VIEW INTERPOLATION WITH STRUCTURED DEPTH FROM MULTIVIEW VIDEO

Pravin Kumar Rana and Markus Flierl

ACCESS Linnaeus Center, School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
{prara, mflierl}@kth.se

ABSTRACT

In this paper, we propose a method for interpolating multiview imagery which uses structured depth maps and multiview video plus inter-view connection information to represent a three-dimensional (3D) scene. The structured depth map consists of an inter-view consistent principal depth map and auxiliary depth information. The structured depth maps address the inconsistencies among estimated depth maps which may degrade the quality of rendered virtual views. Generated from multiple depth observations, the structuring of the depth maps is based on tested and adaptively chosen inter-view connections. Further, the use of connection information on the multiview video minimizes distortion due to varying illumination in the interpolated virtual views. Our approach improves the quality of rendered virtual views by up to 4 dB when compared to the reference MPEG view synthesis software for emerging multimedia services like 3D television and free-viewpoint television. Our approach obtains first the structured depth maps and the corresponding connection information. Second, it exploits the inter-view connection information when interpolating virtual views.

1. INTRODUCTION

The wide availability of low-cost digital cameras enables us to acquire multiview imagery from various viewpoints of dynamic natural three-dimensional (3D) scenes by utilizing camera arrays. Three-dimensional television (3D TV) and free-viewpoint television (FTV) are emerging visual media applications that use multiview imagery [1]. 3D TV aims to provide a natural 3D-depth impression of dynamic 3D scenes, while FTV enables viewers to freely choose their viewpoint of real world scenes. In conventional multiview systems, view interpolation is required for smooth transitions among captured views. Usually, view interpolation uses multiple views and depth maps acquired from different viewpoints. Each depth map gives information about the distance between the corresponding camera and the objects in the 3D scene. Depth maps for chosen viewpoints are estimated by establishing stereo correspondences only between nearby views [2]. However, the estimated depth maps of different viewpoints usually demonstrate only a weak inter-view consistency [3]. These inconsistent depth maps affect negatively the quality of view rendering.

To enable these emerging visual media, two approaches are usually considered. The first approach utilizes multiple views with multiple corresponding depth maps. This requires high bandwidth and storage capacities. The second approach utilizes layered depth images [4]. This is a more efficient data representation. However, the layered structure is defined with respect to a single viewpoint. This may be disadvantageous, if novel viewpoints differ significantly from the reference viewpoint. In contrast to this approach, Muller *et al.* [5] keep the layers at their original positions to maximize data coverage and merge them into a single buffer to obtain a optimal data representation. However, this may lead to distortions due to varying illumination among views.

In this paper, we use depth-based rendering techniques [6] and take advantage of structured depth maps, multiview video (MVV), and inter-view connection information from multiple reference viewpoints to render high quality virtual views. The structured depth map comprises an inter-view consistent principal depth map and additional auxiliary depth information. Based on the consistent principal depth, we explain extraction and use of auxiliary depth information to handle depth discontinuities at virtual viewpoints to improve the visual quality of rendered virtual views. In [3], we test inter-view connection evidence of the estimated depth pixels at multiple reference viewpoints and use the resulting consistent depth to render a virtual view. The experiments show that this testing offers a visually improved rendered view for a given virtual viewpoint. In [3], we also recalculated the inter-view connection evidence for each new virtual view.

In contrast to our previous work in [3], we generate only once the inter-view connection information with respect to the principal viewpoint in order to render multiple virtual views by using structured depth maps and multiview video. The proposed method obtains first the inter-view consistent depth map and extracts the inter-view connection information at the principal viewpoint. Second, the method utilizes the resulting inter-view connectivity information and the structured depth map to render multiple virtual views. Further, this inter-view connectivity information facilitates smooth transitions among multiview imagery with improved visual quality and minimizes the impact of changing illumination on the rendered views.

The remainder of the paper is organized as follows: In Section 2, we summarize the depth consistency testing algorithm (DCTA) to obtain the inter-view consistent principal depth map. In Section 3, we propose structured depth-image-based rendering with the inter-view connection information and its advantage over layered depth approaches. Section 4 presents our experimental results on virtual view rendering. Finally, Section 5 gives concluding remarks.

2. DEPTH CONSISTENCY TESTING

As conventionally estimated depth maps usually show weak inter-view consistency, we proposed a method in [3] to achieve inter-view depth consistency at a given viewpoint. As summarized in Fig. 1, the algorithm warps more than two depth maps from multiple reference viewpoints to the principal viewpoint p. At this stage, the reference depth maps are used for 3D warping. Small holes that occur during warping are filled by using a 3x3 median filter.

In the next stage, the consistency among all warped depth values at the principal viewpoint is examined. To assess consistency, the absolute differences between all possible pairs of depth values for each given principal pixel are determined. For example, with n reference views, there are up to $N = \frac{n!}{(n-2)!2!}$ possible pairs of depth values for a given pixel. This can be represented by the following symmetric matrix

$$ADM = \begin{pmatrix} 0 & \Delta_{1,2} & \dots & \Delta_{1,n} \\ \Delta_{1,2} & 0 & \dots & \Delta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{1,n} & \Delta_{2,n} & \dots & 0 \end{pmatrix}, \qquad (1)$$

where ADM is the absolute difference matrix of all possible pairs of the depth values at the principal pixel for a given frame f, and $\Delta_{j,k} = |d_j - d_k|$ is the absolute difference of depth values between

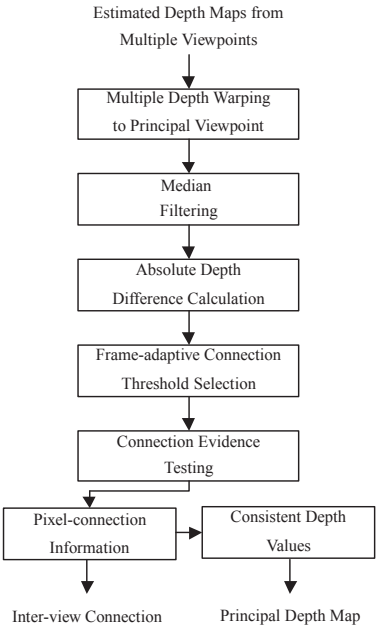


Figure 1: Block diagram for extracting the inter-view connection and the principal depth map.

warped depth map j and warped depth map k at the principal pixel. Indexes $j,k = \{1,\ldots,n\}$ represent the warped views from different viewpoints, where $j < k$. Since the ADM is symmetric with diagonal elements being zero, its upper triangular part is sufficient for testing. Each $\Delta_{j,k}$ is a *inter-view connection evidence*, which is a measure of depth consistency between the corresponding depth pairs (d_j, d_k) at the principal pixel.

2.1 Inter-View Connection Evidence Testing

Now, each inter-view connection evidence is tested by checking the corresponding value of $\Delta_{j,k}$ in the ADM for a given principal pixel. As a result of this testing, we get the inter-view connection information across multiple reference view and the inter-view consistent principal depth map. If a inter-view connection evidence is less than a given connection threshold, the evidence is accepted and it is assumed that the corresponding two depth values in the two warped depth maps are consistent for the given principal pixel. Hence, these consistent depth pixels have a consistent depth representation of the corresponding 3D object point in world coordinates. Otherwise, the connection evidence is rejected and it is assumed that the corresponding two depth pixels in the two warped depth maps do not have a consistent depth representation. The connection threshold T_f relates to the quality of the connectivity and defines a criterion for depth consistency testing for each frame f according to

$$T_f = \mu_f + \lambda\sigma_f, \tag{2}$$

where μ_f and σ_f are the mean and the standard deviation of all $\Delta_{j,k}$ in the ADM per frame, respectively. The appropriate value of the parameter $\lambda \in [0,1]$ is manually chosen for each test sequence. In the experiments, $\lambda = 0.8$ is chosen to maximize the objective quality for most of the test sequences.

Based on the inter-view connection evidence in the ADM per pixel, various cases of inter-view connectivity can arise, as depicted in Fig. 2. The different cases of inter-view connectivity and the corresponding pixel selection from multiple reference viewpoints are tested.

For any accepted connection, the resulting inter-view connectivity information is used to choose consistent depth values for principal pixels. Further, this resulting inter-view connectivity information with respect to the principal view is used for rendering virtual views as described in the Section 3.

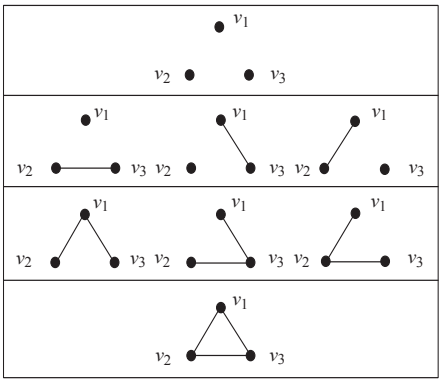


Figure 2: Possible cases of inter-view connectivity for $n = 3$.

To determine the final depth value at a principal pixel, we average the consistent depth values. However, if the reference views are captured by using irregular camera-baseline distances between viewpoints, we estimate the final depth value at the principal pixel by weighted-baseline averaging of the consistent depth values. This exploitation of inter-view connectivity also enhances the depth map of the chosen principal viewpoint.

3. STRUCTURED-DEPTH-IMAGE-BASED RENDERING

Fig. 3 shows a block diagram of the structured-depth-image-based rendering technique which utilizes the inter-view connection information as obtained by the DCTA and multiview video. An inter-view consistent principal depth map and auxiliary depth information for various required viewpoints constitutes the structured depth map. When the principal depth map is warped, areas affected by occlusion result in depth discontinuities. As depth information is not available at the principal viewpoint, we fill occluded areas with depth information from the reference depth maps at the same viewpoint. The information needed to fill these affected areas is usually very small and we call it *auxiliary depth information*. In the following, we describe structured depth maps and auxiliary depth information in more detail. Furthermore, we explain structured-depth-image-based rendering with inter-view connection information.

3.1 Structured Depth Maps

The structured depth map d^* can be denoted by

$$d^* = \{d_p, d'\}, \tag{3}$$

where d_p is an inter-view consistent principal depth map at the principal viewpoint p as obtained by the DCTA in [3] and $d' = \{\ldots, d'_{p-2}, d'_{p-1}, d'_{p+1}, d'_{p+2}, \ldots\}$ is a set of auxiliary depth values at given reference viewpoints. The cardinality $|d'|$ of the set d' depends on the number of reference views used for depth consistency testing to obtain the principal depth map d_p. Let d_r , $r \in \{1,\ldots,n\}$, represent an arbitrary reference depth map from viewpoint r which is used in the depth consistency testing to obtain a principal depth map at viewpoint p. The cardinality of the set d' is defined by

$$|d'| = \begin{cases} (n-1) & \text{if } p = r,\ \forall\, r, \\ n & \text{if } p \neq r,\ \forall\, r, \end{cases} \tag{4}$$

where n is the number of reference views used in the depth consistency testing.

3.2 Extraction of Auxiliary Depth Information

To obtain the auxiliary depth information d'_i at the viewpoint i, we first warp the consistent principal depth map d_p to i. By detecting discontinuities in the warped depth map at i, we identify areas

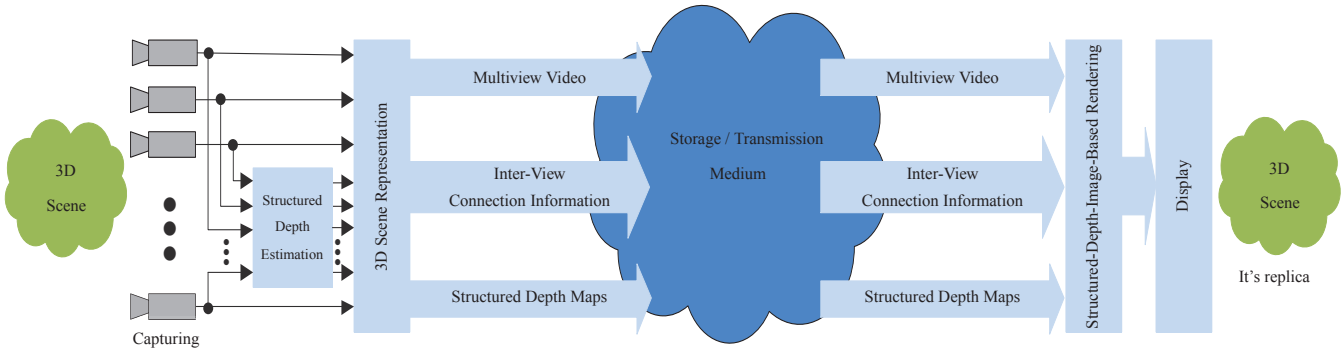


Figure 3: Structured-depth-image-based rendering.

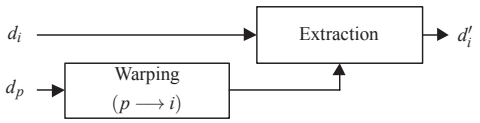


Figure 4: Extraction of auxiliary depth information.

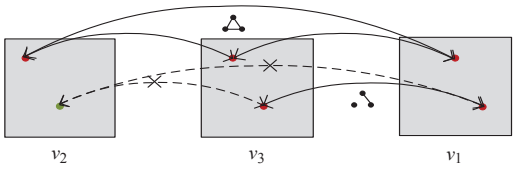


Figure 6: Example for inter-view connection information with three reference views.

which are affected by occlusion with respect to the principal view p and generate an occlusion mask. Using this mask, we extract depth information from the reference depth map d_i that corresponds to occluded areas. With that, we have the auxiliary depth information at i with respect to the principal view p, as shown in Fig. 4.

3.3 Rendering with Inter-View Connection Information

To obtain enhanced depth maps at other viewpoints of a 3D scene, we warp the consistent principal depth map to other viewpoints. The occluded areas in warped depth maps are filled by using auxiliary depth information. The resulting depth maps at multiple viewpoints are inter-view consistent with respect to the principal viewpoint, as shown in Fig 5.

For rendering, we choose a virtual viewpoint v^* which is located between the principal viewpoint and an auxiliary viewpoint. First we warp the principal depth map from p to all required auxiliary viewpoints. By checking discontinuities in the warped principal depth map at the auxiliary viewpoints, we identify areas which are visible due to warping from viewpoint p to auxiliary viewpoints. This location information about the holes and the information in the auxiliary depth information is identical because both are obtained by warping of the same principal depth map. Hence, the newly exposed areas in the principal depth map at a particular auxiliary viewpoint are filled by the corresponding auxiliary viewpoint information obtain from the set d'. Note that this location information does not need to be transmitted or stored additionally. The resulting depth maps are augmented by auxiliary depth information at auxiliary viewpoints, which are inter-view consistent with respect to the principal viewpoint.

Now, by checking discontinuities in the warped principal depth map at v^*, we identify areas which are affected by occlusion with respect to p and generate a principal occlusion mask. The inter-view connectivity information is not available for the occluded areas at v^* because the connectivity information is obtained with respect to p. Hence, we mask the inter-view connection information for the occluded areas by using the principal occlusion mask for rendering the view at v^*.

For each given virtual pixel, the masked inter-view connection information provides information about the connected reference views, as shown in Fig 6. Therefore, we warp texture pixels from multiple reference viewpoints to the viewpoint v^* according to the masked inter-view connection information for viewpoint v^* by using the resulting $|d'|$ depth maps at auxiliary viewpoints and the principal depth map. If there is no inter-view connection information available for a virtual pixel, the proposed rendering is not able to determine a virtual view pixel-intensity from the reference views. In this case, we set a mask for inpainting to determine the pixel-intensity values for such unconnected inter-view pixels at the virtual view. If there is an inter-view connection information available for the virtual pixel, we use the connectivity information to warp the specified pixels in the reference views to the virtual viewpoint.

To determine the final pixel-intensity in the virtual-view, we used various approaches depending upon the baseline scenario and the varying illumination condition among reference views. If the pixel intensities of inter-view connected reference pixels are similar, averaging of the warped pixel intensities is feasible. However, if the pixel intensities among the connected and warped texture views vary significantly due to varying illumination, we assume that the virtual pixel value is best described by the warped texture pixel of the nearest reference view. The reference view which has minimum baseline-distance from the virtual viewpoint is defined as the nearest view. In this case, we simply set the pixel intensity in the virtual view by copying the pixel-intensity information from the warped texture pixel of the nearest reference view that is connected. If the reference views are captured from multiple viewpoints using irregular camera baseline distances between viewpoints, we estimate the virtual pixel intensity by weighted-baseline averaging of the connected and warped pixel intensities. Further, to determine the virtual pixel-intensity, the advantage of color consistency testing could be exploited too.

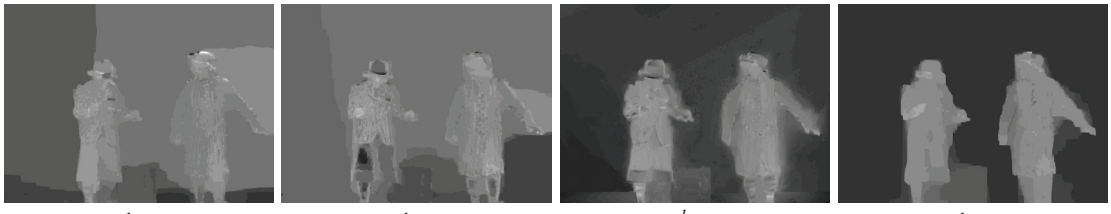
3.3.1 Hole Filling and Inpainting

Increasing the number of reference views is likely to decrease hole areas. However, holes cannot be ruled out completely. Therefore, holes are detected by checking the unconnected inter-view pixels at the virtual view in the inter-view connectivity information. If some holes remain due to unconnected inter-view pixels, they are filled by inpainting [7] using an inpainting mask which is defined for the unconnected inter-view pixels.

3.4 Discussion

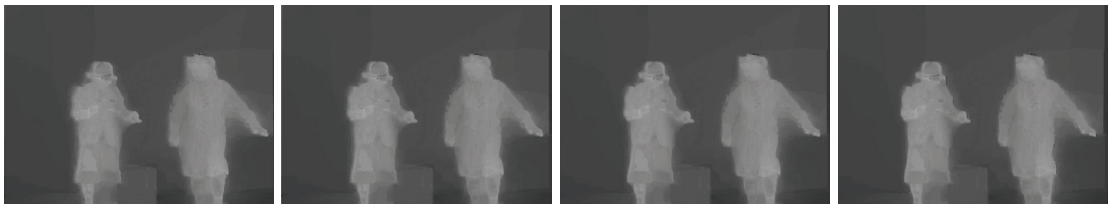
3.4.1 Difference to Layered Depth Approaches

Several approaches have been proposed for the representation of natural 3D scenes and a well-known representation is the layered



d_{p-4} d_{p-2} d_{p+2} d_{p+4}

(a) Estimated depth maps as obtained by MPEG depth estimation reference software 5.0.



$\hat{d}_{p-4}$ $\hat{d}_{p-2}$ $\hat{d}_{p+2}$ $\hat{d}_{p+4}$

(b) Inter-view consistent depth maps as obtained by using a structured depth map.

Figure 5: Comparison of estimated reference depth maps and inter-view consistent depth maps for Pantomime.

depth image (LDI) [4]. LDIs enhance the basic depth image by storing multiple pairs of associated color and depth values for each pixel of the original image. However, the number of layers depends on the scene complexity and the required synthesis quality. LDIs are further explored by Zitnick *et al.* [8], Cheng *et al.* [9], and Muller *et al.* [5] for multiview imagery. For LDI with multiview input imagery, each pixel has to undergo two warping processes for final rendering, which can degrade the quality of rendered views [4]. As investigated in [5], there are two main approaches. In the first approach, views are warped towards the principal viewpoint to obtain a single buffer that leads to a considerable loss of information due to warping, similar to [4]. In the second approach, merging auxiliary views without warping in a single buffers may lead to distortions due to different illumination among the views, which are also visible in the rendered views. We reduce these effects on the rendering by using the original reference views and by keeping the auxiliary depth information at the original viewpoints without merging it into a single viewpoint.

However, the inconsistency among the depth maps affects the visual quality of the rendered views negatively. Therefore, instead of using an individually estimated depth map for rendering, we propose the structured-depth-image-based rendering with inter-view connectivity information. More than two estimated depth maps from multiple viewpoints of a given 3D scene are used to obtain the consistent principal depth map by depth consistency testing [3]. However, Zitnick *et al.* [8] use only a constraint to ensure consistency across layers which is based on a disparity space distribution for each segment under the assumption that all pixels within a segment have the same disparity. In contrast to this, our consistency testing is a statistical approach which is based on a defined value for depth of a 3D point and inter-view connectivity information. The resulting principal depth map and the corresponding resulting structured depth maps provide a inter-view consistent depth representation of the 3D scene across multiple viewpoints, as shown in Fig. 5. The consistently structured depth maps allow for smooth navigation among rendered views at multiple viewpoints with improved visual quality.

4. RENDERING EXPERIMENTS AND RESULTS

To evaluate the proposed structured depth-image-based rendering which exploits inter-view pixel-connectivity, we assess the quality of rendered virtual views. We measure the objective video quality of the rendered view at a given viewpoint by means of the Peak Signal-to-Noise Ratio (PSNR) with respect to the captured view of a real camera at the same viewpoint. We use the standard Motion Picture Expert Group (MPEG) multiview video test sequences, Pantomime, Dog, Newspaper, Lovebird1, and Mobile [10]. We estimate the required depth maps by using the MPEG 3DV/FTV Depth Estimation Reference Software (DERS) 5.0 [11, 12].

In the experiments, we render views at two virtual viewpoints. For example, we use first the estimated reference depth maps at odd reference viewpoints to obtain the principal depth map and the inter-view connectivity information at the viewpoint p by depth consistency testing. Note, the principal view is chosen among the odd reference viewpoints. Second, depth maps at the auxiliary viewpoints are obtained by using the extracted set d' of auxiliary depth information and the principal depth map. Usually, the percentage of auxiliary depth information with respect to the principal depth map depends on many factors, i.e., the number of reference depth maps used to obtain the principal depth map, the baseline distance between the reference depth maps, and the accuracy of the reference depth estimation algorithm. For example, for the three reference view scenario, the average percentage of auxiliary depth information for Dog (averaged over 50 frames) is about 0.6%. However, the average percentage of auxiliary depth information for Lovebird1 (averaged over 50 frames) is about 1.0%. Further, we use the principal depth map and the resulting depth maps at auxiliary viewpoints to warp original texture views according to the masked inter-view connectivity information to render virtual views at two even virtual viewpoints, as described in Section 3.

Table 1: Quality of rendered virtual views.

MPEG Test Data	Virtual View	Proposed Method [dB] (a)	VSRS 3.5 [dB] (b)	△Y-PSNR [dB] (a-b)
Pantomime	38	38.03	36.81	1.22
($p = 39$)	40	39.96	36.62	3.34
Dog	38	35.20	34.05	1.15
($p = 39$)	40	33.10	28.24	4.86
Lovebird1	05	32.80	31.94	0.86
($p = 04$)	07	30.50	30.32	0.18
Newspaper	02	31.29	30.73	0.56
($p = 03$)	04	30.72	29.52	1.20
Mobile	04	42.60	41.36	1.24
($p = 05$)	06	41.97	39.94	2.03

The proposed algorithm is compared to the synthesized virtual view as obtained by the MPEG 3DV/FTV View Synthesis Reference Software 3.5 (VSRS 3.5) [12, 13]. The synthesis reference software is based on the depth-image-based rendering technique provided by Nagoya University for MPEG 3DV/FTV exploration experiments. It uses two reference views, left and right, to synthesize a virtual view by using the two corresponding reference depth

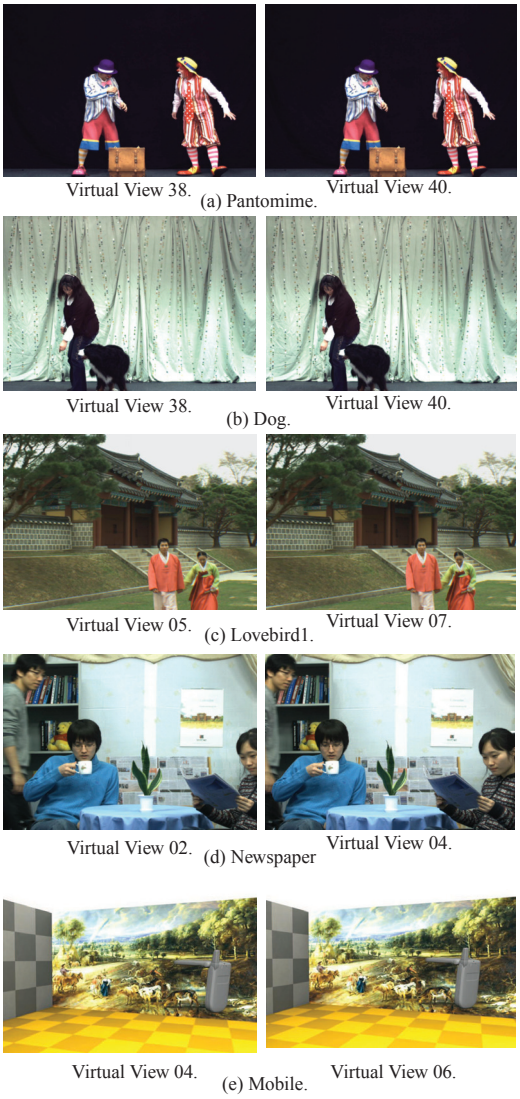


Figure 7: Rendered views at two virtual viewpoints.

maps. The reference software mainly employs pixel-by-pixel mapping for depth maps and texture views, hole filling, view blending, and inpainting for the remaining holes. We synthesize virtual views by using the general synthesis mode with quarter-pel precision.

Table 1 shows a comparison of the average (over 50 frames) luminance signal Y-PSNR (in dB) of the rendered virtual views using (a) the proposed structured view rendering and (b) MPEG 3D/FTV VSRS 3.5, respectively. The presented structured view rendering is based on inter-view connection information and offers improvements of up to 4 dB. The improvement of the quality depends on the input reference depth maps from various viewpoints which are used to obtain the consistent principal depth map. Note that here view interpolation is achieved with only one set of consistently structured depth maps and the corresponding inter-view connection information. This is in contrast to the work in [3] where depth consistency testing has been performed for each interpolated pixel. This demonstrates the efficiency of our consistently structured depth images to represent 3D scenes. Fig. 7 shows rendered views at two viewpoints for subjective evaluation.

5. CONCLUSIONS

In this paper, we proposed structured depth-image-based rendering which exploits the inter-view connectivity information among multiview video and takes advantage of a consistent principal depth map as obtained by depth consistency testing. By using the structured depth map and the connectivity information, we addresses the problems of inter-view depth inconsistencies and varying illumination conditions. Consistently structured depth maps permit an appealing 3D scene representation on the encoder side by avoiding depth consistency testing for each interpolated pixel on the decoder side. Structured depth improves the subjective visual quality as well as the objective quality of rendered views by up to 4 dB when compared to the views rendered by VSRS 3.5.

ACKNOWLEDGMENT

This work was supported by Ericsson AB and the ACCESS Linnaeus Center at KTH Royal Institute of Technology, Stockholm, Sweden.

REFERENCES

[1] M. Tanimoto, "Overview of free viewpoint television," in *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, Jul. 2006.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *International Journal of Computer Vision*, vol. 47, pp. 7–42, Apr. 2002.

[3] P. K. Rana, and M. Flierl, "Depth consistency testing for improved view interpolation," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing*, Saint Malo, France, pp. 384–389, Oct. 2010.

[4] J. Shade, S. Gortler, L. He, and R. Szeliski, " Layered depth images," in *Proc. of the ACM SIGGRAPH*, Orlando, Florida, USA, pp. 231–241, Sept. 1998.

[5] K. Muller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing*, Cairns, Australia, pp. 34–39, Oct. 2008.

[6] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D TV," in *Proc. of the SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, pp. 93-104, Jan. 2004.

[7] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. of the IEEE Conference on CVPR*, Kauai, HI, USA, vol. 1, issue 1063-6919, pp. 355–362, Dec. 2001.

[8] C.L. Zitnick, S.B. Kang, M.Uyttendaele, S. Winder, and R. Szeliski, "High quality video view interpolation using a layered representation," in *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 600-608, 2004.

[9] X. Cheng, L. Sun, and S. Yang, "Generation of layered depth images from multi-view video," in *Proc. of the IEEE International Conference on Image Processing*, vol. 5, pp. V-225–228, Oct. 2007.

[10] "Draft call for proposals on 3d video coding technology," MPEG ISO/IEC JTC1/SC29/WG11, N11679, Guangzhou, China, Oct. 2010.

[11] M. Tanimoto, T. Fujii, M. Panahpour, and M. Wildeboer, "Depth estimation reference software DERS 5.0," ISO/IEC JTC1/SC29/WG11, M16923, Xian, China, Oct. 2009.

[12] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," ISO/IEC JTC1/SC29/WG11, M15377, Archamps, France, Apr. 2008.

[13] "View synthesis software manual," MPEG ISO/IEC JTC1/SC29/WG11, Sept. 2009, release 3.5.