EXTENDING SINGLE-VIEW SCALABLE VIDEO CODING TO MULTI-VIEW BASED ON H.264/AVC

Michael Dröse, Carsten Clemens, and Thomas Sikora

Department of Communication Systems
Technical University of Berlin, 10587 Berlin, Germany
{droese,clemens,sikora}@nue.tu-berlin.de

ABSTRACT

An extension of single-view scalable video coding to multi-view is presented in this paper. Scalable video coding is recently developed in the Joint Video Team of ISO/IEC MPEG and ITU-T VCEG named Joint Scalable Video Model. The model includes temporal, spatial and quality scalability enhancing a H.264/AVC base layer. To remove redundancy between views a hierarchical decomposition in a similar way to the temporal direction is applied. The codec is based on this technology and supports open-loop as well as closed-loop controlled encoding.

The advantage of this approach lies in its compatibility to the state of the art single-view video codec H.264/AVC and its simple decomposition structure. Encoding a base view using H.264/AVC syntax, any standard single-view decoder is able to decode the data. The hierarchical decomposition structure allows efficient access to all views and frames inside a view. This is especially important for video-based-rendering and multi-view displays, which have different requirements. The chosen decomposition structure also supports parallel processing.

Gain in objective as well as subjective quality was achieved for some test sequences using a single layer. The results were compared to JSVM 5.1 (simulcast).

Index Terms— multi-view video, video coding, source coding

1. INTRODUCTION

Applications like ’free viewpoint video’ or ’3-D TV’ provide a more vivid representation of dynamic scenes and enable new kind of interactivity. The user can navigate within the scene or the viewpoint can be changed which might be necessary to fulfill display requirements, e.g. multi-view displays [1], [2].

A 3-D scene can be represented by a geometric model, texture information and light conditions. To reach photorealistic quality very complex models are needed, especially for dynamic surfaces like clothes, hair etc. Even for well defined surfaces, the acquisition of geometric information might be problematic. Besides the model based approach, a 3-D scene is described by its distribution of light, the so called light field [3], which uses an approximation of the plenoptic function [4]. Light fields can easily be captured by recording the scene with several cameras from different positions. Given such a set of 2-D pictures different views can be generated by image based rendering (IBR) [5], [6]. Due to the huge amount of data, compression is crucial to handle light fields [7]. An overview of existing light field compression schemes can be found in [8] and [9].

Multi-view video (MVV) is synchronously captured by multiple cameras, which share more or less information depending on the setup and the content of the scene. Compared to light fields the baseline, i.e., the distance between cameras is rather small. The data rate scales with the number of cameras and can easily be reduced by independent coding for each camera using standard video codecs (simulcast).

The major drawback of simulcast is, that the correlation between the views is not utilised. From intuition it follows, that this redundancy can be reduced by analysing the correlation among views and predictive coding, which is similar to motion compensated coding, where a temporal displacement is estimated for prediction of the next frame. Different approaches have been investigated on multi-view image and video coding (MVC), e.g. [10] and [11] .

To explore the need for standardisation MPEG established an ad-hoc group (MPEG 3DAV) and issued a call for evidence, which identified further demand besides the existing MPEG-2 multi view profile. The coding scheme described in this paper is the response of TUB to a call for proposals issued by MPEG-3DAV [12] with some improvements regarding the decomposition along temporal direction.

The remainder of this paper is organised as follows. Section 2 describes the architecture of the proposed multi-view video codec and gives an overview of the used technology. In Section 3 the decomposition structure of the codec is explained in detail. Section 4 gives the number of reconstructions, which are required to decode one frame, for the worst case scenario. In Section 5, experimental results are compared to JSVM 5.1 (simulcast). Finally, Section 6 follows up with conclusions.

2. ARCHITECTURE OF THE CODEC

The proposed multi-view codec (MVC) is constructed of five main building blocks, as seen in Figure 1. The chosen design is similar to the recently proposed scalable video model [13].

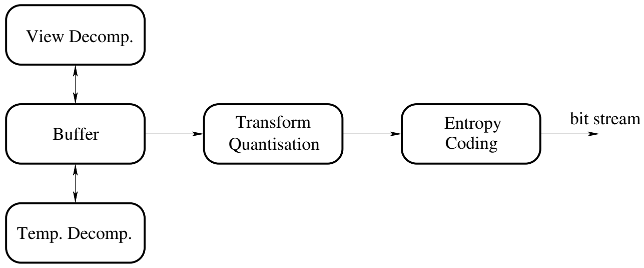

Fig. 1. System overview of presented multi-view video codec.

The multi-view input frames are stored in a predefined buffer. The size of the buffer is specified by the length of the GOP and the number of views. When the buffer is filled, the views are decomposed using disparity compensated filtering, similar to motion compensated temporal filtering (MCTF). This is performed for the first frames of the current group of pictures (GOP) and the first frame of the following GOP, due to an Open-GOP scheme in temporal direction. The remaining frames are decomposed in temporal direction. This leaves one I frame and several P and B frames, which are transformed, quantised, entropy coded and written in a given order to a bit stream.

Due to its relation to the scalable video model, spatial scalability is also supported. The MVC can be run in an open-loop fashion, as well as, in a closed-loop fashion. Furthermore, the open-loop model supports fine granular scalability (FGS) as specified in the SVM 3.0 [13]. The rate-constrained displacement estimation for compensation and predictive coding uses the fast motion search algorithm of the SVM 3.0 software.

3. DECOMPOSITION STRUCTURE FOR CODING

This design of the codec focuses on the decomposition structure. For different requirements, miscellaneous structures and schemes are possible. In this case, view scalability and efficient decoding, in terms of a small number reconstruction to access any frame inside the structure, are the desired requirements. This section gives a detailed description of the chosen mechanism.

Usually, temporal correlation is much higher than correlation inbetween views. Therefore, the views are only further decomposed at I frames of the temporal decomposition. Using the motion compensated lifting framework, also known as motion compensated temporal filtering (MCTF), the energy is concentrated in one low-pass frame, as seen in Figure 2 for single view coding. B1 depicts the first temporal decomposition level. The figure shows the decomposition for an open-loop model. If a closed-loop model is used the decomposition starts at the highest level, as the reconstruction is required for the prediction. The quantisation parameter for each frame is calculated as specified in the scalable video model [13].

For multi-view video, the idea of the single view scheme of Figure 2 is adapted. Figure 3 shows an example for four views and eight frames in a GOP $N_{GOP} = 8$. View decomposition is only applied to the low-pass frames of each view. As an Open-GOP cannot be applied, the views are decomposed using a different prediction scheme, which is explained in the following.

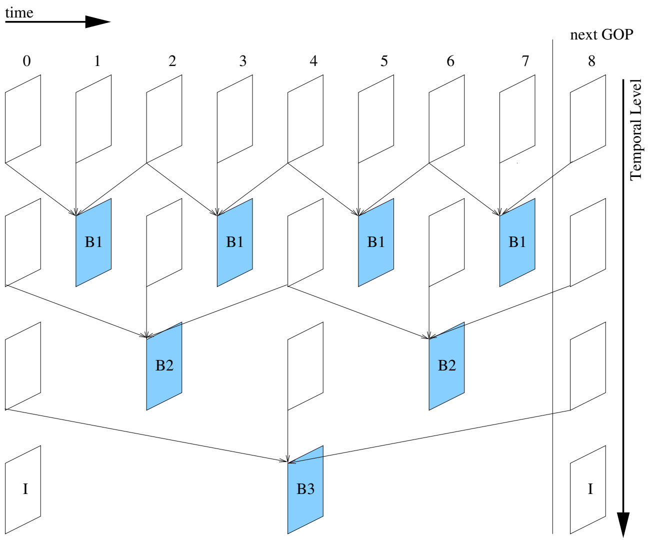

Fig. 2. Temporal decomposition for a single-view for the open-loop case.

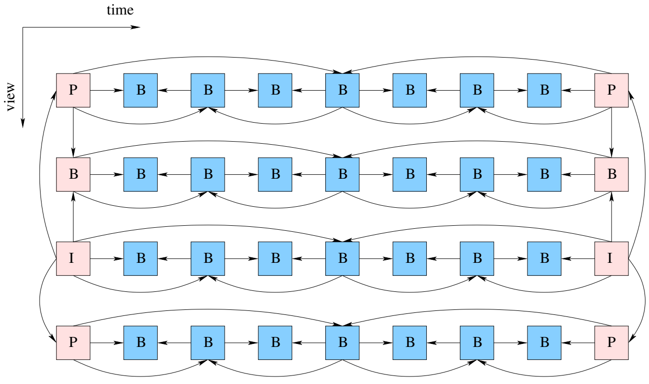

Fig. 3. Multi-view decomposition structure for the presented codec.

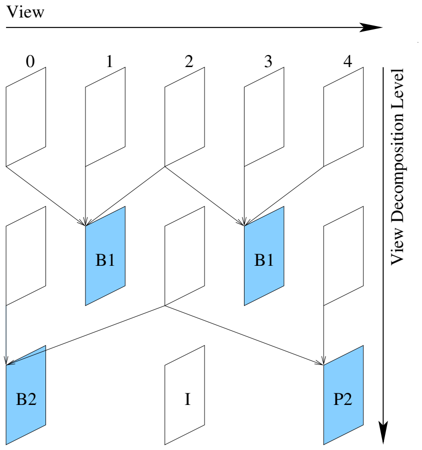

Fig. 4. Prediction scheme for the decomposition of five views.

The position of the I frame depends on the number of views N_{Views}. This can be calculated using the following equation.

$$I_{Pos} = \sum_{i=0}^{M} p(i)$$

with

$$p(i) = \begin{cases} 0 & \text{if } i = 0 \\ 0 & \text{if i is odd} \\ 2^{i-1} & \text{if i is even} \end{cases}$$

$$M = \lceil log_2(N_{Views}) \rceil$$

Depending on the position of the I frame, the prediction scheme is determined. For five views the position of the I frame is at frame two, if the first frame starts with zero, as illustrated in Figure 4. This means, the I frame is shifted further into the center and the energy is more equally distributed over the decomposed frames.

The decomposed frames are transformed, quantised and entropy coded. For the entropy coding, context adaptive binary arithmetic coding is applied, as specified in H.264/AVC [14]. The coded frames are written in a given order to the bit stream. The order is determined by the decomposition level, starting with the I frame and follows up with the P and B frames of the highest view level, and so on. The temporal decomposed frames are added to the bit stream in a hierarchical order, shown in Figure 5.

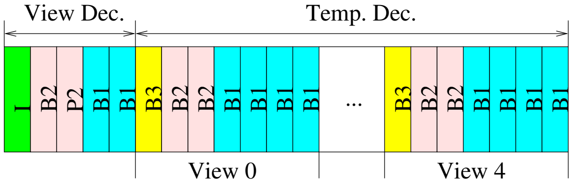

Fig. 5. Bit stream for five views and $N_{GOP} = 8$.

4. DECODING OF A PARTICULAR FRAME

To access any frame in any view, the maximal required number of reconstruction steps depends on the number of views N_{Views} and the number of frames in a GOP N_{GOP}.

$$N_{Rec} = N_{Rec.View} + N_{Rec.Temp.}$$

$$N_{Rec.View} = \begin{cases} M & \text{if } N_{Views} <= \frac{3}{4} 2^M \\ M+1 & \text{if } N_{Views} > \frac{3}{4} 2^M \end{cases}$$

$$N_{Rec.Temp.} = \lceil log_2(N_{GOP}) \rceil$$

5. EXPERIMENTAL RESULTS

Experiments have been carried out for two multi-view sequences using 8 cameras, which have been provided from KDDI and MERL for the Call for Proposal at MPEG on multi-view video coding. The spatial resolution of both sequences used in the experiments is 640x480. 320 and 240 frames, respectively, have been encoded for the experiments. For comparison, the views have been encoded independently using state of the art JSVM 5.1 (simulcast). The size of GOP has been chosen in all experiments to be 16.

	JSVM		MVC	
View	Rate	PSNR	Rate	PSNR
0	227.5	33.0	200.0	33.9
1	217.3	33.4	200.0	33.8
2	218.4	33.1	200.0	34.3
3	212.4	33.5	200.0	33.5
4	179.6	34.4	200.0	34.3
5	177.2	34.6	200.0	34.3
6	230.0	32.8	200.0	33.5
7	214.4	33.3	200.0	34.1
Avg.	209.6	33.5	200.0	34.0

Table 1. Sequence Race1 320 frames encoded using JSVM and the presented MVC (PSNR [dB], Rate[kbit/s]).

The presented MVC is compared against JSVM 5.1 (simulcast). For MVC and JSVM a search range of 96 pels for motion estimation, FRExt, Loop Filter and CABAC have been used in all experiments. For disparity estimation, the search range was set to 96 pels as well. The closed-loop model is applied for the experiments. The chosen sequences have different characteristics. Sequence ’Race1’ contains a lot of motion, due to a camera pan from left to right. The cameras are arranged in a linear setup. Sequence ’Ballroom’ does not contain any camera movement. Instead it contains motion of some dancers and strong reflections on the floor. The cameras are also linearly aligned.

Some gain in quality, up to 0.5 dB, was achieved for both sequences, as tabulated in Table 1 and shown in Figure 6

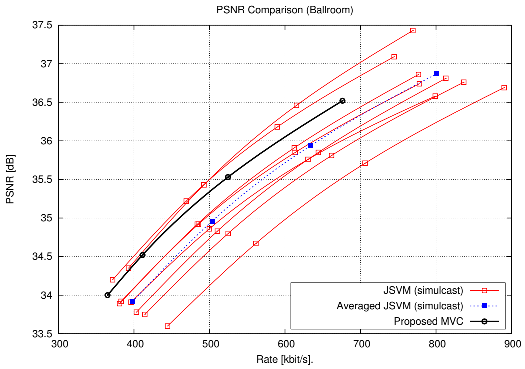

Fig. 6. PSNR comparison of JSVM (simulcast) to the proposed MVC for sequence Ballroom (240 frames).

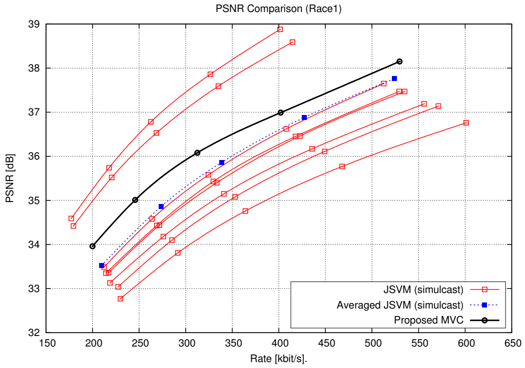

Fig. 7. PSNR comparison of JSVM (simulcast) to the proposed MVC for sequence Race1.

and 7. Subjective tests were also showing improvement compared to JSVM.

6. CONCLUSIONS

In this paper a new multi-view video codec is proposed based on the scalable video model SVM 3.0. It extents the idea of hierarchical decomposition from single-view video to multi-view video, giving full view and temporal scalability. The codec can be run in an open-loop fashion or closed-loop fashion, whereas the open-loop model supports quality scalability.

First experiments using a single-layer, with no spatial scalability and quality scalability, show some gain in subjective and objective quality for certain sequences compared to the state of the art codec JSVM 5.1. Future work will concentrate on inter-view prediction to further improve coding efficiency.

7. ACKNOWLEDGEMENTS

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

8. REFERENCES

[1] N.A. Dodgson, “Autostereoscopic 3d displays,” *Computer Magazine*, vol. 38, no. 8, pp. 31–36, 2005.

[2] M. Wojciech and H. Pfister, “3D TV: A scalable system for real-time acquisition, transmission and autostereoscopic display of dynamic scenes,” Technical report, Mitsubishi Research Labs, 2004.

[3] A. Gershun, “The light field,” *J. Math Phys.*, vol. 18, pp. 51–151, 1939.

[4] E. H. Adelson and J. R. Bergen, “The plenoptic function and the elements of early vision,” *M. Landy and J. A. Movshon, (eds) Computational Models of Visual Processing*, 1991.

[5] M. Levoy and P. Hanrahan, “Light field rendering,” in *Proc. of SIGGRAPH’96*, Aug 1996, vol. 18, pp. 31–42.

[6] Michael Droese, Toshiaki Fujii, and Masayuki Tanimoto, “Ray-space interpolation based on filtering in disparity domain,” in *Proceedings of 3D Image Conference*, Tokyo, Japan, 2004.

[7] Marcus Magnor and Bernd Girod, “Data compression for light-field rendering,” in *IEEE Trans. on Circuits and Systems for Video Technology*, Apr. 2000, vol. 10, pp. 338–343.

[8] M. Tanimoto and T. Fujii, “Comparative evaluation of ray-space representation,” vol. M8892. ISO/IEC JTC1/SC29/WG11, 2002.

[9] Cha Zhang and Tsuan Chen, “A survey on image-based rendering - representation, sampling and compression,” Technical report, Carnegie Mellon University, 2003.

[10] Ru-Shang Wang and Yao Wang, “Multiview video sequence analysis, compression, and virtual viewpoint synthesis,” in *IEEE Trans. on Circuits and Systems for Video Technology*, Apr. 2000, vol. 10, pp. 397–410.

[11] H. Pfister A. Vetro, W. Matusik and Jun Xin, “Coding approaches for end-to-end 3d tv systems,” Technical report, Mitsubishi Research Labs, 2004.

[12] Motion Picture Expert Group (MPEG), “Call for proposals on multi-view video coding,” vol. N7327. ISO/IEC JTC1/SC29/WG11, 2005.

[13] Julien Reichel, Mathias Wien, and Heiko Schwarz, “Scalable video model 3.0,” vol. N6716. Motion Picture Expert Group (MPEG), 2004.

[14] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, “Advanced video coding for generic audiovisual services,” 2005.