Karlsson, L.S.; Sjostrom, M. , "Multiview plus depth scalable coding in the depth domain," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009* , 4-6 May 2009

http://dx.doi.org/ 10.1109/3DTV.2009.5069631

MULTIVIEW PLUS DEPTH SCALABLE CODING IN THE DEPTH DOMAIN

L. S. Karlsson, M. Sjöström

Mid Sweden University
Department of Information Technology and Media
SE-851 70 Sundsvall, Sweden

ABSTRACT

Three dimensional (3D) TV is a growing area that provides an extra dimension at the cost of spatial resolution. The multi-view plus depth representation provides a lower bit rate when it is encoded than multi-view and higher resolution than a 2D-plus-depth sequence. Scalable video coding provides adaption to the conditions at the receiver. In this paper we propose a scheme that combines scalability in both the view and depth domain. The center view data is preserved, whereas the data of the side views are extracted in layers depending on distance to the camera. This allows a decrease in bit rate of 16-39 % for the colour part of a 3-view MV depending number of pixels in the first enhancement layer if one layer is extracted. Each additional layer increases the visual quality and PSNR compared only using center view data.

Index Terms— Three-dimensional displays, Video coding

1. INTRODUCTION

Three-dimensional (3D) TV has been researched for decades; however, the recent progress within display and network technology as well as software has made commercial implementations possible using displays ranging from large TV screens to mobile phones. Of the techniques that provide a 3D experience multi-view is considered one of the most promising as it can provide all necessary depth cues [1]. Multi-view contains a full resolution video sequence for each transmitted view resulting in a huge amount of data. Real-time transmission of multi-view in heterogeneous networks is possible if the redundancy between the views is exploited using video coding. The quality and bit rate can also be adapted to the conditions of the receiver using scalable video coding (SVC), where partial bit streams can be extracted from the transmitted bit stream.

The various methods of transmitting multi-view data range from transmitting all views as they were captured [2] to the 2D-plus-depth representation containing only one view and depth information [3]. Transmitting all views require a high bit rate whereas the 2D-plus-depth representation needs rendering at the receiver and has low quality in occluded parts of the scene. The multi-view plus depth representation [4], which includes multiple views with depth information for each view, is a compromise between the multi-view and the 2D-plus-depth representations. Assuming that multi-view plus depth contains fewer views than multi-view. Another option is the layered depth-image approach [5], which contains information of occluded parts of the sequence at the cost of more complexity. The multi-view video can be compressed using existing standards, including MPEG-C part 3 (ISO/IEC 23002-3) [6], that supports multi-view coding (MVC) and 2D plus depth.

The SVC extension of H.264/AVC [7] that supports temporal, spatial and quality scalability can be applied to multi-view and 2D-plus depth video. Other scalability methods using 3D data include view scalability, which enable extraction of separate views [8] and a method that adapts the multi-view sequence to the depth limitations of the display [9].

In this paper we propose a method that provides scalability in the depth domain to allow parts (macro blocks) of each frame to be extracted depending on their distance from the camera. The priority of a macro block is higher if it is closer to the camera. This allows for objects close to the camera to be rendered with higher detail and less artefacts than if the view was exempted from rendering. In addition we combine this method with view scalability to ensure that the base layer contains the central view and depth map.

The paper is organized as follows: The previous work about multi-view plus depth and SVC is briefly presented in section 2. The proposed algorithm is described in section 3. This is followed by the tests setup in section 4 and test results in section 5.

2. MULTI-VIEW PLUS DEPTH VIDEO

The multi-view plus depth representation [4] is an extension of the 2D-plus-depth representation [3] and multi-view [2]. It contains multiple pairs of conventional colour video and depth maps from different camera positions of the same scene. (See fig. 1)

Thanks to Swedish Graduate School in Telecommunications and the EU Objective 2-programme for funding.

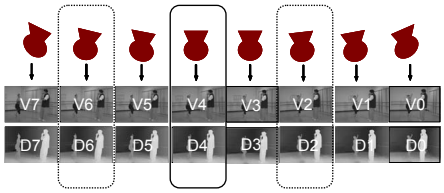

Fig. 1. Multi-view plus depth consists of multiple pairs of colour video and depth of one scene. In this case 8 cameras positioned on a straight line with view 4 as the center.

The video and the depth sequences can be encoded as separate multi-view sequences using for example H.264/AVC, hierarchical b-frames [2] and interview coding using either motion compensation [10] or disparity compensation techniques [11]. The statistical difference between depth data with slow changing surfaces and discontinuities at object borders [12] has motivated further research on new compression methods for depth data.

The SVC methods available in the SVC extension of H.264 have been applied to multi-view video in [13]. A similar approach to temporal scalability is used for view scalability where a set of views can be extracted from the sequence [8]. In addition Ramachandra et al. [9] suggest a method that adapts to the display bandwidth. Regions that are blurry due to limitations in displaying at certain depths can be encoded with less quality. The approach by Shimizu et al [14] provides a solution that uses both video data and geometry information. The base layer contains one view and its view-dependent geometry. Then the enhancement layers contain the geometry needed to transform this view into the other views and the residual of this transform.

3. THE PROPOSED SCALABILITY METHOD APPLIED IN THE VIEW AND DEPTH DOMAIN

The previous works on SVC have mainly focused on 2D relations within multi-view video, except for view scalability and adapting the quality to the depth limitations of the display. In this paper we propose a method that combines scalability in the depth and view domain under the assumption that the central view and objects close to the viewer are important. The central view and depth are assumed to provide the necessary data to render the views at reduced quality. The quality of the rendered views may then be increased by adding enhancement layers. These contain all the colour data at certain distance from the viewer that are found in the side views. Macro-blocks close to the viewer have higher priority.

The central view is encoded as a 2D plus depth video sequence (base layer); the following steps are taken to encode the enhancement layers (called layers in this paper) for each of the side views:

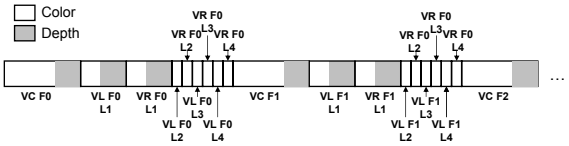

Fig. 2. The first frames $F0, F1$ of the bitstream of the central view (VC), left and right sideviews (VL and VR) are arranged such that the VC (base layer) can be extracted first and thereafter each of the layers l containing side view information.

1. We have defined two criteria to determine to which layer a macro-block (p, q) should be assigned. The first criterion states that the first layer should include objects close to the camera. This is ensured by assigning a fraction N_1 of the pixels in the frame to layer $l = 1$. The second criterion is that the remaining pixels should be equally divided between the other $L - 1$ layers. This is achieved by calculating how many pixels belong to an interval of depth values. The lower boundary of each interval is given by the threshold A_l, based on the probability $p(d)$ that a pixel (m, n) in frame f has depth value $D^{f,(m,n)}$. Thus, if $d = \lceil D^{f,(m,n)}/a \rceil$ then,

$$p(d) = \frac{H(d)}{M \cdot N}$$

where $M \times N$ is the size of the frame and H is the histogram with bin size a of one frame. Thus for all $l = 1, .., L$ we have

$$A_l = a \cdot \{\min x; 1 - P(x) \geq N_1 + (1 - N_1) \cdot \frac{l-1}{L}\},$$

where $P(x)$ is the cumulative distribution function defined as $P(x) = \sum_{d=0}^{d=x} p(d)$.

The inter and intra prediction can only be performed using macro blocks that belong to the same or a lower layer. The layer $l^{f,(p,q)}$ of the macro block (p, q) in frame f is defined as $l^{f,(p,q)} = \{\min l, D^{(f,(m,n))} \leq A_l, (m, n) \in (p, q)\}$, where $m \in [(p-1) \cdot 16, p \cdot 16]$ and $n \in [(q-1) \cdot 16, q \cdot 16]$.) Interview prediction may use any macro block in the center view as reference.

The depth data are encoded using MVC with the center view as a reference for both side views.

2. In the decoding, the center view is extracted first from the bit stream. (See fig. 2.) Thereafter the enhancement layers are extracted until the current bit rate, quality or display related requirements are fulfilled. Each block not extracted is given the YUV-values corresponding to a black macro block and is exempted from the deblocking filter.

3. The views needed by the application are rendered from the encoded colour and depth data using an appropriate

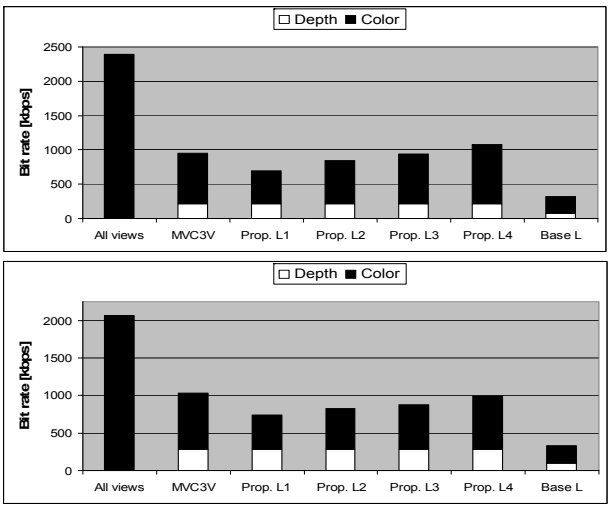

Fig. 3. The bit rate of Breakdance (top) and Ballet (bottom) for the base layer and the four enhancement layers of the proposed method, where $N_1 = 0.33$. These are compared to MSVC3V (MVC of 3 views plus depth) and MVC of 8 views.

rendering algorithm. In this paper a simple algorithm is chosen that performs 3D image warping as in [3] of the two closest views. The result is median filtered to remove small errors, before the two views are blended [15]. Holes due to missing information are filled using bilinear interpolation. Lastly, a median filter is applied to pixels where the neighbouring information does not come from the same view.

4. EXPERIMENTAL SETUP

The data sets Ballet and Breakdance (Interactive Visual Media Group, Microsoft Research) were used in the tests. The sets contains colour and depth data for 8 views, size 1024x768, frame rate 15 fps, 100 frames and a camera description. The method is applied to the first 70 frames of the views from Camera 2, 4, and 6. The remaining views are used a a reference. The encoding was preformed using the multi-view codec (MVC) JVT-X208 [16] in both its original version and a version modified by the authors to enable the scalability in the depth domain (MSVC). The missing views were rendered from the encoded material using the 8 camera position in the original data set. The impact of the rendering was tested by rendering all views using the original data from cameras 2, 4 and 6. The MSVC algorithm is compared to encoding the original three views and corresponding depth data using MVC (refered to as MVC3V in the results). The base layer and the previous enchancement layers are included in the extracted MSVC bitstream. Thus, if layer 2 is used then the base layer, layer 1 and layer 2 are all extracted.

Bit rate and PSNR are used as performance measures. The

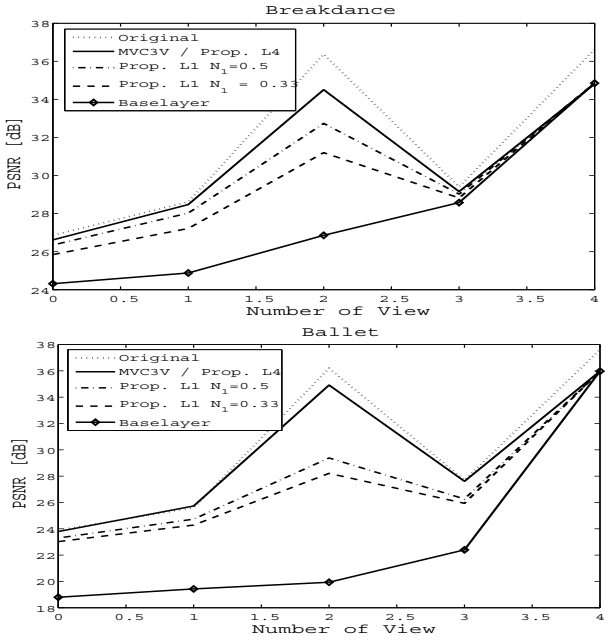

Fig. 4. The PSNR per rendered View for Breakdance (top) and Ballet (bottom). The base layer, layer 1 and 4 of the proposed method are compared to MSVC3V and the original 3 views.

PSNR per view PSNR_w is defined as

$$\text{PSNR}_w = \frac{20}{F} \sum_{f=0}^{F-1} \log_{10} \frac{255}{\sigma_w^f},$$

where F is the total frame number and σ_w^f is the standard deviation of the error of the reconstructed frame f.

5. EXPERIMENTAL RESULTS

The full sequence (layer 4) of the proposed method has an increase in bit rate of 0-17 % compared to MVC3V, assuming that PSNR is not reduced. Using layers 1-3 results in a reduction in bit rate (See fig. 3) compared to MVC3V. There is a reduction of 35 % in bit rate (Breakdance) and 39 % (Ballet), when $N_1 = 0.33$ and layer 1 is used. This reduces the PSNR by 3.3 dB and 6.6 dB, respectively. (See fig. 4). If instead $N_1 = 0.5$ there is a bit rate reduction of 16 % (Breakdance) and 33 % (Ballet). There is a significant improvement in PSNR to use the proposed scheme compared to the base layer only even if only one enchancement layer (layer 1) is added as can be seen in fig. 4. It can also be seen in fig. 4 that the rendering algorithm has a large effect on the PSNR for views 1, 2 and 3 in the original case. This is partially due to the warped pixels are slightly offset, which is not visible to the viewer. Thus, the result for view 2 of the MSVC sequence will also be affected, since it uses warped data from the center view (view 4).

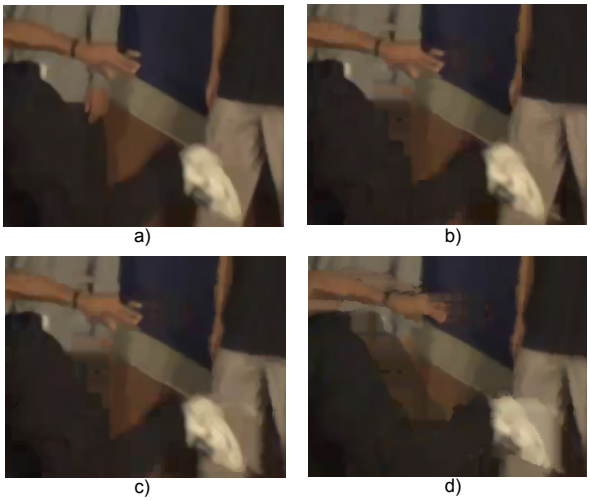

Fig. 5. A part of the rendered frame 10 of view 2. In a) the MVC3V frame is found, whereas in b) layer $l = 2$ and c) layer $l = 1$ of the proposed method with $N_1 = 0.33$ and in d) its base layer has been used in the rendering.

6. CONCLUSIONS

5. This paper presents a method that provides scalability in both the depth and view domain for a multi-view plus depth sequence. Adaptation to the local bit rate is achieved by allowing parts of the side views to be extracted to enhance the quality depending, while considering that objects close to the viewer are of importance. This enables a decrease in bit rate of 16-39 % if only layer 1 is applied in addition to the base layer. The bit rate reduction is proportional to the reduction in PNSR for the rendered views. If at least 33 % of the frontal pixels is included in the first enhancement layer, it improves the quality concerning both PSNR and visual appearance compared to using the base layer only.

7. REFERENCES

[1] J.-Y. Son, B. Javidi, and K.-D. Kwack, "Methods for displaying three-dimensional images," *Proceedings of the IEEE*, vol. 94, pp. 502–523, March 2006.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1461–1473, November 2007.

[3] C. Fehn, "Depth-image-based rendering (dibr), compression and transmission for a new approach on 3d-tv," in *Stereoscopic Displays and Virtual Reality Systems XI*. SPIE, 2004, pp. 93–104.

[4] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *ICIP*, 2007, vol. I, pp. 201–205.

[5] J. Duan and J. Li, "Compression of layered depth image," *IEEE Transactions on Image Processing*, vol. 12, pp. 365–370, March 2003.

[6] ISO/IEC 23003-3, "Mpeg-c part 3: Representation of auxiliary video and supplemental information," October 2007.

[7] H. Schwartz, D. Marple, and T Wiegand, "Overview of the scalable video coding extension of h.264/avc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1103–1120, September 2007.

[8] J.E. Lim, K.N. Ngan, W. Yang, and K. Sohn, "A multiview sequence codec with view scalability," *Signal Processing: Image Communication*, vol. 19, pp. 239–256, 2004.

[9] V. Ramachandra, M. Zwicker, and T.Q. Nguyen, "Display dependent coding for 3d video on automultiscopic displays," in *ICIP*, 2008, pp. 2436–2439.

[10] P. Merkle, K. Muller, and T. Wiegand, "Efficient compression of multiview video exploiting inter-view dependencies based on h.264/mpeg4-avc," in *ICME*. IEEE, 2006, pp. 1717–1720.

[11] J. Lu, H. Cai, J.-G. Lou, and J. Li, "An epipolar geometry-based fast disparity estimation algorithm for multiview image and video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 737–750, June 2007.

[12] K. Muller, P. Merkle, and T. Wiegand, "Compressing time varying visual content," *IEEE Signal Processing Magazine*, vol. 24, pp. 58–65, Nov. 2007.

[13] M. Drose, C. Clemens, and T. Sikora, "Extending single-view scalable video coding to multi-view base on h.264/avc," in *ICIP*, 2006, pp. 2977–2980.

[14] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-d warping with depth map," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1485–1495, November 2007.

[15] E. Cooke, P. Kauff, and T. Sikora, "Multiview synthesis: A novel view creation approach for free viewpoint video," *Signal Processing: Image Communication*, vol. 21, pp. 476–492, 2006.

[16] MVC JMVM5 Reference software JVT-X208, ," http://ftp3.itu.ch/av-arch/jvt-site/2007_06_Geneva/JVT-X208.zip.