EFFICIENT COMPRESSION OF MULTI-VIEW VIDEO EXPLOITING INTER-VIEW DEPENDENCIES BASED ON H.264/MPEG4-AVC

P. Merkle, K. Müller, A. Smolic, and T. Wiegand

Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut
Image Processing Department
Einsteinufer 37, 10587 Berlin, Germany
{merkle/kmueller/smolic/wiegand}@hhi.de

ABSTRACT

Efficient Multi-view coding requires coding algorithms that exploit temporal, as well as inter-view dependencies between adjacent cameras. Based on a spatiotemporal analysis on the multi-view data set, we present a coding scheme utilizing an H.264/MPEG4-AVC codec. To handle the specific requirements of multi-view datasets, namely temporal and inter-view correlation, two main features of the coder are used: hierarchical B pictures for temporal dependencies and an adapted prediction scheme to exploit inter-view dependencies. Both features are set up in the H.264/MPEG4-AVC configuration file, such that coding and decoding is purely based on standardized software. Additionally, picture reordering before coding to optimize coding efficiency and inverse reordering after decoding to obtain individual views are applied. Finally, coding results are shown for the proposed multi-view coder and compared to simulcast anchor and simulcast hierarchical B picture coding.

1. INTRODUCTION

3D and free viewpoint video are new types of natural video media that expand the user's sensation far beyond what is offered by traditional media. The first offers a 3D depth impression of the observed scenery, while the second allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics applications [1]. Target applications include broadcast television and other forms of video entertainment, as well as surveillance. These applications are enabled through convergence of technologies from computer graphics, computer vision, multimedia and related fields, and rapid progress in research covering the whole processing chain from capturing, signal processing, data representation, compression, transmission, display and interaction. Some of these application scenarios may be based on proprietary systems, as for instance already employed for (post-) production of movies and TV content. On the other hand there are also application scenarios that require interoperable systems, such as 3DTV broadcast or free viewpoint video on DVD.

To ensure interoperability between different systems, standardized formats for data representation and compression are necessary; these interchangeable formats are typically specified by international standardization bodies such as the ITU-T Video Coding Experts Group or the ISO/IEC JTC 1 Moving Pictures Experts Group (MPEG). In recent years, the MPEG committee has been investigating the needs for standardization in the area of 3D and free viewpoint video in a group called 3DAV (3D audio-visual) [7]. Thus far, the committee has provided an overview of relevant technologies and has shown that a number of these technologies are already supported by existing standards such as MPEG-4 [8], [9]. For the missing elements, new standardization activities have been launched. Some activities have already been completed, such as the new tools for the efficient and high-quality representation of 3D video objects, which have been adopted as part of the MPEG-4 Animation Framework eXtension (AFX) specification [10].

A common element of many systems described above is the use of multiple views of the same scene that have to be transmitted to the user. The straight-forward solution for this would be to encode all the video signals independently using a state-of-the-art video codec such as H.264/MPEG4-AVC [11]. However, in a "Call for Evidence" [4] it has been shown that specific multi-view video coding (MVC) algorithms give significantly better results compared to the H.264/MPEG4-AVC simulcast solution [5]. The basic idea in all of the submitted proposals is to exploit inter-view and temporal statistical dependencies for compression. Since all cameras capture the same scene from different viewpoints, inter-view statistical dependencies can be expected [3].

To investigate multi-view coding (MVC) technology in-depth, MPEG decided to issue a "Call for Proposals" (CfP) [6] for MVC technology along with related requirements [2]. This paper describes our multi-view coding proposal within this call. The next section shows the statistical analysis that was carried out to design the multi-view coding structure, which is itself described in section 3. The coding results for the multi-view video test set are presented in section 4.

2. TEMPORAL/INTER-VIEW CORRELATION

The main consideration regarding efficiency of MVC is the efficiency gain of inter-view/temporal prediction versus classical temporal prediction. If there is no gain, MVC will not perform better than H.264/MPEG4-AVC simulcast. We therefore set up initial experiments to simulate inter-view/temporal prediction with MVC and performed a statistical analysis to answer the question if it is useful and in which cases a gain can be expected.

For the case of linear camera settings, the inter-view/temporal first order neighbors are shown in Fig. 1. With the exception of left- and rightmost cameras each picture of the multi-view sequence has 8 inter-view/temporal neighbors.

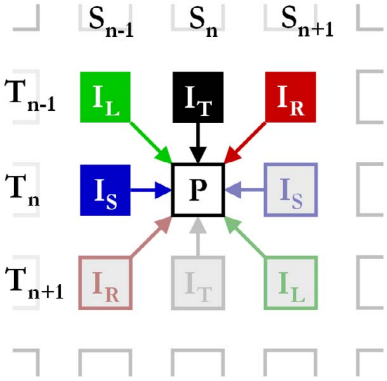

Fig. 1: Prediction modes for first-order neighbor images.

Fig. 1 depicts the resulting prediction modes we used for evaluation, which can be classified into the four color filled basic prediction modes and their four grey shaded symmetric equivalents. In opposite to the basic modes, which use preceding pictures, their equivalents use subsequent pictures for prediction.

The results are shown in the bar graphs of Fig. 2 for the two sequences "Uli" and "Breakdancers" by means of the likelihood of prediction mode selection. Here the prediction mode is chosen with the lowest Lagrangian cost value for Lagrangian motion estimation as described in [12] and repeated here for completeness. Lagrangian motion estimation determines the motion vector $\boldsymbol{m}_i$ for block $\boldsymbol{S}_i$ by

$$\boldsymbol{m}_i = \arg\min_{\boldsymbol{m}\in\boldsymbol{M}} \{D_{DFD}(\boldsymbol{S}_i,\boldsymbol{m}) + \lambda_{MOTION} R_{MOTION}(\boldsymbol{S}_i,\boldsymbol{m})\}$$

where $\boldsymbol{M}$ is the set of possible motion vectors and with the distortion term being given by

$$D_{DFD}(\boldsymbol{S}_i,\boldsymbol{m}) = \sum_{(x,y)\in\boldsymbol{A}_i} \left| s[x,y,t] - s'[x-m_x, y-m_y, t-m_t] \right|^p$$

with p=2 for sum of squared errors and s[] being the current and s'[] being a previously decoded picture that is referenced using the picture reference index m_t. $R_{MOTION}(\boldsymbol{S}_i,\boldsymbol{m})$ is the number of bits to transmit all components of the motion vector (m_x, m_y, m_t). The size of the blocks $\boldsymbol{S}_i$ in the experiment was 16x16 and the Lagrange parameter λ_{MOTION} was chosen to be 29.5. The search range $|\boldsymbol{M}|$ is ±32 integer pixel positions horizontally and vertically.

For the described statistical analysis, the multi-view pictures were coded along all cameras for each time point, by encoding pairs of pictures, consisting of a mode-dependent I and corresponding P picture, for each of the four basic prediction modes, according to Fig. 1. The results that are shown for the two sequences do not change significantly in case of an increased search range or variation of λ_{MOTION}.

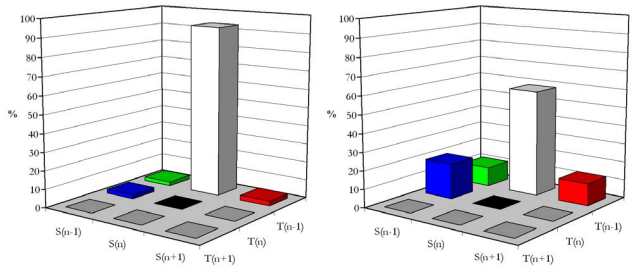

Fig. 2: Probability of chosen predictor when minimizing a Lagrangian cost function in motion estimation for sequences "Uli" and "Breakdancers".

The first conclusion drawn from the analysis over a larger set of multi-view sequences is that temporal prediction is always the most efficient prediction mode. Comparison between inter-view and inter-view/temporal prediction modes shows that inter-view is more efficient than mixed modes at large. However, there are significant differences between the test data sets, regarding the relation between temporal and inter-view prediction.

3. MULTI-VIEW CODING STRUCTURE

Based on the statistical analysis of the above inter-view/temporal prediction, an MVC scheme was set-up, as shown in Fig. 3. This scheme uses the prediction structure of hierarchical B pictures for each view [13]. Hierarchical B pictures provide significantly improved RD performance when the quantization parameters for the various pictures are assigned appropriately [13]. Additionally, inter-view prediction is applied to every 2nd view, i.e. S1, S3 and S5 in Fig. 3. For an even number of views, the last view (S7) is coded as shown, starting with a P picture, followed by hierarchical B-pictures, which are also inter-view predicted from the previous view. Thus, the coding scheme can be applied to any multi-view setting with *number_of_views* ≥ 2. To allow synchronization, pictures start each GOP (S0/T0, S0/T8, etc.). In Fig. 3 and Fig. 4, a GOP length of 8 is shown for illustration purposes, while GOP lengths of 12 and 15 were used in the experiments in order to provide a fair comparison to anchor coding results.

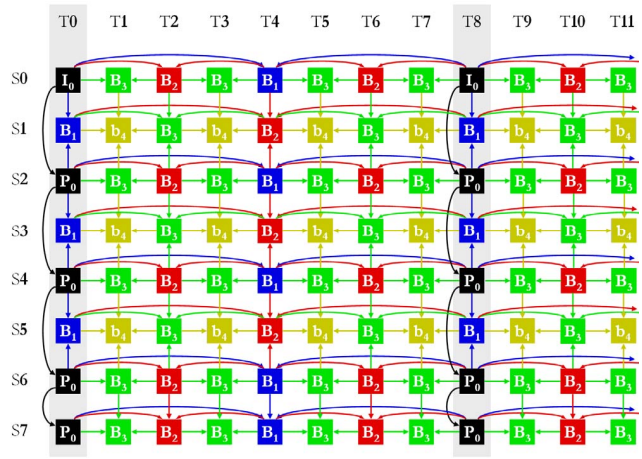

Fig. 3: Spatio-temporal prediction structure based on H.264/MPEG4-AVC hierarchical B pictures.

The used H.264/MPEG4-AVC video encoder utilizes the Lagrangian coder control as described in [12] together with the cascading of QP values as described in [13]. The only change that was applied to the H.264/MPEG4-AVC-coder was the increase of the Decoded Picture Buffer size to 2**GOP_length + number_of_views* to handle the proposed scheme.

The coding scheme itself is specified by configuring our H.264/MPEG4-AVC video codec, where for each picture the level for QP cascading, reference pictures for prediction and memory management commands are set [11]. To allow efficient memory management, picture reordering is applied and reflected in the picture order count values of each picture that are shown in Fig. 4.

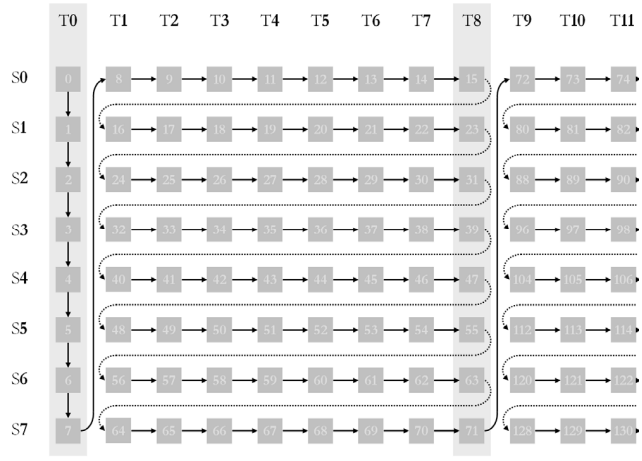

Fig. 4: Memory-efficient Reordering of MV input for compression with H.264/MPEG4-AVC.

4. CODING RESULTS

All multi-view test sequences with 5 to 16 camera views have been evaluated with the developed coding scheme and compared against anchor coding results in terms of PSNR vs. bitrate. The anchor as provided by MPEG was coded with H.264/MPEG4-AVC simulcast and an IBBPBBP temporal decomposition. The best results that we have obtained are shown in Fig. 6 and Fig. 7, where anchor coding results are represented by the black curve. The results produced by our MVC coding scheme, utilizing hierarchical B pictures together with inter-view and temporal dependencies, are shown by the red curve. Additionally, simulcast coding using hierarchical B pictures is represented by the blue curve.

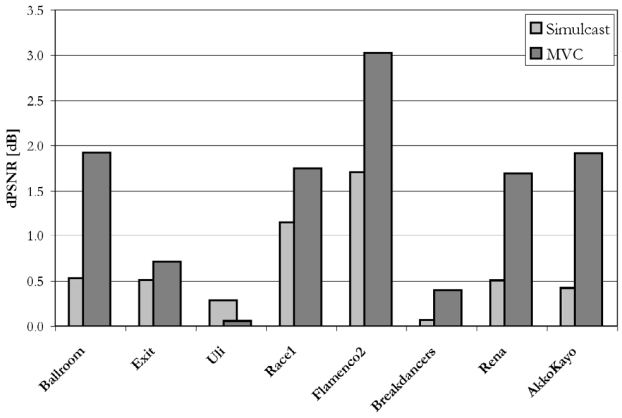

Fig. 5: Average PSNR gains obtained for test sequences

Fig. 5 depicts the average gains in comparison to anchor coding results. Depending on the specific sequence, coding improvements up to 3.2 dB are obtained (red vs. black curve in Fig. 6 and Fig. 7). The gain strongly depends on the original setting of the multi-camera arrangement, namely the temporal and inter-view correlation. As pointed out by the blue curves in Fig. 6 and Fig. 7, a good portion of the coding gain is already provided by using hierarchical B pictures in simulcast.

The quality distribution among the views is sequence dependent. For equal QP-setting across all views, sequences with larger camera distance and higher scene complexity show larger deviations, e.g. the Race1 sequence, while sequences like Ballroom with very small camera distance have only small deviations due to more similar content across all views.

The grey curves in Fig. 6 and Fig. 7 additionally depict the results of the other proposed multi-view coding methods of the Call for Proposal. Here, our presented MVC coding scheme outperforms the other approaches for the majority of tested sequences.

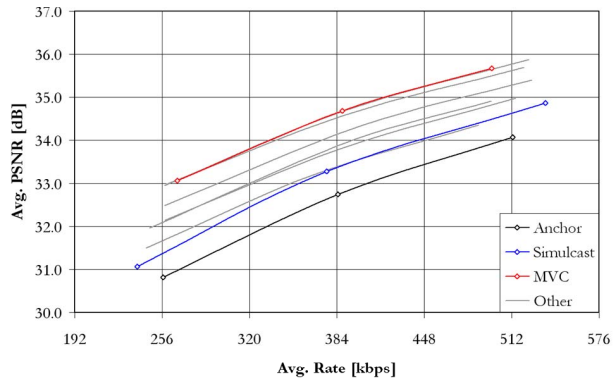

Fig. 6: Coding results for Ballroom-Sequence

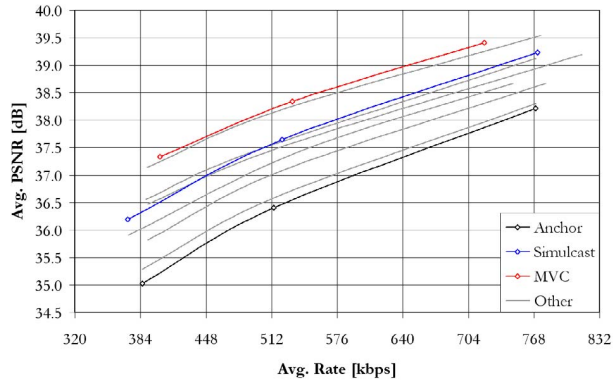

Fig. 7: Coding results for Race1-Sequence

5. CONCLUSIONS

In this paper we have presented a coding scheme for multi-view camera sequences, which is based on H.264/MPEG4-AVC. First, an analysis on temporal and inter-view dependencies was carried out for the different multi-view test sets. Based on the least RD-cost value, the percentage of macroblocks chosen for prediction from all spatiotemporally adjacent pictures was recorded for each sequence. Here, mainly temporal neighbors are selected, although also a certain percentage of inter-view and mixed inter-view/temporal neighbors are selected for prediction. The predictor selection statistics strongly depend on the sequence content and multi-view setup.

Based on this analysis, a coding scheme was developed, that uses hierarchical B pictures to exploit temporal dependencies of each view, as well as inter-view dependencies between neighboring views. The obtained results show that the proposed coding structure performs up to 3.2 dB better than simulcast anchor coding, depending on the multi-view setting. Additionally, simulcast coding with hierarchical B pictures was investigated, showing that roughly half of the coding gain is already obtained by this feature.

6. ACKNOWLEDEMENTS

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

7. REFERENCES

[1] A. Smolic, and P. Kauff, “Interactive 3D Video Representation and Coding Technologies”, Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery, vol. 93, no. 1, Jan. 2005

[2] ISO/IEC JTC1/SC29/WG11, “Requirements on Multi-view Video Coding v.2”, Doc. N7282, Poznan, Poland, July 2005.

[3] ISO/IEC JTC1/SC29/WG11, “Survey of Algorithms used for Multi-view Video Coding (MVC)”, Doc. N6909, Hong Kong, China, January 2005.

[4] ISO/IEC JTC1/SC29/WG11, “Call for Evidence on Multi-View Video Coding”, Doc. N6720, Palma de Mallorca, Spain, October 2004.

[5] ISO/IEC JTC1/SC29/WG11, “Report of the subjective quality evaluation for MVC Call for Evidence”, Doc. N6999, Hong Kong, China, January 2005.

[6] ISO/IEC JTC1/SC29/WG11, “Updated Call for Proposal on Multi-view Video Coding”, Doc. N7567, Nice, France, October 2005.

[7] A. Smolic, and D. McCutchen, “3DAV Exploration of Video-Based Rendering Technology in MPEG”, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No. 3, pp. 348-356, March 2004.

[8] ISO/IEC JTC1/SC29/WG11, “Applications and Requirements for 3DAV”, Doc. N5877, Trondheim, Norway, July 2003.

[9] ISO/IEC JTC1/SC29/WG11, “Report on 3DAV Exploration”, Doc. N5878, Trondheim, Norway, July 2003.

[10] ISO/IEC JTC1/SC29/WG11, “ISO/IEC 14496-16/PDAM1”, Doc. N6544, Redmont, WA, USA, July 2004.

[11] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, “Advanced Video Coding for Generic Audio-Visual Services”, 2003.

[12] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, “Rate-Constrained Coder Control and Comparison of Video Coding Standards,” IEEE Trans. CSVT, vol. 13, pp. 688-703, July 2003.

[13] H. Schwarz, D. Marpe, and T. Wiegand: "Hierarchical B pictures," Joint Video Team, Doc. JVT-P014, Poznan, Poland, July 2005.