Mårten Sjöström and Linda S. Karlsson, "Performance of scalable coding in depth domain," Stereoscopic Displays and Applications XXI, Andrew J. Woods; Nicolas S. Holliman; Neil A. Dodgson, Editors, Proc. SPIE, Vol. 7524, 75240A (2010).

http://dx.doi.org/10.1117/12.838999

Performance of Scalable Coding in Depth Domain

Mårten Sjöström and Linda S. Karlsson

Mid Sweden University, SE-85170 Sundsvall, Sweden

ABSTRACT

Common autostereoscopic 3D displays are based on multi-view projection. The diversity of resolutions and number of views of such displays implies a necessary flexibility of 3D content formats in order to make broadcasting efficient. Furthermore, distribution of content over a heterogeneous network should adapt to an available network capacity. Present scalable video coding provides the ability to adapt to network conditions; it allows for quality, temporal and spatial scaling of 2D video. Scalability for 3D data extends this list to the depth and the view domains. We have introduced scalability with respect to depth information. Our proposed scheme is based on the multi-view-plus-depth format where the center view data are preserved, and side views are extracted in enhancement layers depending on depth values. We investigate the performance of various layer assignment strategies: number of layers, and distribution of layers in depth, either based on equal number of pixels or histogram characteristics. We further consider the consequences to variable distortion due to encoder parameters. The results are evaluated considering their overall distortion verses bit rate, distortion per enhancement layer, as well as visual quality appearance. Scalability with respect to depth (and views) allows for an increased number of quality steps; the cost is a slight increase of required capacity for the whole sequence. The main advantage is, however, an improved quality for objects close to the viewer, even if overall quality is worse.

Keywords: multi-view, 3D video, scalable coding, compression, depth values

1. INTRODUCTION

The technology involved in three-dimensional video (3DV) has matured rapidly in the last couple of years and the interest in 3DV has resulted in a range of applications including 3D cinema,[1] 3DTV[2] and mobile phones.[3] Multi-view presentation techniques can provide all necessary depth cues[4] and is therefore considered a promising technique to provide 3D experience for multiple viewers without discomforting glasses and less restriction on head movement. The quality and transmission bit rate of multi-view in heterogeneous networks can be adapted to network conditions and the receiver presentation device by using scalable video coding (SVC), where partial bit streams can be extracted from the transmitted bit stream. A question that arises is whether *depth* information present in certain encoding formats can be utilized to further improve quality under variable transmission conditions.

The various methods of transmitting multi-view data range from transmitting all views as they were captured[5] to the 2D-plus-depth representation containing only one view and depth information.[6] 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,[7] 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 includes the layered depth video (LDV) approach,[8] which contains information of occluded parts of the sequence at the cost of more complexity, a higher sensitivity to errors in the depth data and that blending data from several views as in multi-view plus depth is not possible. The representation depth enhanced stereo (DES)[9] enhances a stereo pair by providing additional depth and occlusion layers to extend the adaptability of the representation. The multi-view video can be compressed using existing standards, including MPEG-C part 3 (ISO/IEC 23002-3),[10] that supports multi-view coding (MVC) and 2D plus depth.

The SVC extension of H.264/AVC[11] 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

E-mail: marten.sjostrom@miun.se, linda.karlsson@miun.se

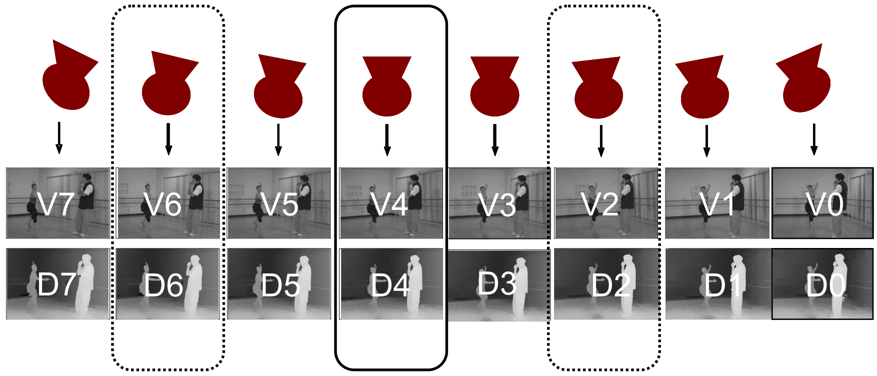


Figure 1. Multi-view plus depth consists of multiple pairs of color video and depth of one scene. In this case 8 cameras positioned on a straight line with view 4 as the center.

enable extraction of separate views[12] and a method that adapts the multi-view sequence to the depth limitations of the display.[13]

In the present paper we consider scalability with respect to depth information on top of the view scalability, which allows for an increased number of quality steps depending on accessible network capacity. Our proposed scheme is based on the multi-view-plus-depth format where the center view data are preserved, whereas side views are extracted in layers depending on depth values. We introduce two new schemes to improve assignment of pixels to the enhancement layers based on distribution of depth information. Furthermore, we investigate how these layer assignment strategies affect the performance of this scalability approach. In particular, we evaluate the number of layers, as well as the distribution of layers in depth. The consequences to variable distortion due to the choice of encoder parameters are also considered. The results are evaluated considering bit rate verses distortion, quality per enhancement layer, as well as visual quality appearance. In addition, the scheme's flexibility is evaluated with respect to its increased use of network resources when transmitting all layers.

2. MULTI-VIEW PLUS DEPTH VIDEO

The multi-view plus depth representation[7] is an extension of the 2D-plus-depth representation[6] and multi-view.[5] It contains multiple pairs of conventional color video and depth maps from different camera positions of the same scene. (See fig. 1)

The video and the depth sequences can be encoded as separate multi-view sequences using for example H.264/AVC, hierarchical b-frames[5] and interview coding using either motion compensation[14] or disparity compensation techniques.[15] The statistical difference between depth data with slow changing surfaces and discontinuities at object borders[16] 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.[17] A similar approach to temporal scalability is used for view scalability where a set of views can be extracted from the sequence.[12] In addition Ramachandra et al.[13] 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[18] 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. PROPOSED METHOD

We have in an earlier paper[19] proposed to utilize the depth information as the basis for a finer division of enhancement layers in scalable coding. The ultimate aim of such an approach is to have a smoother degradation of quality as a result of temporal capacity degeneration in the distribution chain. The cost of introducing this capability is a slight increase in bit rate requirements for the sequence with full quality.

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. Our approach introduces a combination of 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 color 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 (2D plus depth) constitutes the base layer in the multi-view scalable video coding. Enhancement layers consist of color and depth data in the side views that may be predicted from the base layer or enhancement layers of lower order. In the original paper,[19] enhancement layer zero (L_0), was attributed 33 % of the pixels closest to the viewer; the remaining pixels were distributed uniformly between higher order enhancement layers. A uniform distribution sometimes causes layers to have boundaries across objects in the scene, which implies artifacts in the final view as a result of the rendering.

The present paper addresses two novel schemes for assigning color data to the enhancement layers. In both the alternative schemes, the number of pixels assigned to L_0 is based on the characteristics of the depth data distribution, instead of a fixed percentage as applied in the original scheme. Consequently, enhancement layer zero can be adapted to the actual position of the front most objects. The assignment of the pixels to the remaining enhancement layers differs in the two novel schemes: the first applies the uniform distribution scheme of the original paper; the second utilizes the characteristics of the depth data distribution. In particular, the second scheme also determines the number of enhancement layers based on depth data distribution. The schemes can be summarized as follows, where the different schemes only differs at step 2; all other steps use the same code:

1. Compute the distribution $h(d)$ of depth data for each enhancement view.
2. Define thresholds T_i for the enhancement layers L_i for each enhancement view.
3. Assign each macro block $M_k(p,q)$ of size $k \times k$ to an enhancement layer for each enhancement view.
4. Encode using inter- and intra-view prediction based on H.264./MVC.

The distribution in step 1 is computed according to $h(d) = \frac{H_a(d)}{M \cdot N}$, where $M \times N$ is the size of the frame, $H_a(d)$ is the histogram with bin size a, $d = \lceil D^{f,(m,n)}/a \rceil$ is the bin number, and $D^{f,(m,n)}$ is the depth value of each pixel (m,n) of frame f. Step 2 is explained in more detail below. In step 3, each macro block $M_k(p,q)$ is assigned to an enhancement layer L_i by considering what enhancement layers each pixel (m,n) of the macro block ($m \in k \cdot [p-1,p]$, $n \in k \cdot [q-1,q]$) belongs to. The layer closest to the viewer is selected. The inter- and intra-view prediction for color data in step 4 can only be performed using macro blocks that belong to the same or a layer of lower-order. Interview prediction may also use any macro block in the center view as reference. The depth data are not divided into layers but rather encoded using H.264/MVC with the center view as a reference for both side views.

When decoding the sequence, the center view (base layer) 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. Thus, if layer 2 is used, then the base layer, layer 1 and layer 2 are all extracted. Each block not extracted is given the YUV-values corresponding to a black macro block and is exempted from the de-blocking filter of H.264. The views needed by an application are rendered from the encoded color and depth data using a rendering algorithm, at the choice of the user.

3.1 Thresholds of enhancement layers

The enhancement layers in the original paper (*Scheme A*) employs a uniform distribution of pixels between the enhancement layers, except for L_0 that is attributed 33 % of the pixels closest to the viewer. The thresholds can be computed as $T_i = a \cdot \{\min x; 1 - h(x) \geq N_A + (1 - N_A) \cdot \frac{i-1}{L}\}$, where L is the number of enhancement layers, and $N_A = 0.33 \cdot M \cdot N$ is the number of pixels in L_0. (See fig. 3(a).) *Scheme B*, introduced in this paper,

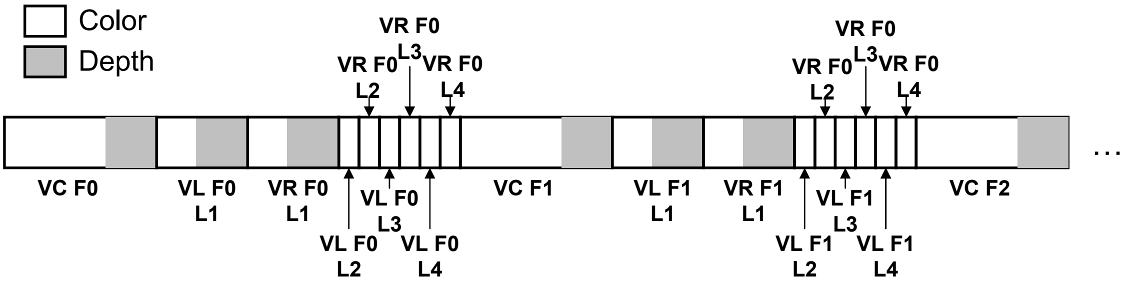
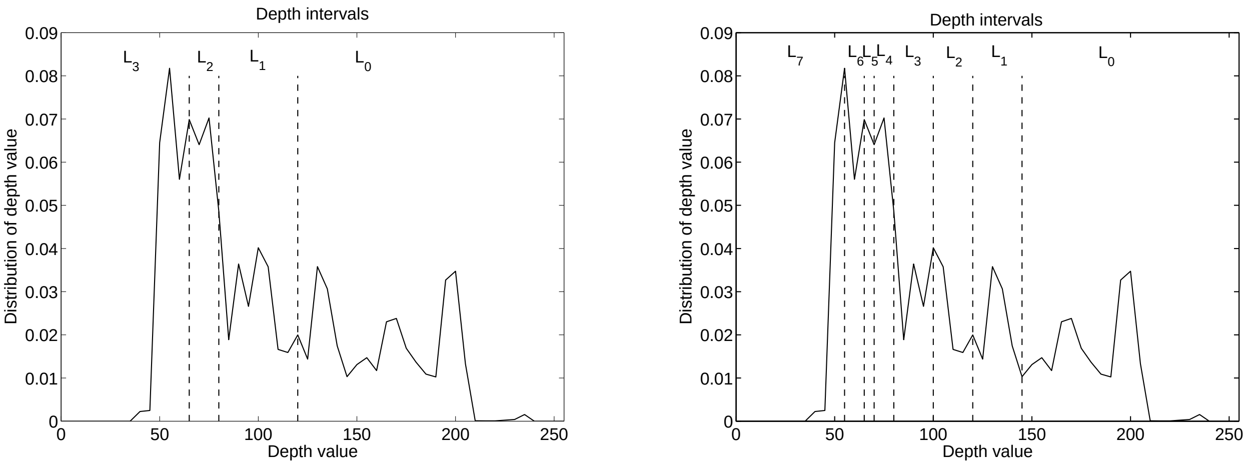


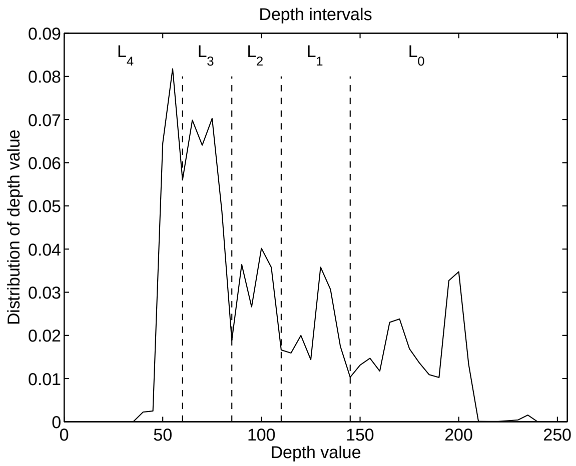
Figure 2. The first frames $F0, F1$ of the bit stream of the central view (VC), left and right side views (VL and VR) are arranged such that the VC (base layer) can be extracted first and thereafter each of the layers l containing side view information.



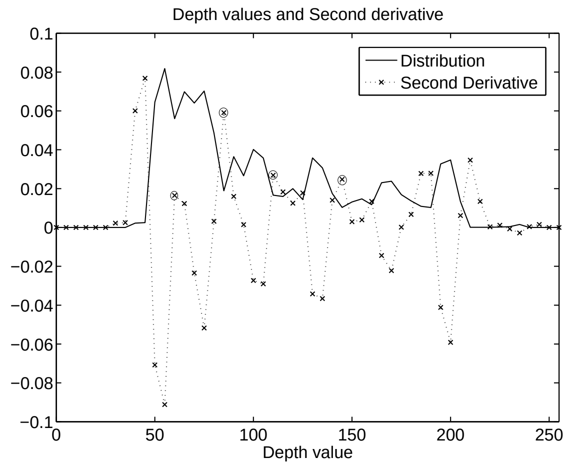
(a) Scheme A

(b) Scheme B

Figure 3. Scheme A employs a uniform distribution of pixels between the enhancement layers, except for L_0 that is attributed 33 % of the pixels closest to the viewer. Here the number of enhancement layers is four. Scheme B employs a uniform distribution of pixels for higher order enhancement layers, and L_0 is defined by the characteristics of the depth distribution $h(d)$. Here the total number of enhancement layers is eight.



(a) Scheme C



(b) Second derivative

Figure 4. Scheme C determines the size of the enhancement layers according to an analysis of the depth distribution $h(d)$. This results in different number of enhancement layers, here five. Local extreme values can be identified by considering the second derivative of $h(d)$. The values in the figure are computed by filtering $h(d)$ with the one-dimensional Sobel operator $[-1, 0, 1]$ twice. The circled values are selected as thresholds, as they are the largest between two maxima of $h(d)$.

determines the first threshold T_0 based on the characteristics of the depth distribution $h(d)$, where the minimal size of L_0 is set to 10 % of the total number of pixels. The higher order enhancement layers are given a uniform distribution of pixels, in the same way as in Scheme A. (See fig. 3(b).) *Scheme C* is an extension to Scheme B in the sense that all thresholds T_i are determined based on the distribution characteristics. Furthermore, it defines the number of enhancement layers as a consequence of the distribution analysis. In fig. 4(a), the number of layers turned out to be $L = 5$.

An analysis of the depth distribution $h(d)$ is carried out in order to identify appropriate thresholds T_i in Scheme B and C. Local minima and maxima of $h(d)$ are identified by considering positive and negative values of its second derivative, respectively. The thresholds T_i are selected as the depth values for the largest value of the second derivative between two local maxima. (See fig. 4(b).)

4. EXPERIMENTAL SET-UP

The three depth scalability schemes for layer assignments presented in the previous section have been evaluated and compared with respect to their effect on quality depending on transmission rates.

The schemes were applied to the data sets Ballet and Breakdance (Interactive Visual Media Group, Microsoft Research). The sets contains color and depth data for 8 views, size 1024x768, frame rate 15 fps, 100 frames and a camera description. The first 60 frames of the views from Camera 2, 4, and 6 were used. The encoding was preformed using the multi-view codec (MVC) JVT-X208[20] in both its original version and a version modified by the authors to enable the scalability in the depth domain (MSVC), with macro block size 16×16. Scheme A was investigated with four and eight enhancement layers; Scheme B with eight enhancement layers. Scheme C was limited to a maximum of eight enhancement layers. Enhancement layer L_0 of Scheme B and C was constrained to contain a minimum of 10 % of the number of pixels per frame.

A simple rendering algorithm was selected for the experimental set-up, which performs 3D image warping as in[6] of the two closest views. The resulting two views were median filtered to remove small errors, before they are blended.[21] Holes due to missing information were filled using bilinear interpolation. Lastly, a median filter was applied to pixels where the neighboring information did not come from the same view.

Views were rendered from the decoded sequences at the positions of camera 1, 2, 3 and 4. (See fig. 1.) These views were then compared to the views at the same positions, but rendered from the original 2D-plus-depth data at camera positions 2 and 4. In this way, the errors introduced by the rendering method are minimized.

As quality metrics, we have applied total PSNR with respect to transmission bit rate, and PSNR per enhancement layer. The total PSNR is computed as the average of the views at camera positions 1, 2, 3 and 4 for the full video sequence. The PSNR per view v is defined as

$$\mathrm{PSNR}_v = \frac{20}{F} \sum_{f=0}^{F-1} \log_{10} \frac{255}{\mathrm{MSE}_v^f}, \tag{1}$$

where F is the total number of frames and MSE_v^f is the mean square error of the reconstructed view v in frame f. The PSNR per enhancement layer i is defined as

$$\mathrm{PSNR}_i = 20 \log_{10} \frac{255}{\mathrm{MSE}_i}, \tag{2}$$

where MSE_i is the mean square error for pixels in layer i, based on data in all frames of the rendered virtual views at camera positions 1, 2, 3 and 4. Note that this evaluation has used a post-calculated pixel-based layer assignment, whereas the layer assignment in the encoding is based on macro block: a pixel (m, n) of frame f in view v belongs to enhancement layer l as decided by

$$\max l; D_{org,v}^{f,(m,n)} \geq T_{l,v}, \tag{3}$$

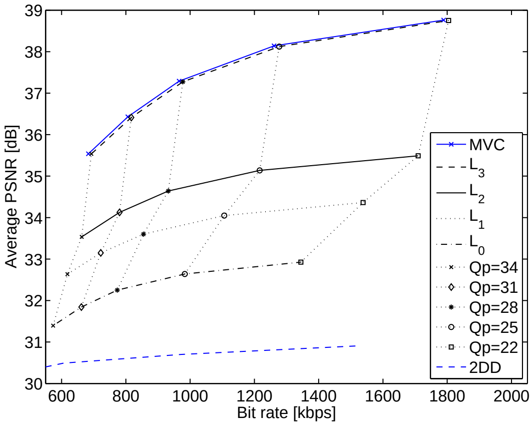
where $D_{org,v}$ are the original depth maps of the sequence and $T_{l,v}$ are the thresholds for each of the views v used in the encoding. The thresholds for the purely rendered views $v = 1$ and $v = 3$ (not part of encoded sequence) are calculated using $D_{org,v}$ and the algorithm in Section 3.1. We have also judged the results with respect to general visual appearance of selected views, based on pixelation, ringing noise manifestation, and blurriness.

5. RESULTS AND ANALYSIS

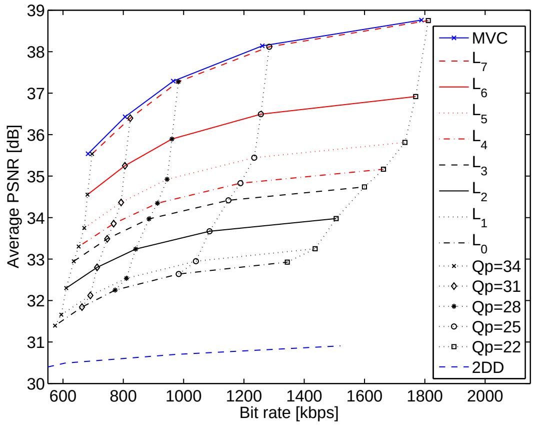
Threshold computations for Scheme B and C resulted in a first enhancement layer that encompassed the objects closest to the viewer. The number of layers for Scheme C varied over the sequence. View 2 had a mean of five and four layers for Ballet and Breakdance, respectively. The limitation to eight layers had no influence in the experiment.

Only plots for the sequence Ballet is presented here and compared between schemes. (See fig. 5). The results based on the sequence Breakdancer are similar. The graphs in the figures should be interpreted as follows. The bottom graph is quality in PSNR for the sequence encoded with 2D plus depth (2DD) only, i.e. the central view only. The MSVC clearly improves quality compared to 2D plus depth. The top graph is quality when encoded with MVC. The graph just below encompasses all enhancement layers using MSVC. Thus, MSVC introduces a slight reduction in coding efficiency compared to MVC; required bit-rate increases with about 15-20 kbit/s with all of four enhancement layers. For eight enhancement layers required, this increase is more than the doubled with eight enhancement layers. A better quality can, therefore, be obtained with MVC, in case the transmission bit rate can be assured. However, at a temporal reduction in available bit rate below a certain limit, the MVC would result in loss of all views. The MSVC, on the other hand, would instead exclude the highest enhancement layer and so result in a sequence with reduced quality. The reduction in quality and bit rate follows the 'vertical' lines, defined by the applied quantization values Qp.

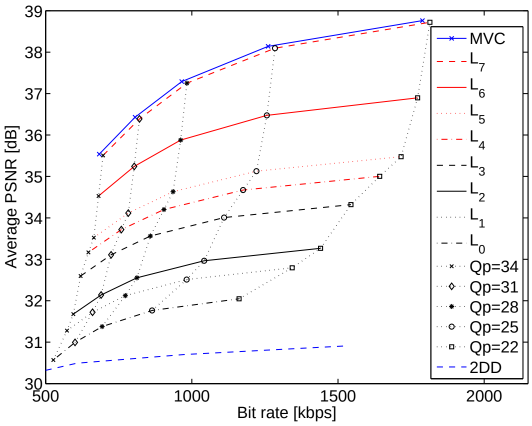
More layers (c.f. fig. 5(a) and 5(b)) imply a smoother degradation in quality at a reduction in transmission capability. Scheme B has the advantage to adjust the first layer to the scene content. This assures enough pixels to be in the first enhancement layer L_0 to produce a good quality of the objects closest to the viewer. Fig. 5(c) shows that L_0 is transmitted at lower bit rates than L_0 for Scheme A with eight layers (fig. 5(b)). Hence, the foremost objects can be presented in good quality at lower bit rates when using Scheme B. Scheme C adjusts the layers to all objects in the scene, such that an object (or group of objects) at a certain depth is in one layer.



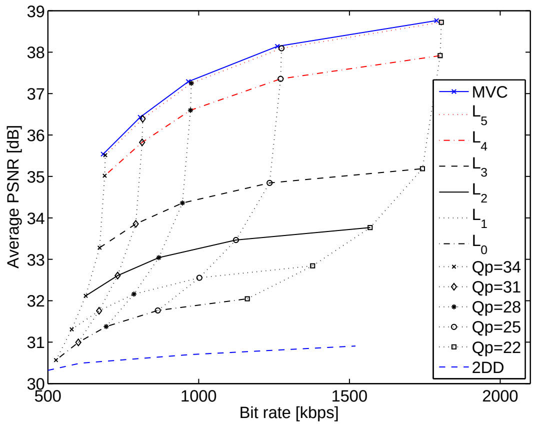
(a) Scheme A, 4 layers



(b) Scheme A, 8 layers



(c) Scheme B, 8 layers



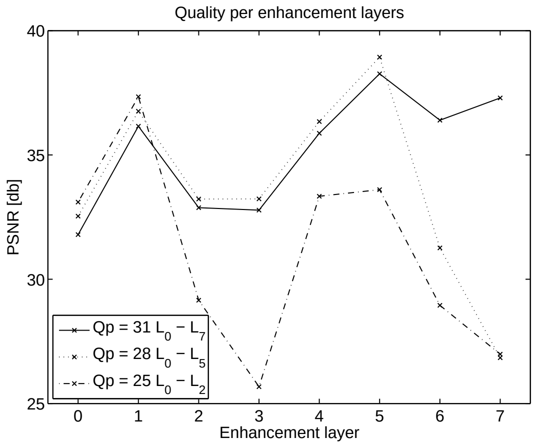
(d) Scheme C

Figure 5. Results for the sequence Ballet.

(a) $Qp = 31$, $L_0 - L_7$

(b) $Qp = 28$, $L_0 - L_5$

(c) $Qp = 25$, $L_0 - L_2$



(d) Quality per enhancement layer

Figure 6. Quality per enhancement layer. The quality drops drastically when the enhancement layers are excluded from rendering. Low values of the quantization parameter Qp results in better quality in enhancement layers that are included in rendering, even if the overall quality is worse. Rendered frame (a) includes all enhancement layers and has the best over all quality, but artifacts are pronounced in objects close to the viewer. In (b) a lower Qp is used. Artifacts are less pronounced in objects close to the viewer even though the overall quality is worse. The frame in (c) exhibits even less artifacts in objects close to the viewer, whereas overall quality has dropped even further due to small value of Qp. The graphs in (d) depicts quality when all enhancement layers are included ($L_0 - L_7$), two layers are excluded ($L_0 - L_5$), and five layers excluded ($L_0 - L_2$).

It has the same advantage as Scheme B, but it also implies that one object at a time (starting with far objects) gets reduced quality as enhancement layers are excluded. In this way, Scheme C optimizes the number of layers.

It should be noted that the drop in PSNR is fairly large when an enhancement layer is excluded. For instance when using Scheme B with $Qp = 28$, the PSNR drops to just below 36 dB when excluding enhancement layer L_7. The graphs reveal that a higher PSNR would be obtained using MVC with $Qp = 31$, at an even lower bit rate. In this sense, the proposed depth scalability method is less advantageous. However, the visual examination discloses that objects closer to the viewer are given a better visual quality with the proposed method, even though the MVC with a larger quantization parameter Qp results in a larger overall PSNR. (See fig. 6)

This conclusion is manifested when quality is measured for pixels per enhancement layer. As an example, the overall PSNR quality using Scheme A using all eight layers ($L_0 - L_7$) is 36.4 dB with $Qp = 31$; it is 34.2 dB using $L_0 - L_5$ with $Qp = 28$; and 33.3 dB using $L_0 - L_2$ with $Qp = 25$. (See fig. 5(b).) Fig. 6 shows that the quality is higher in the layers closest to the viewer for smaller quantization parameters, even though the overall quality is worse due to excluded enhancement layers. The figure also confirms how the quality drops drastically for pixels belonging to excluded enhancement layers.

6. CONCLUSIONS

Multi-view scalable video coding (MSVC) has been investigated, where scalability in both views and depth has been introduced. Two novel schemes to assign the enhancement layers of side views have been proposed, where the depth distribution decides the thresholds between layers. Hence, the layers are adjusted to the scene content; the second scheme so assigns objects at different depths to each layer.

MSVC introduces a slight reduction in coding efficiency compared to multi-view coding (MVC). However, the quality and transmission bit rate of multi-view can then adapt to network conditions and the receiver capability. The scalability in depth and views allows for finer adjustment steps. More importantly, depth scalability implies higher quality to objects closer to the viewer, even though an overall quality may be poorer than when using MVC. The adjustment of layers to depth distribution has the advantage of just enough data is assigned to each layer. A temporal decrease in transmission capacity will then imply a reduced quality reproduction of one object at a time, starting with far objects. In particular, the objects closest to the viewer may be reproduced with good quality at a lower transmission bit rate.

ACKNOWLEDGMENTS

This work is supported by the EU Objective 1 - programme Södra Skogslän region and the local government of Västernorrland.

REFERENCES

[1] Umble, E. A., "Making it real: The future of stereoscopic 3D film technology." ACM Siggraph Computer Graphics, volume 40, 2006.

[2] lifestyle technology guide., S. T., "Sony Vs Panasonic 3D HD TV Battle Emerges." http://www.smarthouse.com.au/TVs_And_Large_Display/3D _TV/D5C4Q7U4, 2009-09-29.

[3] Flack, J., Harrold, J., and Woodgate, G., "A prototype 3d mobile phone equipped with a next generation autostereoscopic display," in [*SPIE*], **6490A-21**, 502–523 (2007).

[4] Son, J.-Y., Javidi, B., and Kwack, K.-D., "Methods for displaying three-dimensional images," in [*IEEE*], **94**, 502–523 (2006).

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

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

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

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

[9] Smolic, A., Mueller, K., Merkle, P., Kauff, P., and Wiegand, T., "An overview of avaliable and emerging 3d video formats and depth enhanced stereo as efficient generic solution," in [*Picture Coding Symposium*], IEEE (2009).

[10] 23003-3, I., "Mpeg-c part 3: Representation of auxiliary video and supplemental information," (October 2007).

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

[12] Lim, J., Ngan, K., Yang, W., and Sohn, K., "A multiview sequence codec with view scalability," *Signal Processing: Image Communication* **19**, 239–256 (2004).

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

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

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

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

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

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

[19] Karlsson, L. and Sjöström, M., "Multiview plus depth scalable coding in the depth domain," in [*2009 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video : Potsdam, Germany, May 04-06, 2009*], (2009).

[20] software JVT-X208, M. J. R. http://ftp3.itu.ch/av-arch/jvt-site/2007_06_Geneva/JVT-X208.zip.

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