps were estimated. Using the images and depth information, new views can be generated. Further information on the calibration and rendering techniques can be found in Ref. 6.

For comparison, we also study the compression of multi-view video sequences. This case is equivalent to dynamic light fields, where the recorded object or scene changes over the time, and therefore a whole video sequence needs to be recorded for each camera position. This case has attracted increasing interest, especially for being applied in three-dimensional television (3D TV) or free-viewpoint television (FTV).

In Ref. 7, Fehn et al. compared different MPEG codecs concerning their coding performance when they are applied to depth maps. The results show that the H.264/AVC codec significantly outperforms older MPEG standards such as MPEG-2 or MPEG-4 for the compression of depth images. That is why H.264/AVC was also chosen for our coding experiments.

In a first step, we compress the depth maps and analyze which compression factors and data rates can be achieved. For that, the quality of the decoded depth maps is compared to the uncoded depth maps in terms of PSNR. This is done for the case of sequences acquired by a hand-held camera used for unstructured lumigraph rendering, as well as for the case of multi-view sequences.

In the rendering process, the depth maps are used to reduce or eliminate artifacts that appear when the original images are interpolated in order to generate the image seen from a novel viewpoint. Wrong depth information causes the original images to be misaligned, which results in a blurring of the rendered image. Therefore, we analyze in a second step how the compression of the depth information affects the rendering quality. New images are rendered using the decoded, distorted depth maps and compared to images rendered using the original depth maps.

2. TEST SEQUENCES

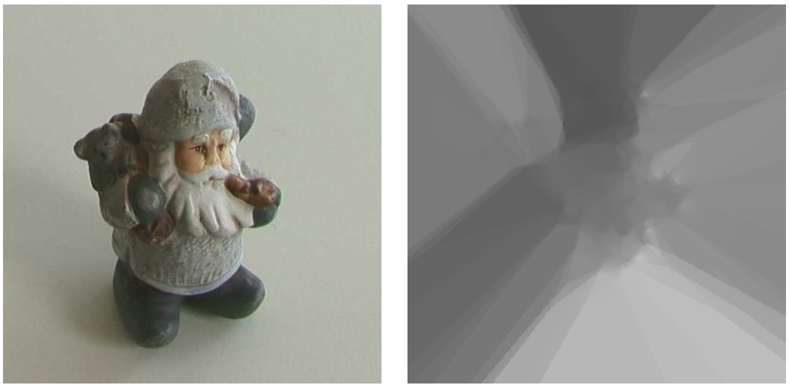
2.1. Hand-Held Camera Sequences for Unstructured Lumigraph Rendering

Exemplarily, four test sequences are used in this paper to show the coding results. Two of them, "Santa" and "milk", cover the case of unstructured lumigraph rendering. They show static objects and were recorded with a hand-held camera which was moved during the recording process, so that views from different positions were obtained. They both have a resolution of 512×512 pixel and a frame rate of 15 frames per second. The "Santa" sequence consists of 67 frames, the "milk" sequence of 190 frames. Figure 1(a) contains the first frame of the "Santa" sequence, Fig. 2(a) the first frame of the "milk" sequence.

In Fig. 1(b) and 2(b), the corresponding depth maps are shown. They are generated from the 3-D reconstruction of the scene geometry. The resolution is the same as for the images, since a corresponding depth value is assigned to each pixel. As it can be seen, the depth maps are rather coarse and do not show much detail. However, they are sufficient for significantly increasing the rendering quality.[6, 8]

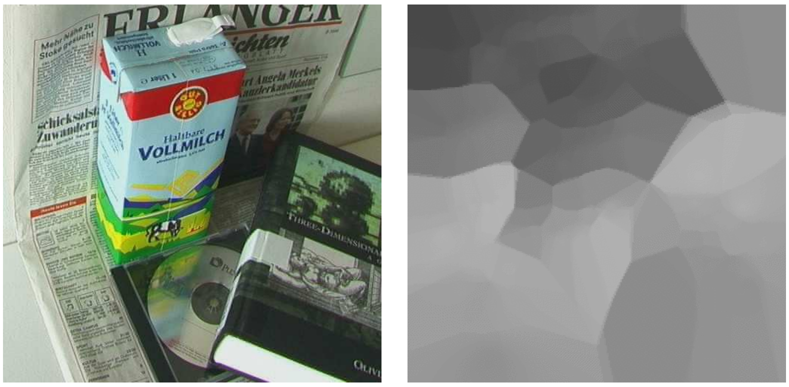
2.2. Multi-View Test Datasets

Two other test sequences, "ballet" and "breakdancers", have been recorded using a multi-camera system consisting of eight video cameras. They therefore show a dynamic scene from eight different viewpoints and serve as examples for multi-view video sequences. These sequences and the corresponding depth maps have kindly been provided by the Interactive Visual Media Group at Microsoft Research.[9] The resolution is 1024×768 pixel and the frame rate is 15 frames per second. The length of each sequence is 100 frames. The first frame of the "ballet" sequence is shown in Fig. 3(a), the first frame of the "breakdancers" sequence in Fig. 4(a). The corresponding depth maps are shown in Fig. 3(b) and 4(b), respectively.



(a) Image (b) Depth map

Figure 1. First frame of the “Santa” sequence



(a) Image (b) Depth map

Figure 2. First frame of the “milk” sequence

As can be seen from Fig. 3(b) and Fig. 4(b), the depth maps for these test datasets are more precise than those for the “Santa” and “milk” sequences and give a much clearer impression of the scene. They can therefore serve as an example for depth maps with a higher level of detail.

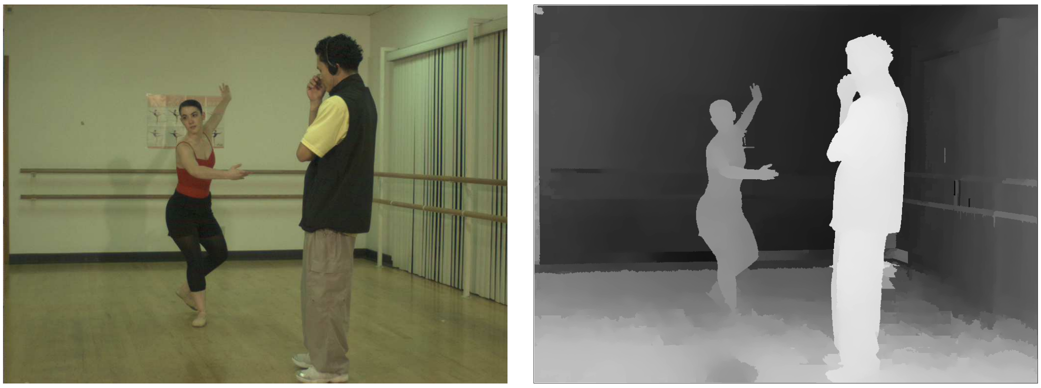
3. CODING TESTS USING H.264/AVC

For the coding tests described in this section, version JM 9.3 of the H.264/AVC reference software has been used. Only the first frame of the depth sequence was coded as an I-frame, the remaining frames as P-frames. The search range was set to 16 pixels and the number of reference frames to 5. For comparison, the recorded image sequence was also coded in addition to the corresponding depth maps with the same parameter set.

3.1. Simulation Results for the Hand-Held Camera Sequences

Figure 5 shows the resulting rate-distortion curves for the “Santa” and “milk” sequence. For both datasets, the uncoded image sequence was stored in YUV 4:2:0 format with a bit rate of 47 185.92 kbit/s, the uncoded sequence of depth maps in YUV 4:0:0 format with a bit rate of 31 457.28 kbit/s.

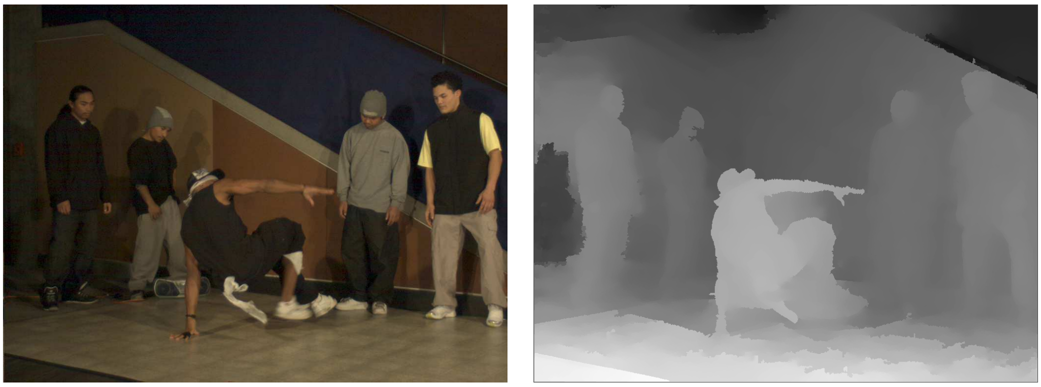
From the curves, it is obvious that the depth maps can be compressed very well. For the same peak-signal-to-noise-ratio (PSNR), the remaining bit rates after compression are much smaller for the depth maps than for the



(a) Image

(b) Depth map

Figure 3. First frame of the “ballet” sequence



(a) Image

(b) Depth map

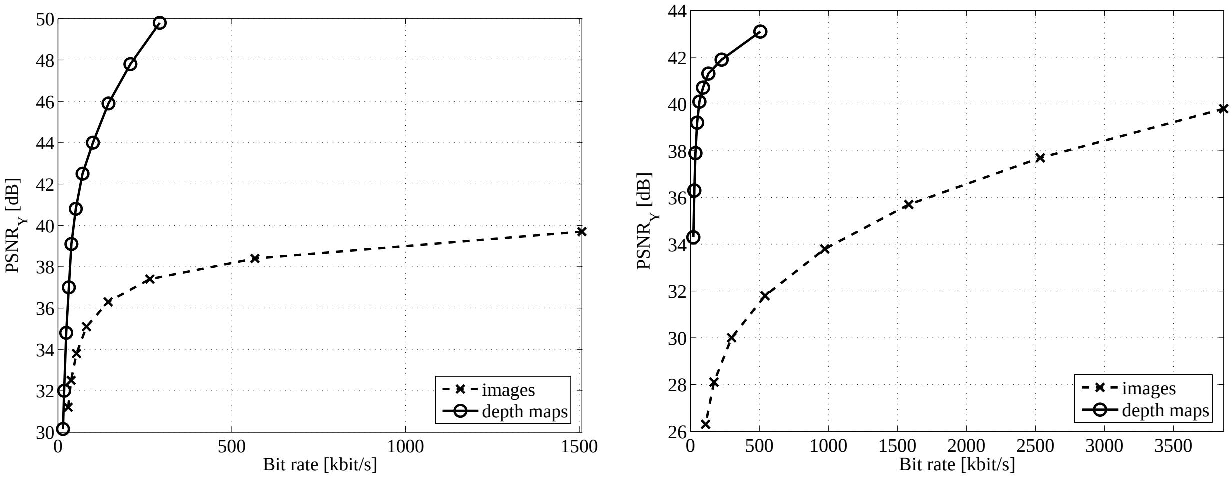
Figure 4. First frame of the “breakdancers” sequence

image sequences, and the achievable compression factors are much higher. If, e. g., the “Santa” sequence shall be transmitted with a PSNR of at least 38 dB, the image sequence would require a coded bit rate of 566.8 kbit/s (compression 1 : 83.3), the depth map sequence a coded bit rate of 38.71 kbit/s (compression 1 : 812.6). This can be explained by the fact that the depth maps used mainly consist of smooth areas with little high-frequency content.

3.2. Simulation Results for the Multi-View Sequences

In Fig. 6, the coding results for the multi-view test datasets “ballet” and “breakdancers” are shown. The uncoded bit rates for a single view of the multi-view dateset are 141 557.76 kbit/s for the image sequence (YUV 4:2:0 format), and 94 371.84 kbit/s for the sequence of depth maps (YUV 4:0:0 format).

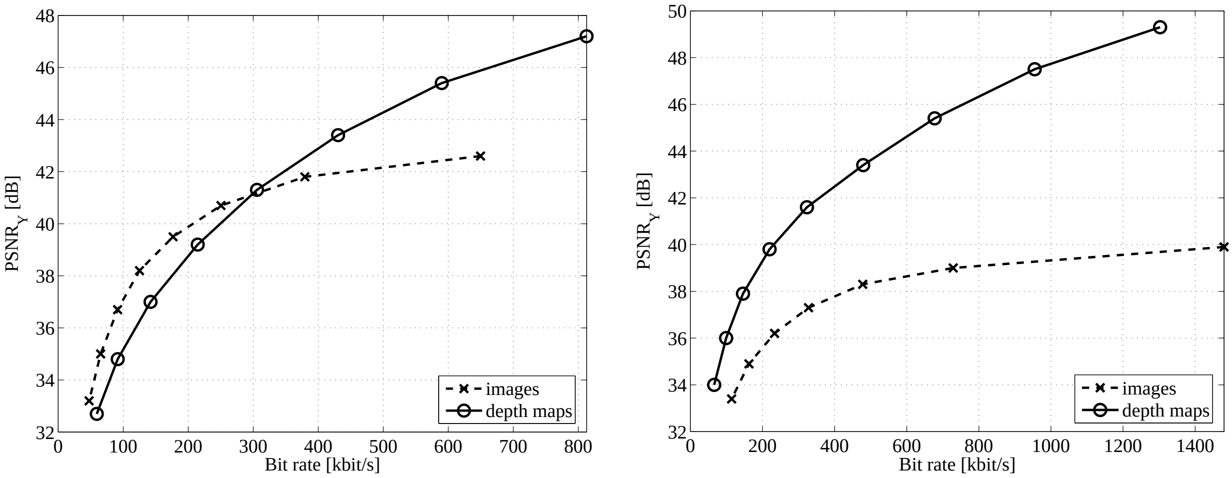
As the depth maps used here are rather detailed, they cannot be compressed as good as the depth maps of the “Santa” and “milk” sequence. However, the compression still performs very well. For a transmission of the “breakdancers” sequence with a PSNR of about 38 dB, the bit rate for the image sequence would be 478.0 kbit/s (compression 1 : 296.1), the bit rate for the depth map sequence would be 145.8 kbit/s (compression 1 : 647.3). The achieved compression factors are still about twice as high than the compression factors achieved for the corresponding image sequences.



(a) “Santa” sequence

(b) “milk” sequence

Figure 5. Rate-distortion curves for the hand-held camera sequences and their associated depth maps



(a) “ballet” sequence

(b) “breakdancers” sequence

Figure 6. Rate-distortion curves for the multi-view sequences and their associated depth maps. The results are shown for the first camera view of the dataset.

For the “ballet” sequence, depth map compression performs slightly worse than image compression in the case of low bit rates. For high bit rates, in contrast, it still outperforms image compression.

When depth maps corresponding to multi-view sequences are coded, the coding efficiency is not necessarily homogeneous between the different camera views. Figure 7 compares the coding efficiencies obtained when compressing the depth maps for the different viewpoints of the “ballet” sequence. It can be seen that the performance differs quite significantly between the different views.

4. EFFECT OF DEPTH MAP COMPRESSION ON THE RENDERING QUALITY

In a practical application, depth maps are usually not shown to the viewer but used as additional information during the rendering process. Therefore, distortions which are introduced in the depth maps themselves by lossy

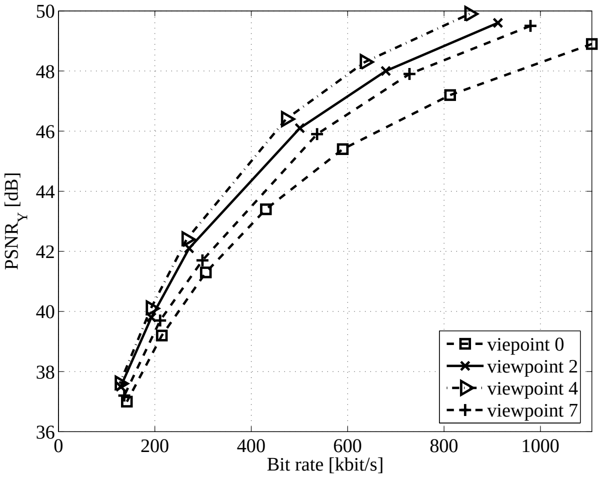


Figure 7. Coding efficiency for the depth maps corresponding to different viewpoints of the "ballet" sequence. Some views have been omitted for the purpose of clarity.

coding are only of limited interest for a real rendering system. Distortions in the final, rendered views of the object or scene as seen by the viewer are of greater impact. That is why in this section, we analyze how the quality of rendered images is affected by the distortions in the depth maps introduced by compressing them. For that, new views are generated from the original sequence together with either the uncoded or coded version of the associated depth maps. This is done for the case of unstructured lumigraph rendering and sequences recorded by a hend-held camera. The results are again exemplarily shown for the "Santa" and "milk" sequences.

4.1. Evaluation Method

The rendering quality is measured in terms of PSNR of the rendered view compared to the original image recorded at the same viewing position and with the same viewing angle. That is why a PSNR calculation is only possible when an image is rendered for one of the original camera positions. If the whole recorded sequence was used for rendering, an optimal light field renderer would use all pixels from the originally recorded image at the chosen position without applying any depth information and deliver an exact copy, which would make the PSNR calculation useless. To overcome this issue, the quality estimation is done as follows:

1. Choose one of the originally recorded images.
2. Leave out the chosen image and do not use it for rendering. Optionally, images preceding and succeeding the chosen image may also be left out, which leads to a greater importance of the depth maps for the rendering process.
3. From the remaining images, render a new view for the viewpoint and viewing direction of the chosen image.
4. Calculate the PSNR between the rendered image and the chosen original.
5. Repeat steps 1 to 4 for all originally recorded images in the sequence and calculate the average PSNR.

More information on this process can be found in Ref. 8.

4.2. Results

The results are shown in Table 1 for the "Santa" sequence and in Table 2 for the "milk" sequence. From the results, it can be seen that the rendering quality marginally deteriorates when the depth maps are compressed. Even for very high compression factors, the influence on the rendering quality can be neglected.

Table 1. Rendering qualities (PSNR values) for the "Santa" sequence using uncompressed and compressed depth maps with different quality levels. QP denotes the quantization parameter used for coding the depth maps, C denotes the achieved compression factor.

Number of images left out	**No compression** ($C = 1$)	**QP = 22** ($C = 107.4$)	**QP = 31** ($C = 312.2$)	**QP = 43** ($C = 1006.7$)	**QP = 51** ($C = 2164.1$)
1	30.31 dB	30.30 dB	30.30 dB	30.26 dB	30.19 dB
3	26.10 dB	26.10 dB	26.09 dB	26.03 dB	25.96 dB
5	23.82 dB	23.81 dB	23.80 dB	23.76 dB	23.69 dB
7	22.51 dB	22.51 dB	22.51 dB	22.49 dB	22.39 dB
9	21.05 dB	21.04 dB	21.04 dB	21.00 dB	20.91 dB

Table 2. Rendering qualities (PSNR values) for the "milk" sequence using uncompressed and compressed depth maps with different quality levels

Number of images left out	**No compression** ($C = 1$)	**QP = 22** ($C = 62.1$)	**QP = 31** ($C = 345.8$)	**QP = 40** ($C = 854.0$)	**QP = 46** ($C = 1462.0$)
1	24.49 dB	24.50 dB	24.50 dB	24.46 dB	24.33 dB
3	21.99 dB	22.01 dB	21.99 dB	21.92 dB	21.71 dB
5	20.33 dB	20.35 dB	20.33 dB	20.25 dB	19.99 dB
7	18.98 dB	19.01 dB	18.97 dB	18.87 dB	18.59 dB
9	17.86 dB	17.88 dB	17.85 dB	17.74 dB	17.47 dB

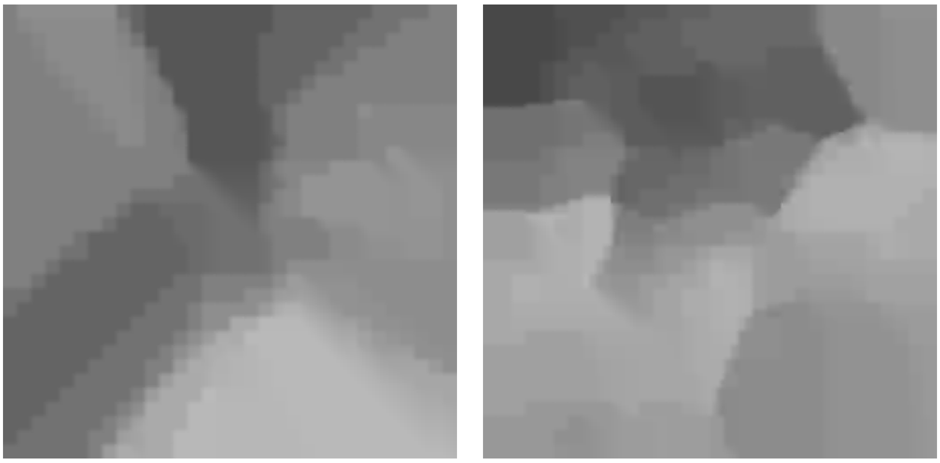
Examples for depth maps which have been compressed with high compression factors are shown in Fig. 8. They show severe blocking artifacts. However, the coding artifacts alter the absolute depth values only to a limited amount, and the original depth maps are already very coarse. That is why the artifacts do not have much impact on the quality of the rendered images.

The results indicate that depth information could be stored or transmitted together with the original luminance and color information using very little additional data rate. As it could be seen in Sect. 3, the data rate however depends on the complexity of the depth maps. When light field data is e. g. transmitted and the receiver needs depth information for the rendering process, it might be more suitable to generate depth maps at the sending side and to transmit them in a compressed form instead of generating them at the receiver side of the system. This would then free the receiver from the time-consuming process of depth estimation and would assure that the depth maps are based on the original, uncompressed light field images.

5. SUMMARY

Compression of depth maps used for image-based rendering was analyzed. For the case of unstructured lumigraph rendering, test sequences recorded using a single, hand-held camera were used. For the case of multi-view video, sequences containing dynamic scenes from eight different camera viewpoints together with their corresponding depth maps were available. H.264/AVC was chosen for the compression of the depth maps, and the coding efficiency was compared to the efficiency when the original image data is coded. High compression factors could be achieved for the depth maps, and in most cases, they were much higher than those for the corresponding images. The achievable coding efficiency however depends on the complexity of the depth maps. In the case of multi-view sequences, it may also depend on the camera view.

For the case of unstructured lumigraph rendering, the effect of depth map compression on the rendering quality was analyzed. For that, images were generated using the original depth maps as well as the compressed, distorted depth maps. Although for low bit rates, the quality of the depth maps is severely distorted, only a marginal deterioration of the rendering quality could be observed, even for very high compression factors. This leads to the conclusion that depth maps could be stored or transmitted together with the original images with



(a) First frame of the “Santa” sequence (QP = 51)

(b) First frame of the “milk” sequence (QP = 46)

Figure 8. Distorted depth maps after compression with low bit rates

limited additional data rate. When light field data is transmitted, it might therefore be suitable to generate the depth information at the sending side and to transmit it in addition to the compressed light field data. By doing so, the receiver could be freed from the complex process of depth estimation.

ACKNOWLEDGMENT

This work was funded by the German Research Foundation (DFG) under grant SFB 603/TP C8 and TP C2. Only the authors are responsible for the content.

REFERENCES

1. M. Levoy and P. Hanrahan, “Light field rendering,” in *Proceedings SIGGRAPH 96*, (New Orleans, Louisiana, USA), Aug. 4–9, 1996.
2. S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, “The lumigraph,” in *Proceedings SIGGRAPH 96*, pp. 43–54, (New Orleans, Louisiana, USA), Aug. 4–9, 1996.
3. U. Fecker and A. Kaup, “H.264/AVC-compatible coding of dynamic light fields using transposed picture ordering,” in *2005 European Signal Processing Conference*, (Antalya, Turkey), Sept. 4–8, 2005.
4. U. Fecker and A. Kaup, “Statistical analysis of multi-reference block matching for dynamic light field coding,” in *10th International Fall Workshop Vision, Modeling and Visualization 2005*, (Erlangen, Germany), Nov. 16–18, 2005.
5. C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, “Unstructured lumigraph rendering,” in *Proceedings SIGGRAPH 2001*, (Los Angeles, CA, USA), Aug. 12–17, 2001.
6. B. Heigl, *Plenoptic Scene Modeling from Uncalibrated Image Sequences*, ibidem-Verlag Stuttgart, January 2004.
7. C. Fehn, K. Schüür, P. Kauff, and A. Smolic, “Coding results for EE4 in MPEG 3DAV,” in *ISO/IEC JTC1/SC29/WG11, Document MPEG2003/M9561*, (Pattaya, Thailand), Mar. 2003.
8. H. Niemann and I. Scholz, “Evaluating the quality of light fields computed from hand-held camera images,” *Pattern Recognition Letters* **26**, pp. 239–249, February 2005. In Memoriam: Azriel Rosenfeld.
9. C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, “High-quality video view interpolation using a layered representation,” in *ACM SIGGRAPH and ACM Transactions on Graphics*, pp. 600–608, (Los Angeles, CA, USA), Aug. 2005.