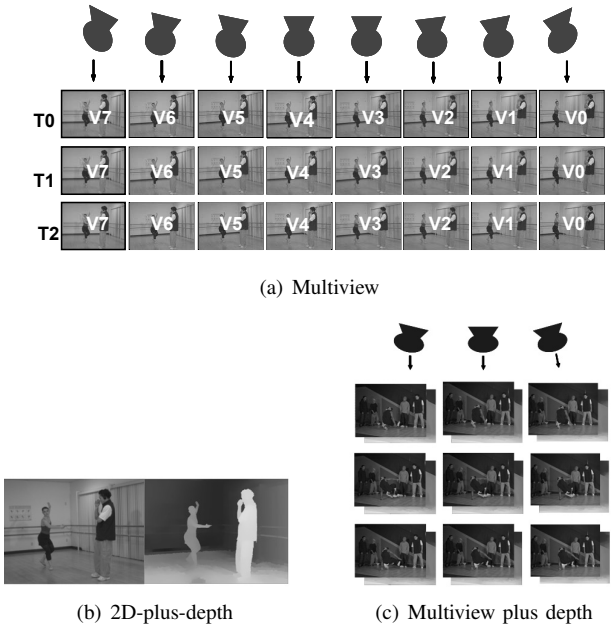
Karlsson, L.S.; Sjostrom, M. , "Layer Assignment Based on Depth Data Distribution for Multiview-Plus-Depth Scalable Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, Final manuscript accepted for publication 2010.

Layer Assignment Based on Depth Data Distribution for Multiview-Plus-Depth Scalable Video Coding

Linda S. Karlsson and Mårten Sjöström*, *Member, IEEE,*

***Abstract*—Three dimensional (3D) video is experiencing a rapid growth in a number of areas including 3D cinema, 3DTV and mobile phones. Several problems must to be addressed to display captured 3D video at another location. One problem is how to represent the data. The multiview plus depth representation of a scene requires a lower bit rate than transmitting all views required by an application and provides more information than a 2D-plus-depth sequence. Another problem is how to handle transmission in a heterogeneous network. Scalable video coding enables adaption of a 3D video sequence to the conditions at the receiver. In this paper we present a scheme that combines scalability based on the position in depth of the data and the distance to the center view. The general scheme preserves the center view data, whereas the data of the remaining views are extracted in enhancement layers depending on distance to the viewer and the center camera. The data is assigned into enhancement layers within a view based on depth data distribution. Strategies concerning the layer assignment between adjacent views are proposed. In general each extracted enhancement layer increases the visual quality and PSNR compared to only using center view data. The bit-rate per layer can be further decreased if depth data is distributed over the enhancement layers. The choice of strategy to assign layers between adjacent views depends on whether quality of the fore-most objects in the scene or the quality of the views close to the center is important.**

***Index Terms*—H.264/AVC, Multiview, Scalable video coding**



(a) Multiview

(b) 2D-plus-depth

(c) Multiview plus depth

Fig. 1. Exampels of multiview representations.

I. 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]. In the 3D movies of today, two views are processed, stored and rendered on the cinema screen, on the display of a PC and several companies are planning to provide solutions for 3D TV in 2010 [2]. These approaches provide a 3D experience through anaglyphic, polarized or shutter glasses without a motion parallax. Multiview can provide all necessary depth cues [4] and is therefore considered one of the most promising techniques to provide 3D experience for multiple viewers without discomforting glasses and less restriction on head movement. The huge amount of data necessary to depict a full resolution multiview video sequence can be reduced if the redundancy of the data is exploited using source 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.

In a 3D video system several problems need to be addressed in order to display captured 3D video at another location. One of the first problems that must be solved is how to represent the 3D data. The choice of representation influences the compression efficiency and the quality that may be achieved in the view synthesis, in addition to imposing restrictions on the capturing process. An overview of the current 3D video formats [5] show that the various methods of representing multiview data range from transmitting all views as they were captured [6] to the 2D-plus-depth (video plus depth) representation. The latter contains only one view and depth information [7]. An example of these two representations is found in fig. 1. Transmitting all views at high quality require a high bit rate, since the bit-rate increases linearly with the number of views. In case of low disparity, the bit rate can be reduced substantially by rendering the necessary views from a 2D plus depth representation.

High disparity, between existing data and the rendered

Manuscript received March 18, 2010; revised July 28, 2010

This work was supported by the Swedish Graduate School in Telecommunications and the EU Objective 2-programme.

L.S. Karlsson is with the Department of Information Technology and Media, SE-851 70 Sundsvall, Sweden e-mail: Linda.Karlsson@miun.se.

M. Sjöström is with the Department of Information Technology and Media, SE-851 70 Sundsvall, Sweden e-mail: Marten.Sjostrom@miun.se.

views, increase the impact of artifacts due to disoccluded parts of the scene. A solution is the multiview plus depth (MVD) representation [9], [10], which includes multiple views with depth information for each view as can be seen in fig. 1(c). The MVD sequence can be encoded and transmitted at a lower bit rate than if all necessary views were transmitted, under the assumption that some views were rendered at the receiver. An option is the layered depth-video (LDV) approach [11], [12] that reduces the amount of data further. LDV contains information of occluded parts of the sequence at the cost of more coding complexity, a higher sensitivity to errors in the depth data. Blending data from several views as in MVD is not possible with LDV. The representation depth enhanced stereo (DES) [5] complies with the trend in industry to provide stereo video. It enhances the stereo pair by providing additional depth and occlusion layers to extend the adaptability of the representation.

The multiview video can be compressed using existing standards, including MPEG-C part 3 (ISO/IEC 23002-3) [13] that supports 2D plus depth and multiview supported by the MVC extension of H.264/AVC [14].

The SVC extension of H.264/AVC [15] that supports temporal, spatial and quality scalability can be applied to MVD video. Other scalability methods using 3D data include view scalability, which enable extraction of separate views [16]–[18] and a method that adapts the multiview sequence to the depth limitations of the display [19].

The present work is an extension to authors' previous works [20] and [21], where an approach depth and view scalability of a MVD sequence of three views was proposed. The depth scalability concerns scalability of the color data in relation to the distance to the camera. The view scalability only considers scalability in relation to the center view. The focus of the papers was on the assignment of layers within a view. This approach enables objects close to the camera to be rendered with higher detail and fewer artifacts than from a single 2D-plus-depth sequence, but at a reduced bit rate compared to a full MVD sequence. For clarity, the main results of the previous works are explained in the present paper. The novelty of this paper is a scheme that provides depth and view scalability of more than three views. Two strategies to assign layers in adjacent views are proposed and analyzed. We also propose to include depth data in the same enhancement layer as the corresponding color data, instead of the first enhancement layer only.

The paper is organized as follows: The previous work concerning MVD and SVC is briefly presented in section 2. An overview of the proposed algorithm is found in section 3 and the details concerning the layer assignment are described in section 4 and 5. The experimental part of the paper is divided into the set up in section 6 and the result in section 7.

II. Previous works

The MVD representation [10] is an extension of the 2D plus depth representation [7] and multiview [6]. It contains multiple color video sequence, which are viewing the same scene from different camera positions, and a depth map for each image in the color sequences. (See fig. 1.) Each depth map provides the depth value per pixel of one view represented by the corresponding 2D video sequence. The depth value is converted such that the minimum $Z_{near} = 0$ and maximum $Z_{far} = 255$.

A. Coding of MVD

The MVD sequence can be encoded using methods for multiview video coding where the color sequences and the depth sequences are encoded as separate multiview sequences. H.264/AVC and hierarchical b-frames has shown to provide the highest coding efficiency [6], [8]. Interview coding at key frames can increase the coding efficiency further if the cameras are not to sparsely placed. This can be achieved using both motion compensation [22] and disparity compensation techniques [23], where the latter uses the interview statistics.

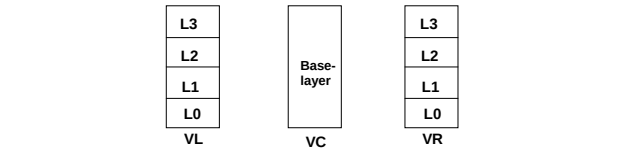
The depth data differs statistically from the color 2D video sequence due to its slow changing surfaces and discontinuities at object borders [24]. H.264 can be applied if the compression is limited to avoid severe artifacts at the discontinuities. Further improvements are possible if the statistical properties of depth data are included. The motion vectors can be estimated based on both color and depth data to allow the same set of motion vectors to be used for both [25]. The coding efficiency of motion vectors is increased in [26] by adapting the mode selection in H.264 to depth data characteristics. Another approach is to extract and encode the edges separately. The ROI image coding in JPEG2000 has been used in [27] and in [28] a scheme that segments and encodes the edges and main objects was proposed. A method aimed at segmenting and encoding of edges in video is found in [29].

Compression of MVD has been improved by using platelet-based depth coding [30]. This gives a higher rendering quality than for H.264 intra coding of depth images. Pre-processing in the form of adaptive smoothing of the depth data [31] can also be used. In addition a temporal sub sampling scheme has been proposed in [32] that uses inter-view prediction to reconstruct removed depth data.

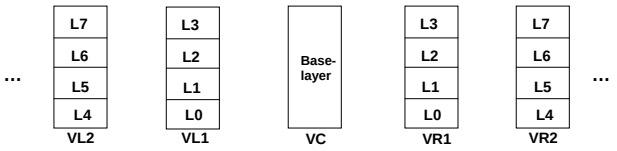
B. Scalable video coding

Multiple MVD sequences would be required to provide an overall high perceived quality in a heterogeneous network with various types of receivers. Scalable video coding produces one single video sequence from which parts of the sequence can be extracted. Hence, the video sequence can be adapted to the characteristics of a part of the network and to the receiver.

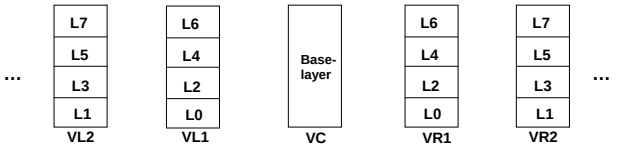
1) 2D video: Scalability can be performed in various ways and combinations. In temporal scalability the video sequence can be extracted at a reduced frame rate, whereas in spatial scalability it is the size of the picture and thus the spatial resolution that can be varied. These two types of scalability is found in the SVC extension of H.264 [15] in addition to quality scalability. Quality scalability means that the fidelity of the sequences may be varied either in clearly defined layers, CGS (coarse grain quality scalability), or continuously within these layers, MGS (medium grain quality scalability). Wavelets can be used for spatial and quality scalability; however if temporal



(a) 3 Views



(b) Center View Priority (CVP)



(c) Front Most Priority (FMP)

Fig. 2. Layer assignment between views. (a) In the case of 3 views the center view (VC) is assigned to the base layer and both the side views (VL and VR) are divided into enhancement layers depending on the depth data. In the case of more than 3 views this is extended by either prioritizing data close to the center view (b) or the front most data (c) in the layer assignment.

prediction is used drift problems appear unless the prediction is restricted to the base layer.

2) Multiview and MVD: The scalability methods available in the SVC extension of H.264/AVC has been applied to multiview video in [33], where the coding structure using hierarchical b-frames enables decomposition into layers of different frame rates (temporal scalability). These methods can be used directly on the multiview part of MVD and in [34] spatial scalability has been applied to the depth data. A similar approach to temporal scalability is used to provide view scalability where a set of views can be extracted from the sequence [16]. The approach by Shimizu et al [17] provides a scalable solution that uses both monochrome data and geometry information. A base view and its view-dependent geometry is given the highest priority. In the enhancement layers the geometry needed to transform this view into the different views are found as well as the residual of this transform. Depth scalability in the sense of providing scalability related to the depth limitations of a display is suggested in [19], since high frequency components outside of the depth range of a display may cause distortion. Wavelet transforms has also been suggested to enable temporal, spatial and quality scalability [18] and in some cases even view scalability [35], [36], but the wavelets still has some problems that reduces the coding efficiency compared to block-based approaches [37].

III. Scalability in the depth and view domains

The previous works on SVC have mainly focused on 2D relations within multiview video except for view scalability and adapting the quality to the depth limitations of the display. Depth is an important factor in 3D video, where existence of natural depth enhances the perceived quality. Additionally, the position of objects in the scene influences their contribution to the overall perceived quality. Distortion will have a higher impact on the perceived quality of objects close to the part of the scene in focus of the viewer, than on objects that are further away. The proposed algorithm for scalability in this paper is based on the assumption that the viewer focuses on the front most objects. In the case of a camera setup with zero disparity for the farthest object, objects closer to the camera will have larger disparity. The front most objects are then more sensitive to errors in the view synthesis.

The scheme proposed in this paper is mainly aimed at the encoding of MVD displayed on a 3D display or similar equipment. The intended displays have a limited viewing angle and therefore the maximal disparity between the two outmost views is limited. The minimal data that are transmitted in the scheme consist of all the color and depth data in the center view, gathered in the base layer. This ensures that background objects can be rendered even if enhancement layers containing data of that object is not included. Random access is not considered in this scheme.

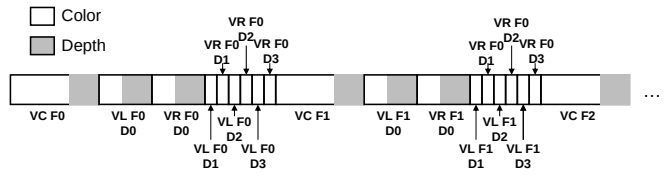
The scalability of the color data is accomplished by dividing the data into separate layers, where the base layer contains the necessary data to render the views at reduced quality. The quality of the rendered views may then be increased by adding enhancement layers. In the proposed scheme, the base layer contains the central view and the corresponding depth data. The color data of remaining views are divided into the enhancement layers as can be seen in fig. 2(a). Depth data may either be assigned to the layer of the corresponding color data, or to the first enhancement layer of the view in question.

The proposed algorithm consists of two main parts, source coding and view synthesis.

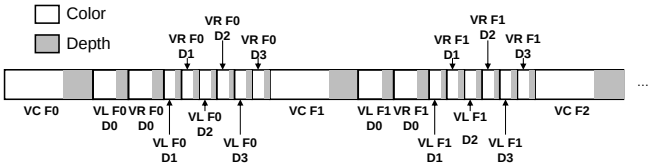
A. Source coding

The color data of the MVD sequence is encoded using a modified version of H.264/MVC [14]. The modified version provides scalability depending on the distance of objects to the camera in all views except the center view. The depth data can either be encoded using the original MVC or the modified version. In the modified version an extra step is added in the encoding process of each macro-block, where the macro-block is assigned to an enhancement layer using one out of three schemes described in section IV. The predictive coding (intra, interframe, interview) of a macro-block is then restricted to macro-blocks of the same layer or a layer of lower order. Errors due to missing data are then avoided in the decoding process. The macro-blocks are rearranged into a bit stream when all the frames at that time instance have been encoded. Fig. 3(a) shows the arrangement of the enhancement layers when all the depth data are encoded using MVC, i.e depth data is a part of the first enhancement layer of each view. The option of including the depth data in the scalable encoding is depicted in fig. 3(b). In the decoding, the center view is extracted first from the bit stream.

Thereafter the enhancement layers are extracted until the current bit rate, quality or display related requirements are fulfilled. Thus, if layer 1 is used, then the base layer, layer 0



(a) Color data encoded using MSVC, Depth data encoded using MVC



(b) Both color and depth data encoded using MSVC

Fig. 3. Extract of the bit stream. The figure depicts the bit stream of two frames $F0, F1$ of the central view (VC), along with the left and right side views (VL and VR). The bit stream is arranged such that the VC (base layer) can be extracted first and thereafter each of the enhancement layers l containing side view information. The depth data are either encoded using MVC (a) or MSVC (b).

and layer 1 are extracted. Each block not extracted is exempted from the deblocking filter of H.264 and the view synthesis.

Other types of scalability could be applied to the central view (and each of the enhancement layers), such as quality scalability, but are not included in this scheme. Hence, all blocks are treated equally concerning the level of compression.

B. View synthesis

Depending on the type of 3D display views at other positions than the camera positions available in the provided MVD sequence may be required. These views are synthesized from the decoded color and depth data using an appropriate view synthesis algorithm. In this paper we have chosen an algorithm that applies 3D image warping as in [7]. The two transmitted views closest to the desired view are used for synthesis. The views are median filtered to remove small errors, before they are blended [38]. The closest view is given priority in the blending if the requested pixel information is found in more than one view. Missing information results in holes in the blended view. These are filled using linear interpolation of the two closest pixels in the current frame and the corresponding pixel in the previously rendered frame. Lastly, a median filter is applied to pixels, whose neighbors either come from different views or bilinear interpolation.

The algorithm used in this paper could be improved by applying the recent advances in view synthesis. These include various ways to improve the handling of holes and other artifacts in the synthesized image by pre-filtering the areas with depth discontinuities [31] and inpainting [39]. The boundaries of objects are treated as separate regions in [29], [40] and in [41] three regions are used. The regions in [29], [40], [41] are warped separately and thereafter merged into one view.

IV. Layer assignment within a view

The assignment of the color data (and the depth data) to the enhancement layers is based on two criteria.

1) The front most objects should be included in the first layer.
2) The division into layers should comply with objects such that most of an object is within one layer only or partitioned in a logical way. Otherwise artifacts in the rendering of the final view may occur.

The two criteria could be fulfilled by using object-based methods [42]–[44]. However, the segmentation of objects increases the complexity of the algorithm substantially and may introduce errors. The layer assignment is therefore performed using the depth distribution layer assignment scheme (DLA) that was first proposed by the authors in [21] using the name Scheme C. DLA utilizes the characteristics of the depth data distribution both to determine the position in depth and the number of enhancement layers. DLA was shown in [21] to provide a improved performance compared to using layer assignment schemes that are based on a uniform distribution of the pixels over the enhancement layers. DLA enables L_0 to be adapted to the actual position of the front most objects.

DLA can be summarized as follows:

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 frame in the enhancement view.

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.

In step 2 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, 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 DLA. 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 which the largest value of the second derivative is between two local maxima. (See fig. 4.) The threshold for layer 0, T_0, has been selected such that at least 10 % of the pixels are assigned. This lower limit for layer 0 ensures that front most objects are assigned to layer 0.

In step 3, each macro block $M_k(p,q)$ is assigned to an enhancement layer L_i, where $p = 1, 2, ..P_k$ and $q = 1, 2, ..., Q_k$ are the indexes of a macro block in a frame with $P_k \times Q_k$ macro blocks. The pixel (m,n) is assigned to highest enhancement layer L_i for which its original depth value $D_{org}^{f,(m,n)}$ is equal or larger than the threshold T_i as described in eq. 1.

$$\max i; D_{org}^{f,(m,n)} \geq T_i \tag{1}$$

The macro block ($m \in k \cdot [p-1, p]$, $n \in k \cdot [q-1, q]$) is then assigned to the lowest layer (front most layer) of the pixels

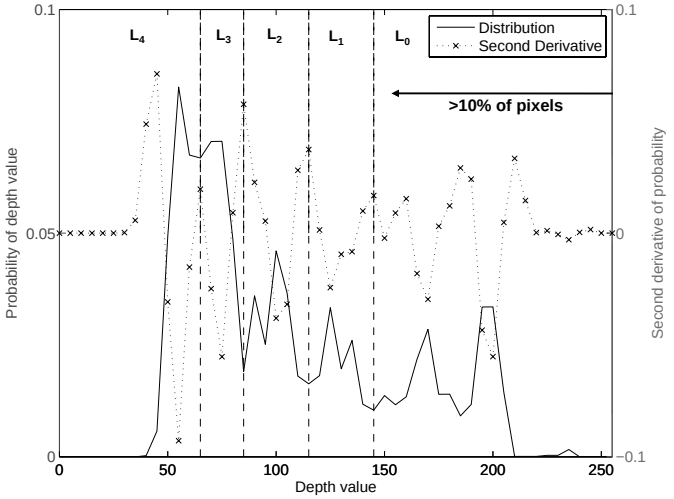
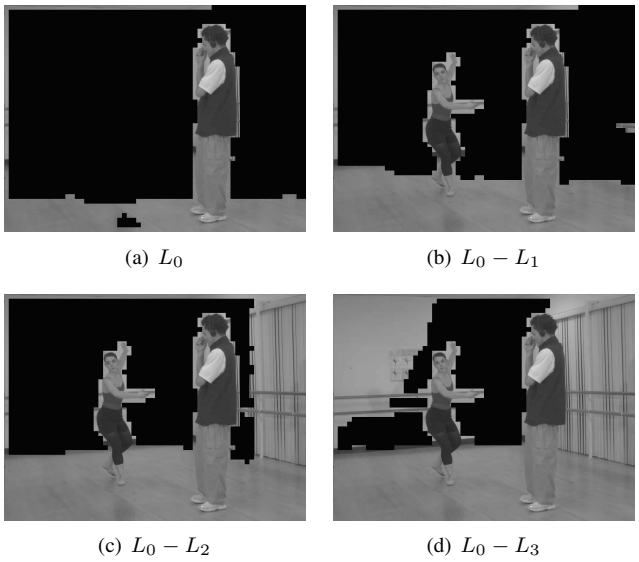


Fig. 4. DLA is based on the depth distribution. DLA determines the size of the enhancement layers according to an analysis of the depth distribution $h(d)$. This results in different number of enhancement layers, here five. Local extreme values can be identified by considering the second derivative of $h(d)$. 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 $h(d)$ and assigns at least 10 % of the pixels to layer 0.



(a) L_0 (b) $L_0 - L_1$

(c) $L_0 - L_2$ (d) $L_0 - L_3$

Fig. 5. An exampel of the layer assigment of a frame in view 2 of the Ballet sequence. The images demonstrate how more objects are included by increasing the number of layers. The black parts does not contain any data.

included in the macro block. An example is found in fig 5.

V. Layer assigment between views

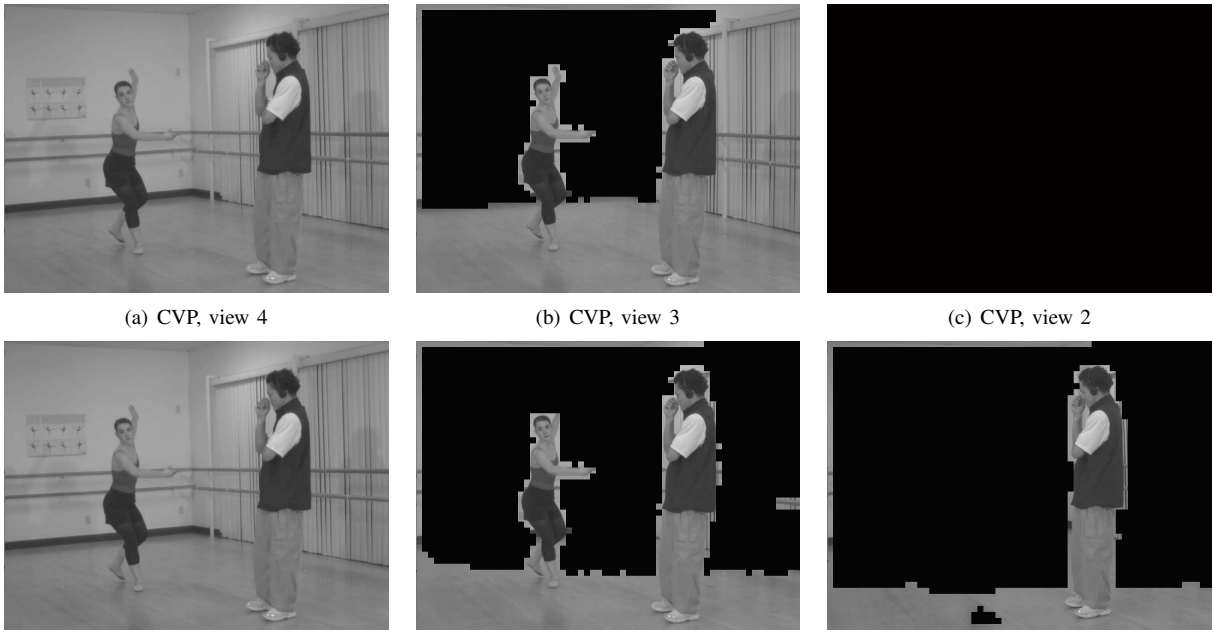
The layer assignment for three views described in section IV can be extended to include more views. The relation between the enhancement layers of adjacent views depends on which layers are given priority over the others. In this paper we propose two strategies:

1) Center View Priority CVP. For this scalability the relation to the center view is considered the most important. Thus, all layers of a view closer to the center view are extracted before the layers in the adjacent views. The priority of each layer can be seen in fig. 2(b) and an example can be found in fig. 6(a)-6(c).
2) Front Most Priority (FMP). In this scheme the front most layers are given a higher priority. Therefore the front most layers, in all views, are extracted before the background. This strategy is depicted in fig. 2(c) and an example can be found in fig. 6(d)-6(f). The view synthesis algorithm must also be modified to handle the case when the two closest views lack enhancement layers. Hence, data from views further away are necessary to provide the missing information. This strategy requires additional restrictions on interview prediction, to ensure that only data from the same or lower layers are used in the prediction.

VI. Experimental setup

The schemes and strategies proposed in this paper were applied to the data sets Ballet and Breakdance (Interactive Visual Media Group, Microsoft Research) and Book Arrival (HHI) [45]. The sets contain color and depth data for more than 5 views, size 1024x768, frame rate 15 fps, 100 frames and a camera description for each camera position. The methods are applied to the first 60 frames of each view. The encoding was performed using the version of the multiview codec (MVC) JVT-X208 [46] that was modified by the authors to enable the scalability in the depth domain (MSVC). The original JVT-X208 codec was used as a reference for the full sequence (MVC) and 2D plus depth (2DD) using the center view only. The focus in the tests was on the encoding of the color data, since it requires the most bit rate. The quantization parameters $Qp = 34, 31, 28, 25, 22$ were used for the color data of all views in the rate-distortion tests. All the depth data was encoded using $Qp = 32$, which provides the ratio 1:1 (Ballet, Breakdance) and 1:2 (Book Arrival) between the rates of the color and depth data for $Qp = 34$. A rather small compression of the depth data is chosen, since the focus of the evaluation is on the consequences of the difference in CVP and FMP. An adjustment of the compression level of the depth data to avoid noticeable rendering errors is out of the scope of this investigation.

Views were rendered from the decoded color and depth sequences of the center view and the views to one side. The synthesis algorithm described in section III-B was used. In addition to the original camera positions intermediate virtual camera positions between the views and to the right of the rightmost camera were also used. The actual camera positions of the intermediate views are interpolated from the camera parameters of the encoded views using bilinear interpolation. The aim of the tests was to determine the impact of the MSVC schemes. All reference data are therefore rendered using the synthesis algorithm mentioned in section III-B so that the resulting sequences after encoding and decoding by the MVSC schemes are compared with rendered views. The impact of the rendering algorithm on the quality metrics is thereby minimized, even if all errors due to rendering distortion remain in the resulting sequences.



(a) CVP, view 4 (b) CVP, view 3 (c) CVP, view 2

(d) FMP, view 4 (e) FMP, view 3 (f) FMP, view 2

Fig. 6. An example of a frame from the decoded Ballet sequence when layers $L_0 - L_2$ are extracted for CVP in (a)-(c) and FMP in (d)-(f). The black parts do not contain any data.

A. Layer assignment within a view

A subset of three view sequences from camera 2, 4, 6 (Ballet, Breakdance) and 7,9,11 (Book Arrival) was used to test the layer assignment within each view. View 4 is the center view for Ballet and Breakdance, whereas view 9 is the center view for Book Arrival. The DLA scheme was evaluated. The option of including the depth data in the depth scalability was also investigated. The quality analysis was made using rendered views at camera positions 1, 2, 3, 4 for Ballet, Breakdance and 6, 7, 8, 9 for Book Arrival.

B. Layer assignment between views

The two strategies CVP and FMP to assign layers between intermediate views were tested using a subset of five view sequences from camera 2, 3, 4, 5, 6 (Ballet, Breakdance) and 7,8,9,10,11 (Book Arrival). Views at camera positions 1.5, 2, 2.5, 3, 3.5, 4 (Ballet, Breakdance) and 6.5, 7, 7.5, 8, 8.5, 9 (Book Arrival) were rendered and used in the quality tests. The view positions with non-integer values are between the original camera positions. E.g., the intermediate view at position 2.5 is defined as the camera position with equal distance to position 2 and 3 in the original sequence.

C. Evalutation criteria

As quality metrics, we have applied total PSNR with respect to transmission bit rate, PSNR per view, PSNR per depth value and temporal PSPNR. The total PSNR over all rendered views and the PSNR per view v is defined as:

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

where MSE is the mean square error over all frames in all rendered views or over one view for total PSNR and PSNR per view, respectively.

The PSNR per depth d_v is defined as

$$\text{PSNR}_{d_v} = 20\log_{10}\frac{255}{\text{MSE}_{d_v}}, \tag{3}$$

where MSE_{d_v} is the mean square error for all depth value $D^{f,(m,n)}$ within the interval $[d_v - 5, d_v + 5)$ for $d_v = 0, 10, 20, ..., 250$. The MSE_{d_v} is calculated from data in all frames of the rendered virtual views. When a pixel is rendered the depth value from the original decoded views (view position 2, 3, and 4) that are used in the rendering process is stored. The original depth value is then used as the depth value $D^{f,(m,n)}$ when calculating PSNR for the rendered views.

The temporal PSPNR was measured for each view using the PSPNR tool 2.1 [47] with the rendered original view as a reference. The mean over all rendered views of a sequence is calculated. We have also judged the results with respect to general visual appearance of selected views, based on pixellation, blurriness and rendering errors.

D. Subjective tests

The visual quality was evaluated using a subjective test with the aim to evaluate the user experience of the CVP and FMP layer assignment strategies. The test was design such that it would reflect a situation with a limited transmission bit rate: The number of layers in each tested video clip was chosen to the maximum number of layers that can be extracted without exceeding a particular bit rate for both CVP and FMP. The bit rates 1600 and 1900 kbps were used for Ballet, 1600 and 2140 kbps for Book Arrival. Each video clip contains data that were

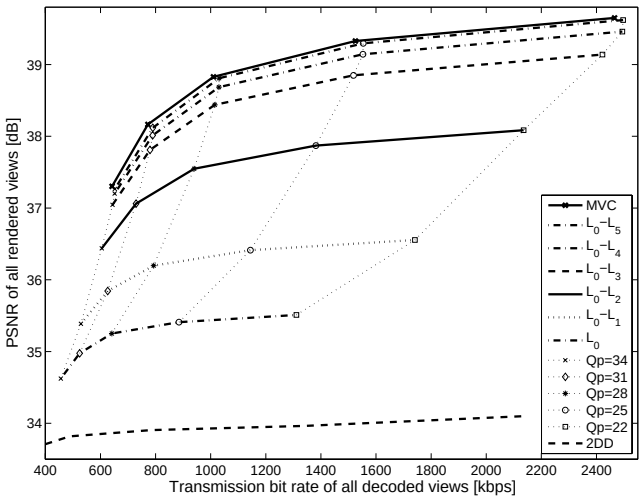


Fig. 7. Results for the subset of three views of Breakdance. The result for the DLA scheme is shown, for each extracted enhancement layer. The result for the complete tree views (MVC) and for the center view only (2DD), all encoded using MVC, are depicted as a reference.

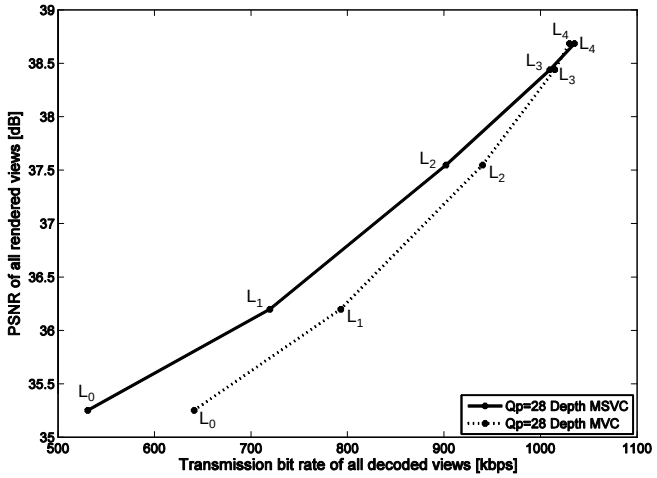


Fig. 8. PSNR for different layer assignment schemes for the depth data. The solid curve: The depth data has been encoded using MSVC and is distributed in the corresponding layer of as the color data (DLA used). The dotted curve: The depth data is encoded using MVC and the depth data for all views are included in L_0. The color data is encoded using MSVC with DLA and $Qp = 28$ in both graphs.

rendered for the camera positions 2, 3, 4, 5, 6 (Ballet) and 7 ,8 ,9 ,10 ,11 (Book Arrival). Only one position was displayed at each moment and it was changed regularly to display all positions during the video clip

The Double Stimulus Continuous Quality-Scale (DSCQS) method was applied as defined in ITU-R BT.500 [48]. Each observer was shown a series of video clips that were displayed in pairs, denoted video 'A' and video 'B'. The order within the pair was randomized for each observer. The observer was asked to assess the quality on a continuous scale in the terms of 'Bad', 'Poor', 'Fair', 'Good', and 'Excellent' as defined in [48]. The difference of the scores of video 'A' and 'B' was determined for each pair. The scores were normalized such that the maximum possible difference in the score sheet was 100. In addition to the quantitative metric, each observer was asked to answer questions concerning which kind of distortions they perceived to have the highest and lowest effect on visual quality. The test was performed using 12 test subjects with a visual acuity over 0.8. The majority of the test subjects were non-experts within this field. The tests were performed on a 22 inch screen with a resolution of 1680x1050 pixels that were viewed from approximately a distance of 1 meter.

VII. Results and analysis

The results based on the Ballet and Breakdance sequences are similar. Hence, only plots from one of the sequences are presented in this paper in addition to the results of the Book Arrival sequence.

A. Layer assignment within a view

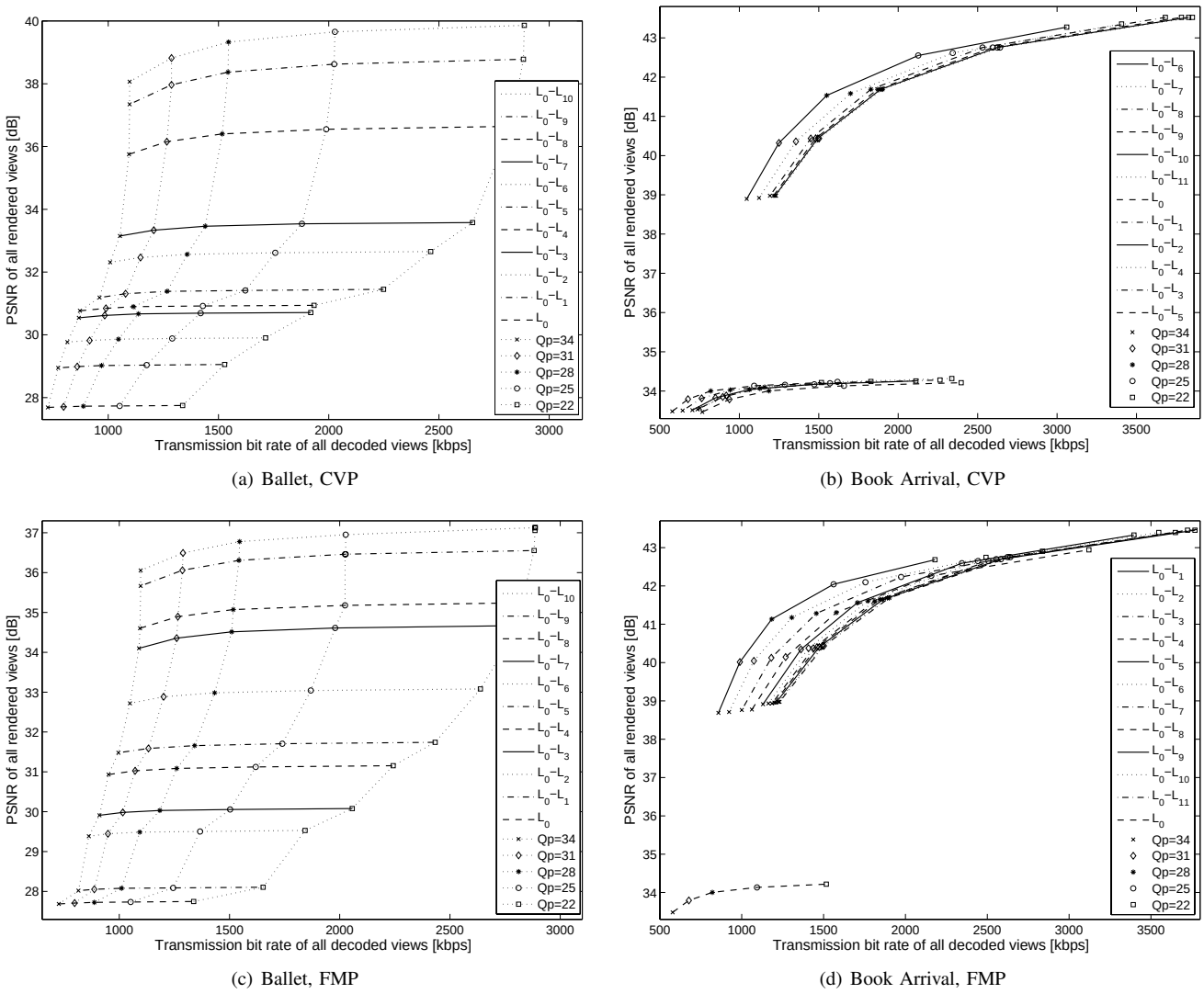
The results concerning PSNR with respect to bit rate using DLA of the Breakdance sequence are found in fig. 7 .

The PSNR diagrams in fig. 7 should be interpreted as follows. The bottom curve is quality in PSNR for the sequence encoded with 2D plus depth (2DD) only, i.e. the central view only. The top curve is quality when encoded with MVC. These curves are given as references. The curve just below encompasses all enhancement layers using MSVC. The reduction in quality and bit rate of removing one layer follows the 'vertical' lines, defined by the applied quantization values Qp. The MSVC clearly improves quality compared to 2D plus depth (See fig 7.), but compared to MVC it introduces a slight reduction in coding efficiency. A better quality is, therefore, obtained if 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 exclude the highest enhancement layer and subsequently result in a sequence with reduced quality.

The visual examination disclosed that objects closer to the viewer contain less pixelation and blurriness with the proposed method than when using all layers with a larger quantization parameter Qp. The visible rendering errors in the background are mainly in the form of discolored areas and in some instances flickering. Rendering errors due to excluded layers still influence the quality of the background as measured in PSNR despite the use of rendered sequences as a reference. These errors are due to that data are rendered from a camera at a farther distance than in the reference sequence. The PSNR measure is sensitive to displacement errors that may not be visible in a visual examination.

The complete MSVC encoded Ballet, Breakdance and Book Arrival sequences require 1 - 1.6 % more bit-rate than using the MVC for the tested Qp. The number of layers for DLA varied over the sequence. View 2 had a mean of five and four layers for Ballet and Breakdance, respectively. The corresponding view in the Book Arrival sequence, view 7, had a mean of five layers.

Fig. 8 depicts the difference of including depth data in each enhancement layer or in L_0 only. The figure shows that



(a) Ballet, CVP

(b) Book Arrival, CVP

(c) Ballet, FMP

(d) Book Arrival, FMP

Fig. 9. PSNR results for the layer assignment strategies CVP and FMP. (a) CVP for Ballet, (b) CVP for Book Arrival, (c) FMP for Ballet and (d) FMP for Book Arrival. The curves are in the same order as described in the legend as seen from above in the graph.

the quality (PSNR) remains unaltered, whereas the bit rate is reduced when the depth data are placed in the same layer as the corresponding color data. Hence, the bit-rate can be decreased by including the depth data in each enhancement layer. A visual inspection of the data verified that there was no effect on visual quality. The coding efficiency of the total MSVC encoded sequence will be slightly reduced compared to including the depth data in L_0 only.

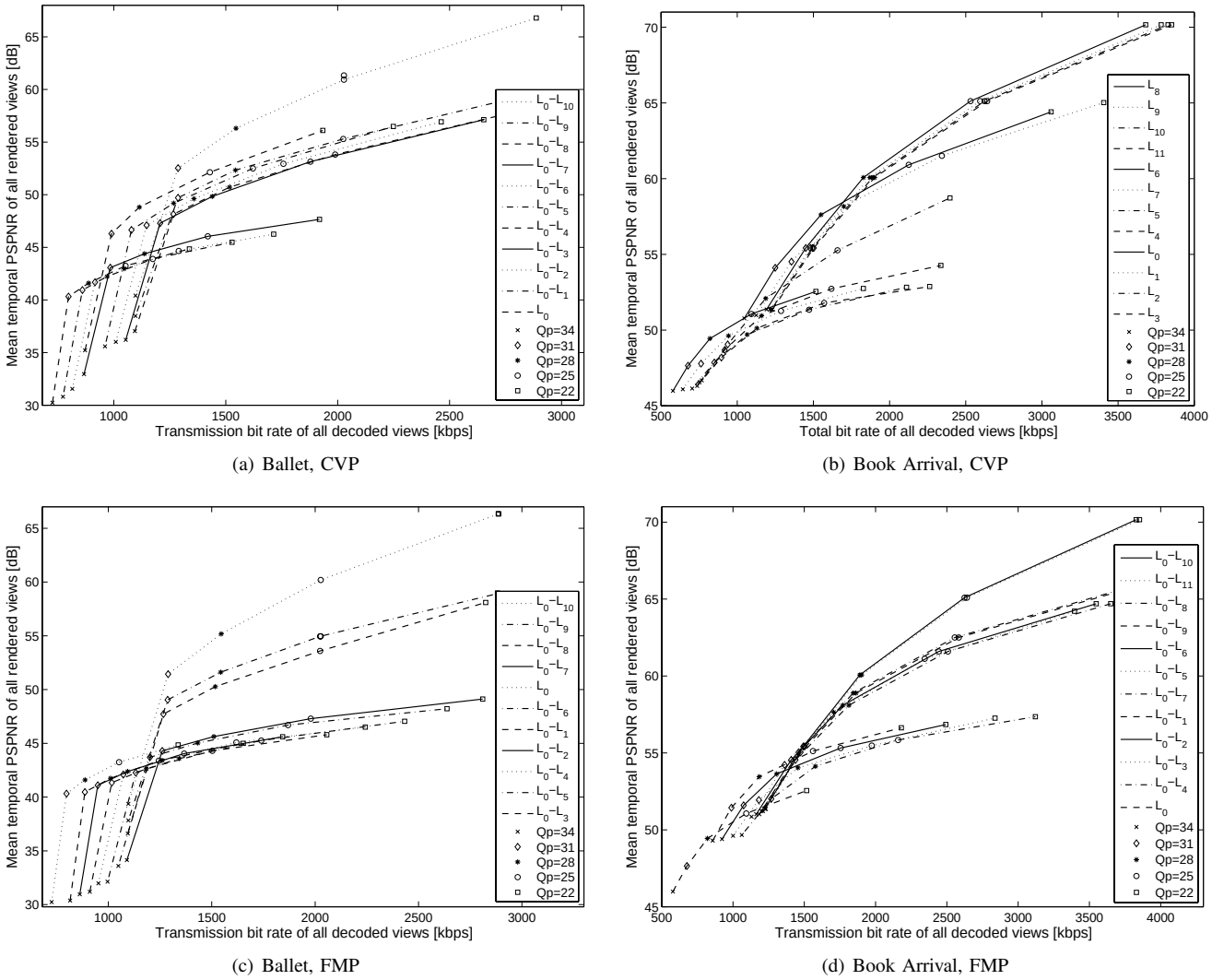
B. Layer assignment between views

The results of the tests of the MVD sequences containing five views are presented for the two strategies in fig. 9 - 13. The PSNR and bit rate for DLA of Ballet and Book Arrival are found in for CVP in fig. 9(a)-9(b) and FMP in fig.9(c)-9(d). The corresponding results for temporal PSPNR are found in fig. 10. CVP provides a better result considering the PSNR and temporal PSPNR of the total sequence with respect to the bit rate for the Ballet sequence.

The result of the Book Arrival sequence differs from the Ballet sequence in two ways. Firstly, FMP has the best performance for lower layers. Secondly, an altered Qp has relatively larger influence on quality (PSNR, temporal PSPNR) than excluded layers have. The removal of the top-most layers appears to give a better rate-distortion curve than if all layers are extracted. This is due to that the reduction of PSNR and PSPNR is minor in proportion to the reduction of bit rate.

The two layers L_0 and L_1 includes the front most pixels for all views in the case of FMP. Hence, those layers contain a larger part of the pixels for FMP than CVP for both Ballet and Book Arrival. The bit-rate will therefore be higher for FMP compared to CVP when only a few layers are extracted.

The main difference between the Ballet and Book Arrival sequences is the disparity between the views. We draw the conclusion that the FMP and CVP schemes performs better with a limited disparity between views. Furthermore, the lower the disparity, the better the FMP performs over CVP. In addition, an improved view synthesis algorithm with less sensitivity to disparity is likely to improve the results.



(a) Ballet, CVP

(b) Book Arrival, CVP

(c) Ballet, FMP

(d) Book Arrival, FMP

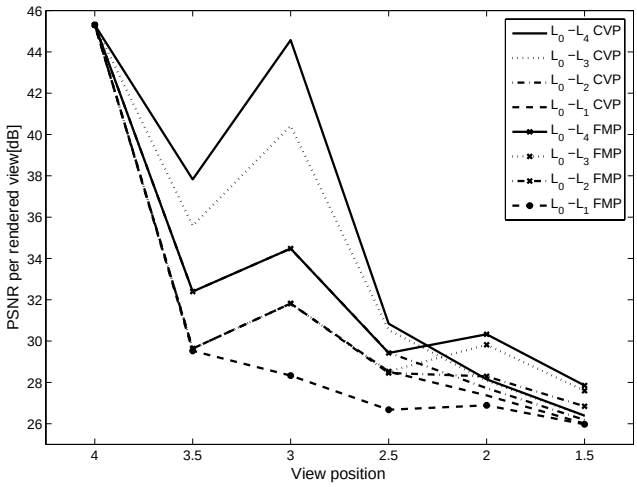
Fig. 10. Temporal PSPNR results for the layer assignment strategies CVP and FMP.(a) CVP for Ballet, (b) CVP for Book Arrival, (c) FMP for Ballet and (d) FMP for Book Arrival. The curves are in the same order as described in the legend as seen from above in the graph.

The effect of the two strategies on the quality for each view is presented in fig. 11(a)-11(b) (Ballet) and fig. 11(c)-11(d) (Book Arrival) for $Qp = 28$. The figures show the improvement of using CVP over FMP as measured in PSNR of each of the rendered views. View 4 and view 9 are the center views in Ballet and Book Arrival, respectively. The impact of the strategies on quality per depth is presented in fig. 12. The graphs contain the PSNR per depth of for $L_0 - L_2$ with $Qp = 25$, $L_0 - L_3$ with $Qp = 28$ and $L_0 - L_4$ with $Qp = 31$ for Ballet and Book Arrival. A visual example is also found in fig. 13, where parts of a frame for view 2 and 3 for $L_0 - L_3$ with $Qp = 28$ are depicted for both CVP and FMP.

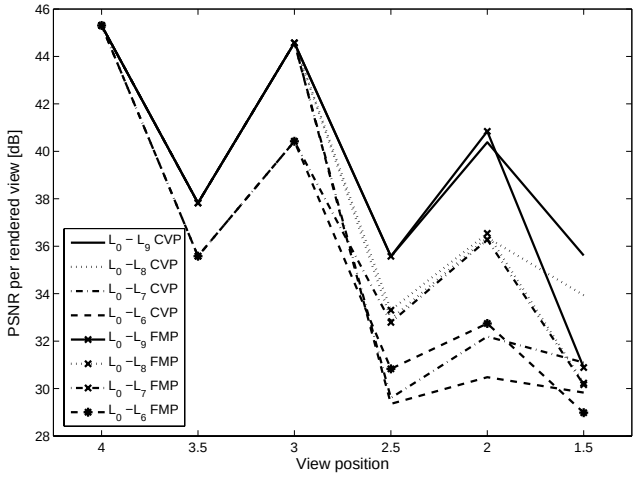
The PSNR per view in fig. 11 shows that the quality of each of the rendered views is affected by the choice of strategy. CVP provides a large increase in quality of the views close to the center view (view 4 and view 9, respectively) for each added layer, in particular, when there is a higher distance between the cameras as in the case of the Ballet sequence. FMP, on the other hand, shows a better performance than CVP for the Book Arrival, and FMP also gives a more even distribution of quality for lower layers for Ballet. These results are verified by the extracts of a frame from the two views in fig. 13. In the case of CVP of Ballet, view 3 in 13(a) has a high quality. The quality of view 2 in fig 13(b) is poor as all data must be rendered from views 3 and 4, which introduces additional rendering errors. FMP of Ballet provides an even visual quality in the two extracts in fig. 13(c) and 13(d), since the front most pixels are available in both views. However, if only view 3 is considered the CVP in 13(a) has the least rendering errors. Similar results can be seen in fig 13(e)-13(h) for Book Arrival. The test results demonstrate that the choice of strategy should be made depending on disparity between views and where quality is desired. CVP should be used in the case of high disparity or when quality of views close to the center is more important, whereas FMP should be used in the case of low disparity or when all views are of importance to the viewer.

C. Subjective test

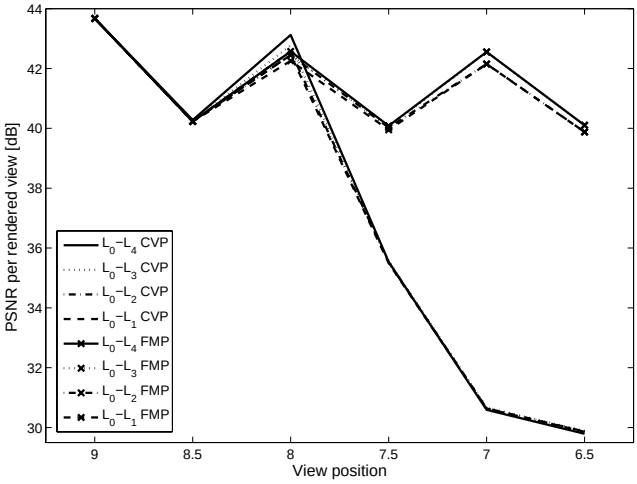
The results of the comparison of the two layer assignment strategies CVP and FMP can be found in fig. 14. The graph



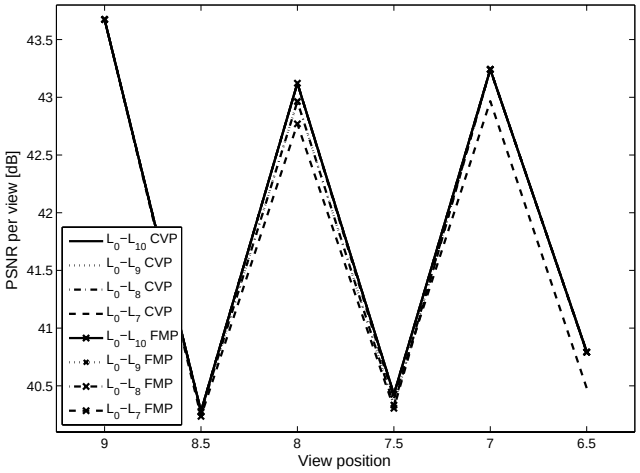
(a) Ballet, Lower layers



(b) Ballet Higher layers

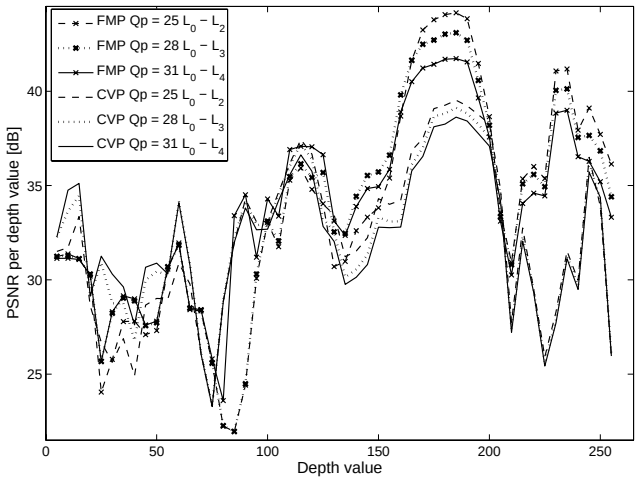


(c) Book Arrival, Lower layers

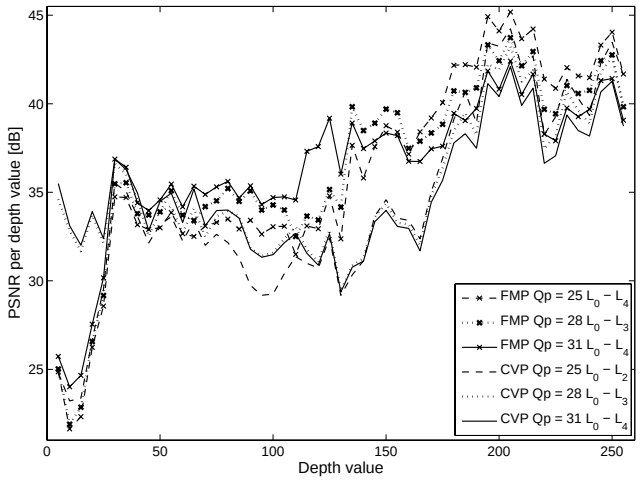


(d) Book Arrival Higher layers

Fig. 11. The PSNR per view of CVP and FMP for $Qp = 28$. The results for Ballet are found in (a) for the lower layers and (b) for the higher layers. The results for Book Arrival are depicted in (c) for the lower layers and (d) for the higher layers. Note that the center view is view 4 for Ballet and view 9 for Book Arrival.



(a) Ballet



(b) Book Arrival

Fig. 12. The curves in (a) shows quality for both CVP and FMP of the Ballet sequence when only three layers are included ($L_0 - L_2$), four layers are included ($L_0 - L_3$) and five layers are included ($L_0 - L_4$). The Qp is increased for each additional layer. The depth value 255 represents the position of the camera. The corresponding curves for the Book Arrival sequence are found in (b).

shows the difference between the scores of the CVP and FMP strategies for each video sequence, maximum bit rate and Qp. The mean score with a 95% confidence interval is depicted. The result indicates that when data from all views are shown equally to an observer, most observers prefer to have strong distortions in a few views (CVP) rather than a even distribution of the distortion over several views (FMP). However, the large confidence interval and answers to the qualitative questions indicate that the choice of priority strongly differs on an individual basis.

The subjective test, including the qualitative questions, also showed that the main visual distortions in the tested sequences are due to the view synthesis algorithm. In particular, the filling of larger holes in the view synthesis was prominent. Hence, an improved view synthesis algorithm is likely to provide a higher visual quality. The problem with the view synthesis further depends on the distance between the cameras. The tests showed that the Ballet sequence (larger camera distance) experienced more flickering and other rendering errors than the Book Arrival sequence.

The main difference between MVC and the proposed MSVC schemes is the behavior when the transmission bit rate is too low for the complete sequence to be transmitted. The views of an MVC sequence can be extracted separately depending on the choice of interview coding. Hence, the sequence may then only be transmitted if one of the views of less importance is dropped. The proposed MSVC scheme, on the other hand, provides the option of extracting parts of a view. The additional data provided in these layers improve the visual quality of the outmost rendered views compared to the case when the complete outmost view is dropped.

VIII. Conclusions

Multiview plus depth (MVD) scalable video coding has been investigated, where scalability in relation to the center view and distance to the camera has been introduced. Scalability in relation to the center view favors quality in views close the scene center, whereas scalability with respect to the distance to the camera preserves the quality of objects close to the viewer in all views.

A scheme to assign the enhancement layers of a view has been proposed. Depth distribution layer assignment (DLA) aims to assign complete objects to each layer; it decides the total number of layers and what pixels belong to what layer, depending on the depth distribution within each view. The DLA has the advantages to assign just enough data to each layer and to avoid the division of objects into multiple layers.

Two strategies has been proposed addressing layer assignment between adjacent views. The choice of strategy is based on what is the most important to visual quality of a particular application: the cameras views in relation to the scene center, or front most objects in all viewing positions. The evaluation in the paper indicates that center view priority has the better performance in the case of high disparity or when quality of views close to the center is more important. The front-most priority scheme, on the other hand, should be used in the case of low disparity. However, the subjective tests also showed that peoples experience of the two strategies vary.

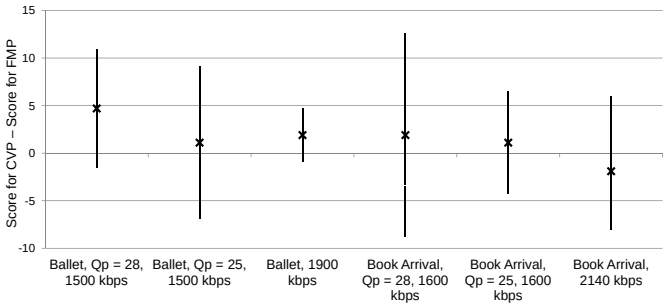


Fig. 14. A subjective comparison of the CVP and FMP strategies for Ballet and Book Arrival. The graph shows the mean values (marked as x) with a 95% confidence interval concerning the difference between the CVP and the FMP score.

A. Future work

The approach suggested in this paper is intended for applications with a limited viewing angle and disparity between the views. A larger number of views and other layer assignment strategies, including the assignment of more views to the base layer, are subject to future investigation.

The main focus in all of the tests was the scalability of the color data, since it requires a higher bit rate for high quality than the depth data. It was further assumed that depth data of a sufficient quality was provided. Future work includes a more extensive test concerning the impact of errors in the depth data, including other depth coding algorithms. The combination with other types of scalability is subject to further evaluation.

Acknowledgment

The authors would like to thank the Interactive Visual Media Group of Microsoft Research for providing the Ballet and Breakdance video sequences and HHI for providing the Book Arrrival sequence. We would also like to thank the people at Mid Sweden University that participated in subjective tests.

References

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

[2] C. Chinnock, “3D comming home in 2010,” http://www.3dathome.org, Oct. 2009.

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

[4] J.-Y. Son, B. Javidi, and K.-D. Kwack, “Methods for Displaying Three-Dimensional Images,” in *Proc. IEEE*, vol. 94, 2006, pp. 502–523.

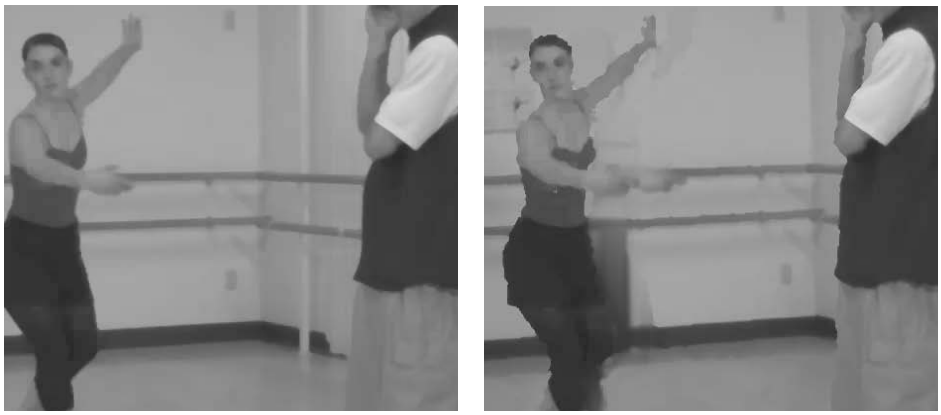
[5] A. Smolic, K. Müller, P. Merkle, P. Kauff, and T. Wiegand, “An overview of avaliable and emerging 3d video formats and depth enhanced stereo as efficient generic solution,” in *Picture Coding Symp.* IEEE, 2009.

[6] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, “Efficient prediction structures for multiview video coding,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, pp. 1461–1473, Nov. 2007.

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

[8] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G.B. Akar, G. Triantafyllidis and A. Koz “Coding Algorithms for 3DTV–A Survey,” in *Int. Symp. Circuits and Sys.* IEEE, 2007, pp. 1606 – 1621.

[9] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, “Depth map creation and image-based rendering for advanced 3dtv services providing interoperability and scalability,” *Signal Process.: Image Commun.*, vol. 22, pp. 217–234, 2007.



(a) Ballet, CVP, view 3

(b) Ballet, CVP, view 2

(c) Ballet, FMP, view 3

(d) Ballet, FMP, view 2

(e) Book Arrival, CVP, view 8

(f) Book Arrival CVP, view 7

(g) Book Arrival, FMP, view 8

(h) Book Arrival, FMP, view 7

Fig. 13. The extracts from a frame for CVP and FMP for the sequences Ballet and Book Arrival has been coded using the proposed MSVC for $QP = 28$ with DLA and rendered using layers $L_0 - L_3$. CVP provides the best quality of view 3 (Ballet) and view 8 (Book Arrival), which is closer to the center view than view 2 (Ballet) and view 7 (Book Arrival). FMP, on the other hand, provides a more even quality over both views.

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

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

[12] S.-U. Yoon, and Y.-S. Ho, "Multiple Color and Depth Video Coding Using a Hierarchical Representation,"*IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, pp. 1450–1460, Nov. 2007.

[13] I. 23003-3, "Mpeg-c part 3: Representation of auxiliary video and supplemental information," Oct. 2007.

[14] ISO/IEC 14496-10, "Advanced Video Coding," Telecommunications Union, 2009.

[15] H. Schwartz, D. Marple, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Trans. Circuits and Systems for Video Techn.*, vol. 17, pp. 1103–1120, Sep. 2007.

[16] J. Lim, K. Ngan, W. Yang, and K. Sohn, "Multiview sequence CODEC with view scalability," *Signal Process.: Image Commun.*, vol. 19, pp. 239–256, 2004.

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

[18] W. Yang, F. Wu, J. Cai, K. N. Ngan, and S. Li, "Scalable Multiview Video Coding Using Wavelet," in *Int. Symp. Circuits and Sys.* IEEE, 2005, pp. 6078 – 6081.

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

[20] L. Karlsson and M. Sjöstrom, "Multiview plus depth scalable coding in the depth domain," in *Proc. IEEE 3DTV-CON*, 2009.

[21] M. Sjöstrom and L. Karlsson, "Performance of scalable coding in depth domain," in *Conf. Stereoscopic Displays and Applications XXI.* SPIE, 2010.

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

[23] 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 Trans. Circuits and Systems for Video Technology*, vol. 17, pp. 737–750, June 2007.

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

[25] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Motion vector sharing and bitrate allocation for 3d video-plus-depth coding," *EURASIP J. Advances in Signal Processing*, 2009.

[26] M. Pourazad, P. Nasipoulos, and R. Ward, "An h.264 based video encoding scheme for 3d tv," in *Proc. EURASIP EUSIPCO*, 2006.

[27] R. Krishnamurthy, B.-B. Chai, H. Tao, and S. Sethuraman, "Compression and transmission of depth maps for image based rendering," in *Proc. IEEE ICIP*, 2001.

[28] P. Zanuttigh and G. Cortelazzo, "Compression of depth information for 3d rendering," in *Proc. IEEE 3DTV-CON*, 2009.

[29] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szelinski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graphics*, vol. 23, pp. 600 – 608, August 2004.

[30] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. de With, and T. Wiegand, "The effect of depth compression on multiview rendering quality," in *Proc. IEEE 3DTV-CON*, 2008.

[31] Y. Park, K. Jung, Y. Oh, S. Lee, J. Kim, G. Lee, H. Lee, K. Yun, N. Hur, and J. Kim, "Depth-image-based rendering for 3DTV service over T-DMB," *Signal Process.: Image Commun.*, vol. 24, pp. 122–139, 2009.

[32] E. Ekmekcioglu, S. Worrall, and A. Kondoz, "A temporal subsampling approach for multiview depth map compression," *IEEE Trans. circuits and systems for video technology*, vol. 19, pp. 1209–1213, August 2009.

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

[34] J. Cho, S. Cho, N. Hur, H. Lee, and J. Jeong, "Effective multiview video coding using a scalable depth map," in *Int. Conf. Computational Intelligence for Modelling, Control and Automation*. IEEE, 2008, pp. 255 – 259.

[35] J.-U. Garbas, U. Fecker, T. Troger, and A. Kaup, "4D Scalable Multi-View Video Coding Using Disparity Compensated View Filtering and Motion Compensated Temporal Filtering," in *Workshop Multimedia Signal Proc.* IEEE, 2006, pp. 54 – 58.

[36] Y. Liu and K. N. Ngan, "Fully scalable multiview wavelet video coding," in *Int. Symp. Circuits and Sys.* IEEE, 2009, pp. 2581 – 2584.

[37] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Processing Magazine*, vol. 24, pp. 66–76, Nov. 2007.

[38] E. Cooke, P. Kauff, and T. Sikora, "Multi-view synthesis: A novel view creation approach for free viewpoint video," *Signal Process.: Image Commun.*, vol. 21, pp. 476–492, 2006.

[39] K. Oh, S. Yea, and Y.-S. Ho, "Hole-filling method using depth based in-painting for view synthesis in free viewpoint television (ftv) and 3d video," in *Picture Coding Symp.* IEEE, 2009.

[40] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3d video systems," *EURASIP J. Image and Video Process.*, 2008.

[41] Y. Huang and C. Zhang, "A layered method of visibility resolving in depth image-based rendering," in *Int. Conf. Pattern Recogn.* IEEE, 2008.

[42] T. Sikora, "The MPEG-4 video standard verification model," in *Int. Symp. Circuits and Sys.* IEEE, 1997, pp. 19 – 31.

[43] C. Cigla and A. Aydin Alatan, "Depth Assisted Object Segmentation in Multi-View Video," in *Proc. IEEE 3DTV-CON*, 2008.

[44] S.-C. Chan, Z.F. Gan, K.-T. Ng, K.-L. Ho and H.-Y. Shum "An Object-Based Approach to Image/Video-Based Synthesis and Processing for 3-D and Multiview Televisions," in *Int. Symp. Circuits and Sys.* IEEE, 2009, pp. 821 –831.

[45] I. Feldmann, M. Müller, F. Zilly, R. Tanger, K. Müller, A. SmolicP. Kauff, and T. Wiegand, "HHI Test Material for 3D Video," in *ISO/IEC JTC1/SC29/WG11, MPEG08/M15413* , 2008.

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

[47] Y. Zhao, L. Yu, "PSPNR Tool 2.1, R. software JVT-X208,in *ISO/IEC JTC1/SC29/WG11, MPEG09/N10879* ,Kyoto 2009.

[48] "Recommendation ITU-R BT.500-11 - Methodology for the subjective assessment of the quality of television pictures, , 2002.

Linda S. Karlsson recieved the M.Sc degree in Mediatechnology and Engineering from Linköping University in 2002 and the licenciate degree from Mid Sweden University in 2007.

She is now a PhD Student at the Department of Information Technology and Media at Mid Sweden University.

Her current research interests include region-of-interest video coding and scalabe coding of 3D video.

Mårten Sjöström received the MSc in Electrical Engineering and Applied Physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in Signal Processing from KTH, Stockholm, Sweden, in 1998, and the Ph.D. degree in Modelling of Nonlinear Systems from EPFL, Lausanne, Switzerland, in 2001.

He worked as an Electrical Engineer at ABB, Sweden, 1993-1994, was a fellow at CERN 1994-1996, and a PhD-student at EPFL, Lausanne, Switzerland 1997-2001. He joined the Department of Information Technology and Media, Mid Sweden University in September 2001 as a Senior Lecturer. During 2002-2006, he was appointed Head and Assistant Head of Division, respectively. As of 2008 he is Associate Professor.

His current research interests are within system modelling and identification, as well as 2D and 3D image and video processing . He is a member of IEEE since 1992.