Suryanarayana M. Muddala; Mårten Sjöström; Roger Olsson; and Sylvain Tourancheau. "Edge-aided virtual view rendering for multiview video plus depth", Proc. SPIE 8650, Three-Dimensional Image Processing (3DIP) and Applications 2013, 86500E (March 12, 2013);

http://dx.doi.org/10.1117/12.2004116

Edge-aided virtual view rendering for multiview video plus depth

Suryanarayana M. Muddala, Mårten Sjöström, Roger Olsson and Sylvain Tourancheau

Department of Information Technology and Media, Mid Sweden University
Holmgatan10, 85170 Sundsvall, Sweden

ABSTRACT

Depth-Image-Based Rendering (DIBR) of virtual views is a fundamental method in three dimensional 3-D video applications to produce different perspectives from texture and depth information, in particular the multi-view-plus-depth (MVD) format. Artifacts are still present in virtual views as a consequence of imperfect rendering using existing DIBR methods. In this paper, we propose an alternative DIBR method for MVD. In the proposed method we introduce an edge pixel and interpolate pixel values in the virtual view using the actual projected coordinates from two adjacent views, by which cracks and disocclusions are automatically filled. In particular, we propose a method to merge pixel information from two adjacent views in the virtual view before the interpolation; we apply a weighted averaging of projected pixels within the range of one pixel in the virtual view. We compared virtual view images rendered by the proposed method to the corresponding view images rendered by state-of-the-art methods. Objective metrics demonstrated an advantage of the proposed method for most investigated media contents. Subjective test results showed preference to different methods depending on media content, and the test could not demonstrate a significant difference between the proposed method and state-of-the-art methods.

Keywords: View rendering, 3DTV, multiview plus depth (MVD), depth-image-based-rendering (DIBR), warping, image quality assessment.

1. INTRODUCTION

The attention towards future TV and entertainment using 3D technology has been increasing rapidly. Today, glasses are required to separate left and right views to create perception of depth. Multiview displays, on the other hand, provide stereoscopic experience without the need to wear glasses as the display itself distributes several perspective views into the viewing zone. Depth-image-based rendering (DIBR) is an efficient method to render virtual views from an image and its corresponding depth per pixel information. Despite several improvements to general DIBR method, there are still artifacts in the resulting virtual views, which raise the question whether there are alternative DIBR formulations that result in competitive quality.

An efficient and common way to distribute the 3D contents to the end user is to use video-plus-depth (V+D) and multiview video-plus-depth (MVD) formats, which enable rendering of virtual views for stereo and multiview displays. MVD contains two or more number of V+D that increase angular information such that artifacts are reduced in the rendered virtual views: disoccluded areas of one view are mostly filled with information from the other view. Even so, rendering artifacts appear in virtual views, such as cracks (due to rounding the projected pixel coordinates), disocclusions (large missing areas in the virtual view), corona-like effects around foreground objects (due to erroneous depth information), empty regions (due to errors in depth map) and unnatural contours (due to pixilation).

DIBR relies on perspective 3D warping.[1] Different kinds of DIBR methods have been presented to avoid its inherent artifacts. Zitnick et al. divided the original data into layers to separate data depending on the reliability of the information and then the view rendered by 3D mesh structure.[2] Layering approach has been further improved by constraints on layers selection.[3,4] Several improvements considering sub-pixel accuracy, boundary aware processing, noise removal and inpainting have led to the view synthesis reference software

Further information:
Mårten Sjöström: E-mail: marten.sjostrom@miun.se, Telephone: +(46) 060 14-8835, Fax: +(46) 060 14-8830

Three-Dimensional Image Processing (3DIP) and Applications 2013, edited by Atilla M. Baskurt, Robert Sitnik, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8650, 86500E
 · doi: 10.1117/12.2004116

(VSRS).[5] Additional processing steps such as creating the reliability map and the similarity enhancement were implemented in VSRS 1D fast mode.[6] Even though these methods imply significant quantitative improvements, there is still room for further enhancement with simple and direct approaches.

In this paper, we propose an alternative DIBR view synthesis method for MVD data, based on forward warping and interpolation of the actual pixel values from projected pixel positions. So called *edge pixels* are introduced in order to manage discontinuities between objects in the scene in a similar way as in,[7] which addressed V+D format. The novelty of this paper lies in the way information from two adjacent views is merged into the virtual view before the interpolation process. The weighting is done with respect to the distance to the contributing adjacent views.

The outline of the paper is as follows. Section 2 describes the problem description and the proposed method is presented in Section 3. The test arrangement and evaluation criteria are described in Section 4. The results and analysis are given in Section 5 and finally, Section 6 concludes the paper.

2. PROBLEM DESCRIPTION

Depth image based rendering is a method for rendering virtual views at different perspectives using original view plus depth per pixel information and camera parameters. It is based on perspective projection; the mathematical equation for virtual view coordinate from the 3D warping is described by

$$z_v \mathbf{m}_v = z_o \mathbf{P}_v \mathbf{P}_o^{-1}(\mathbf{m}_o), \tag{1}$$

where $\mathbf{P} = \mathbf{KI}[\mathbf{R}|\mathbf{t}]$; $\mathbf{P}$ is the perspective projection matrix, $\mathbf{K}$ is the intrinsic parameter matrix, which describes the focal distance, image centre and camera pixels sizes, $\mathbf{R}$ is the rotation matrix, $\mathbf{t}$ is the translation vector, $\boldsymbol{m} = (u/z, v/z, 1)^T$ is the camera pixel coordinate and z corresponds to its depth information. The subscripts v and o in the Eq. (1) denote virtual and original camera views, respectively.

The basic 3D warping method produces a number of artifacts in the virtual view. Different methods have been described in the literature to handle each of these artifacts. The purpose of this study is to reduce DIBR inherent artifacts in a simple and straightforward approach rather than looking into each artifact individually.

This study is restricted to the rendering of horizontally displaced virtual views with respect to the original views using the MVD format, because the common scenario of rendering a virtual view from two adjacent views in MPEG 3DV EEs is the 1D parallel camera setup. In this setup, the assumption is that the optical axes of the cameras are in parallel. This setup further avoids vertical disparity that creates keystone distortions, a cause for visual discomfort.

The objective of the study is to propose and evaluate a method for rendering virtual views that handles artifacts without specific processing. This work intends to compute a virtual view using principles of 3D warping actual information and to investigate the objective and subjective quality of the produced view by comparing the obtained results and those from the state-of-the-art methods.

3. PROPOSED METHOD

We propose a method based on 3D warping using actual projected pixel information to solve the stated problem. Firstly, we project the original views pixels into the virtual view while retaining the projected pixel information (floating point position) without considering the nearest integer position nor sub-pixel accuracy as in other methods. Next, the method combines the projected view pixels from two adjacent views followed by linear interpolation to get pixel values at the target pixel grid of the virtual view. Empty regions occur in virtual views when rendering from one-view-plus-depth. These empty regions are mostly filled by the data projected from the adjacent view when using MVD format. Nonetheless, there may still exist empty regions after this filling. These artifacts are removed by interpolation using the information of specially introduced edge pixels. Finally, we apply the edge smoothing using a low pass filter.

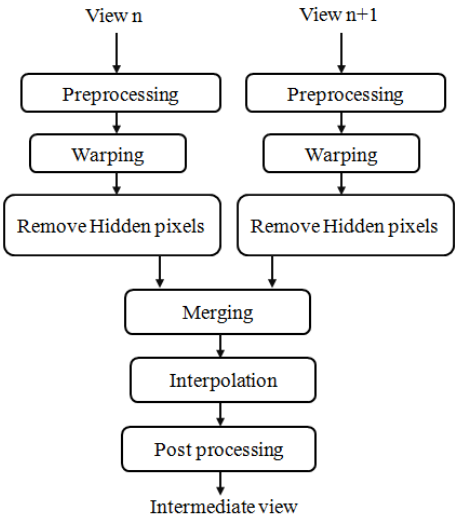


Figure 1. Block diagram of the proposed method.

The assumption of horizontal displacement implies that pixels are projected on the same line in a virtual view. Thus, a sequential, line-by-line, one-dimensional processing is possible and the proposed method (M1) can be described using the block diagram shown in Fig. 1.

In the *preprocessing* step (see Fig. 1), we add an edge pixel with horizontal shift of $\epsilon = \pm 0.01$ from the edge $\mathbf{x}$. The positive sign is used when the edge is on the right-hand side of the foreground object, and the negative sign when the edge is on the left-hand side. The edge pixel is assigned with foreground depth so that it is rendered equally much as its adjacent foreground pixel. The edge pixels become a part of interpolating background information in empty regions. Therefore, edge pixels should be assigned the background color. The background color values are selected outside the transition area between foreground and background colors; this transition in range of about 1 or 2 pixels is due to the averaging of colors in the camera pixel sensor.

The *warping* step applies the 3D warping to find the projected floating point coordinates for each pixel in the rendered view. From the assumption of 1D parallel camera arrangement, the general 3D warping equation can be simplified into:

$$u_v = u_o + \frac{f \cdot (t_{x,v} - z_{x,o})}{z_o} + (o_{x,v} - o_{x,o}), \tag{2}$$

where f is the focal length; $t_{x,v}$, $t_{x,,o}$ are the horizontal components of translation vector $\mathbf{t}$, for the virtual and original views; $o_{x,v}$, $o_{x,o}$ are the principal component offsets for the virtual and original views, respectively.

Removal of hidden pixels eliminates the occluded pixels. The occluded pixels are identified by the depth difference between two neighboring pixels: the difference greater than a threshold indicates an occluded pixel. The occluded pixels would cause errors in the interpolation process that correspond to translucent cracks in other DIBR methods.

The *merging* step (see Fig. 1) combines projected pixels from the two views by applying weighted averaging of subsets: First the two projected view coordinates are arranged in ascending order. Then, if the horizontal difference $d = x_i - x_{i-1}$ between any two of the projected coordinates is less than a threshold d_0 the pixel closer to the camera (smaller $|z|$) is selected. The pixels origin (left or right view) is not considered in this process. See Fig. 2. Thereafter, the total image width is divided into one-pixel-wide bins, i.e. the virtual view pixel position $x \pm 0.5$. Finally, a weighted average is computed over all pixels projected in each of these bins, where the weight is based on the distance to each pixel's original image. All pixels within each bin is assigned this averaged value $\otimes$ before interpolation (see Fig. 2).

The *interpolation* step allocates values to the pixel grid of the intermediate view by linearly interpolate the rendered pixels of floating point coordinates (see Fig. 2).

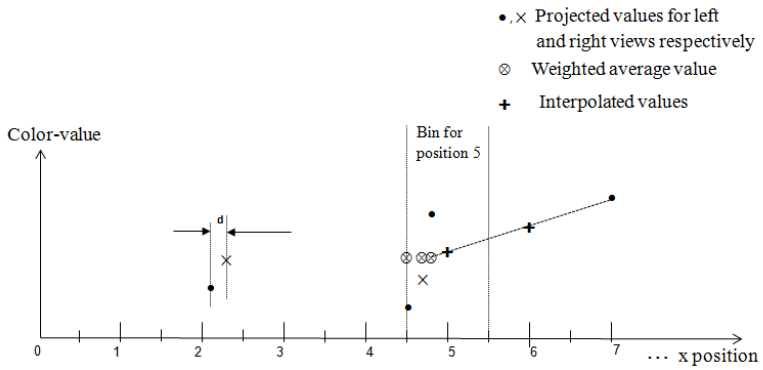


Figure 2. Weighting illustration.

The *post processing step* is applied on the virtual view and includes a low pass filtering over a small area around all edges in order to give a more natural appearance to the edges in the virtual view, and to counteract the pixilation that occurs at the edges. The low pass filter is a combination of bicubic interpolation and an average filter. The edge information is obtained from the depth map.

4. TEST ARRANGEMENT AND EVALUATION CRITERIA

The proposed method was assessed by computing given intermediate views and comparing them with given original views. Three input MVD test sequences were used for the assessment: "Poznan Hall" (SRC1), "Poznan Street" (SRC2) and "Lovebird1" (SRC3).[8,9] The sequence details are (1920x1088 pixels, 9 cameras with 13.5 cm spacing) for SRC1 and SRC2 and (1024x768 pixels, 12 cameras with 3.5 cm spacing) for SRC3. SRC1 has structured background with large depth discontinuities. SRC2 contains gradual depth transitions and lots of edges with complicated scene in the background and the last sequence SRC3 has very complex texture in the background with large depth discontinuities. For each SRC, two intermediate views were rendered at $\lambda = 0.25$ and $\lambda = 0.5$ between the two original cameras, where λ is the interpolation parameter such that $\lambda = 0$ corresponds to the left original view and $\lambda = 1$ corresponds to the right original view. The first 10 frames of the selected sequences were considered for the objective measurements and a few key frames were manually selected based on large disocclusions and scene content for the subjective test.

The proposed method (M1) was compared to the following state-of-the-art methods: MPEG-VSRS 3.5[5] and fast 1-dimensional view synthesis algorithm software,[6] which we denote as M2 and M3, respectively. Both these reference methods have incorporated many tools in the rendering process to remove the artifacts. The proposed method was implemented in MATLAB.

4.1 Evaluation criteria

The results when applying the proposed method were assessed by using objective metrics and subjective tests. The quality assessment of the rendered view through objective metrics gives detailed information about the differences to a reference image and subjective evaluations reflect the end user preferences.

4.1.1 Objective evaluation

The evaluation metric used in the objective test setup was Mean Structural Similarity index (MSSIM) as this metric has shown good agreement with subjective tests and is commonly used to measure quality of images.[10] MSSIM measures the similarity between the two images; closer value to 1 demonstrates better quality. This metric was applied at $\lambda = 0.5$ view position since no reference images are available for $\lambda = 0.25$.

In case of SRC1, the sequence views 5 and 7 were used to compute the rendered view at 6th view position. For SRC2, the sequence views 3 and 5 were used to compute the rendered view at position 4 and for SRC3, the sequence views 4 and 8 were used to compute the rendered view at position 6.

4.1.2 Subjective evaluation

Test Procedure: The subjective quality test procedure was chosen according to the goal of the study. The most commonly used test procedures from ITU-T Rec. P.910 are absolute categorical rating (ACR) and pair comparison (PC).[11] In our experiment, PC subjective test methodology was utilized in order to get reliable quality ratings. PC is the suitable method when there exist small differences between the images from various test conditions.

Apparatus and environment: The test content was presented to the observers in monoscopic mode using Alienware display (Optx AW2210, 1920x1080 full HD LCD). The subjective assessment session was conducted according to the ITU test environment including the viewing distance: 1-8 display image height, peak luminance of the screen: 100-200 cd/m^2, ratio of luminance of background behind picture monitor to luminance of picture: 0.2, chromacity of background: D_{65} and background illumination less than or equal to 20 lux.[11]

Test material and error conditions: The test images were rendered using the proposed and reference methods at two different viewpoints $\lambda = 0.5$ and $\lambda = 0.25$. The test sources (SRCs) were SRC1 (frame 150), SRC2 (frame 1) and SRC3 (frame 1) and hypothetical reference circuits (HRCs) were the proposed method M1 and reference methods M2 and M3 respectively.

Test subjects, training and randomization: A total 16 naive test observers participated in the test; they were engineering and science students with the age range 20-35 years old. A pre-screening was conducted for all participants to check for the visual acuity and color blindness by using the Snellen chart and the Ishihara chart. A training session was conducted before the test with 4 test pairs to understand the task. Two test images with different rendering methods were paired and presented to the subjects in random order. The images were presented one after the other and subjects were asked to pick one image out of each pair as their preferred image. Subjects were free to toggle between images in each pair as many times as they like before making their choice.

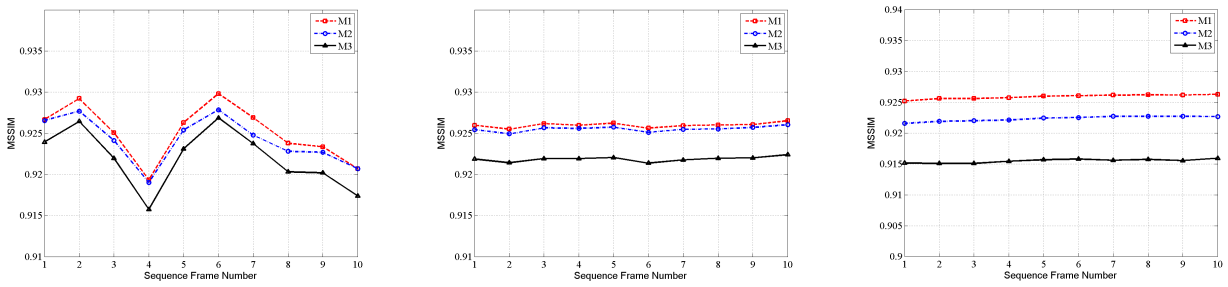
Analysis: Preferences from all subjects were then converted into a quality score using the Bradley-Terrys model. This model gives maximum likelihood estimators for scale parameters with confidence intervals, hypothesis test for model fit, uniformity and preferences among groups.[12]

5. RESULTS AND ANALYSIS

The objective measurements using MSSIM are shown in Fig. 3. The subjective test results from the pair comparison are shown in Fig. 4. The quality scores are presented for each SRC using the three rendering methods M1, M2 and M3.

The MSSIM values show improvements in the results from the proposed method M1 compared to other state-of-the art methods M2, M3 for all three test sequences (see Fig. 3).

According to the subjective scores, the proposed method M1 performs better than M3 at rendered view position $\lambda = 0.5$, but at the other investigated view position $\lambda = 0.25$, no significance difference can be noted



(a) (b) (c)

Figure 3. Objective metric MSSIM of the photographic sequences; (a) MSSIM for each rendered frame at view position 6 of "Poznan Hall"; (b) MSSIM for each rendered frame at view position 4 of "Poznan Street"; (c) MSSIM for each rendered frame at view position 6 of "Lovebird1".

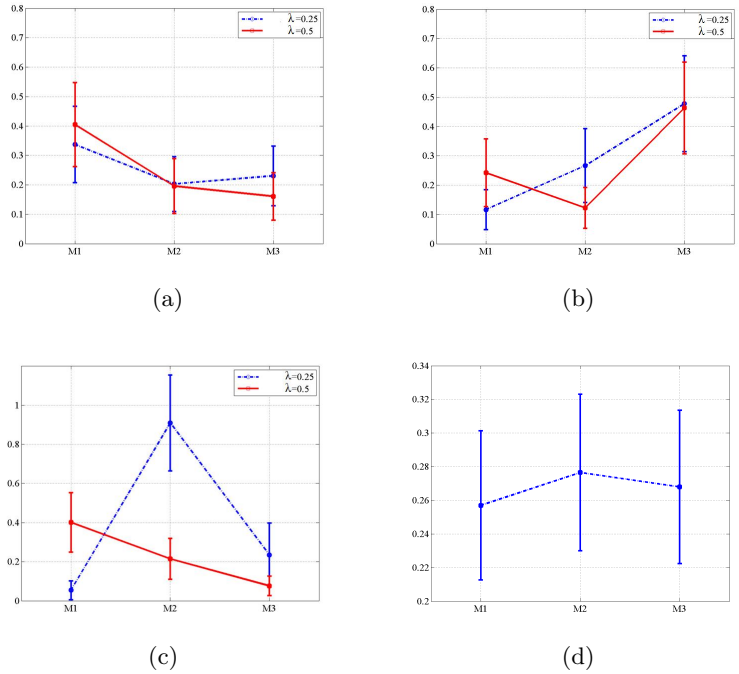


Figure 4. Subjective quality score for each sequence at different view positions; (a) "Poznan Hall"; (b) 'Poznan Street"; (c) "Lovebird1"; (d) Overall score for each method.

between the results from the proposed method and those from the reference methods. The reason is that the rendered view is closer to the original view in the case of $\lambda = 0.25$ (seeFig. 4(a)). In Fig. 4(b), no significant difference can be observed between the results from the proposed method and those from the reference methods at either rendered view positions. This may be due to the few but distinct depth changes in the scene, as there is too little information that depends on the edge aided pixels. Fig. 4(c) demonstrates that the proposed method M1 performs better than the reference method M3 at $\lambda = 0.5$, but not at other view positions, which is due to linear weighting for the nearest original view in the merging when the two input views have slightly different depth characteristics.

The results from the three SRCs reveal that the obtained subjective scores depend on the test material as well as the rendered view positions. However the overall scores in Fig. 4(d) confirm that the results from the proposed method M1 are comparable to the state-of-the-art methods. The proposed method is, however, straightforward and requires less dedicated processing for each error appearing in the DIBR process.

The objective evaluation shows that the proposed method improves quality to a certain extent, especially for sequences with a background with low frequency texture. The subjective test results could not determine a significant difference between the proposed method and the state-of-the-art methods. Nonetheless, we find this result encouraging since the proposed method employs a simple and straightforward processing structure, where the reference methods include specific processing steps to remove different artifacts.

6. CONCLUSIONS

We have proposed a depth-image-based rendering method that reduces DIBR inherent artifacts and thus eliminates the additional processing steps required to search for and address those artifacts. The proposed method introduces edge-aiding pixels before projecting pixels into their actual (floating point) positions. The projected pixels from adjacent original views are then merged by weighted averaging, followed by linear interpolation to give the values on the virtual view pixel grid. The objective evaluation showed a slightly improved quality for the rendered views using the proposed method. The subjective evaluation could not determine a significant

difference to state-of-the-art methods. Nonetheless, we find the results encouraging as the proposed method omits specific processing steps to remove different artifacts.

7. ACKNOWLEDGEMENT

This work has been supported by grant 00156702 of the EU European Regional Development Fund, Mellersta Norrland, Sweden, and by grant 00155148 of Lnsstyrelsen Vsternorrland, Sweden.

REFERENCES

[1] Fehn, C., "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI* , 93–104 (Jan. 2004).

[2] Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R., "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.* **23**, 600–608 (Aug. 2004).

[3] Muller, K., Smolic, A., Dix, K., Merkle, P., Kauff, P., and Wiegand, T., "View synthesis for advanced 3D video systems," *EURASIP Journal on Image and Video processing* **2008** (2008).

[4] Sjöström, M., Härdling, P., Karlsson, L. S., and Olsson, R., "Improved depth-image-based rendering algorithm," in [*3DTV Conference: The True Vision Capture, Transmission and Display of 3D Video (3DTV-CON)*], 1–4 (2011).

[5] "Report on experimental framework for 3D video coding," Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2010/N11631 (October 2010). Guangzhou, China.

[6] "Test model under consideration for HEVC based 3D video coding," Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2011/N12559 (February 2012). San Jose, CA, USA.

[7] Muddala, S. M., Söström, M., and Olsson, R., "Edge-Preserving Depth-Image-Based Rendering Method," in [*International Conference on 3D Imaging*], (2012).

[8] Domañski, M., Grajek, T., Klimaszewski, K., Kurc, M., Stankiewicz, O., Stankowski, J., and Wegner, K., "Poznañ multiview video test sequences and camera parameters." ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050 (2009). Xian, China.

[9] Um, G. M., Bang, G., Hur, N., Kim, J., and Ho, Y. S., "3d video test material of outdoor scene." ISO/IEC JTC1/SC29/WG11/M15371 (April 2008).

[10] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).

[11] "ITU-T p.910 subjective video quality assessment methods for multimedia applications," tech. rep., ITU-T Study Group 12 (1997).

[12] Handley, J. C., "Comparitive analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in [*IS and TS PICS Conference*], 108–112 (2001).