CODING EFFICIENCY AND COMPLEXITY ANALYSIS OF MVC PREDICTION STRUCTURES

Philipp Merkle, Aljoscha Smolic, Karsten Müller, and Thomas Wiegand

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

ABSTRACT

Based on the idea to exploit the statistical dependencies from both temporal and inter-view reference pictures for motion compensated prediction, this paper presents a systematic evaluation of multi-view video coding with optimized prediction structures. The compression method is based on the multiple reference picture technique in the H.264/AVC video coding standard. The advantages of hierarchical B pictures for temporal prediction are combined with inter-view prediction for different temporal hierarchy levels, starting from simulcast coding with no inter-view prediction up to full level inter-view prediction. When using inter-view prediction at key picture temporal level average gains of 1.4 dB PSNR are reported, while additionally using inter-view prediction at non-key picture temporal levels average gains of 1.6 dB PSNR are reported. For some cases gains of more than 3 dB, corresponding to bit rate savings of up to 50%, are obtained.

1. INTRODUCTION

The development of new applications for natural video scenes is one of the most promising fields for the employment of 3D techniques. Rising interest in 3D television (3DTV) and free viewpoint video (FVV) lead to the promotion of these types of new media [1][2]. While FVV allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics, 3DTV offers three-dimensional depth impression of the observed scenery. Both technologies do not exclude each other and can be combined within a single system. Such applications are enabled through convergence of technologies from computer graphics, computer vision, multimedia and related fields on one hand and by research and development, regarding the complete processing chain from capturing, representation, compression, transmission to interactive presentation on the other.

Efficient compression techniques are essential for realizing such applications, because most of them employ multiple camera views of the same scene, often referred to as multi-view video (MVV). Basically MVV creates large amounts of data to be stored or transmitted, but contains a large amount of inter-view statistical dependencies, since all cameras capture the same scene from different viewpoints. These can be exploited for combined temporal/inter-view prediction,

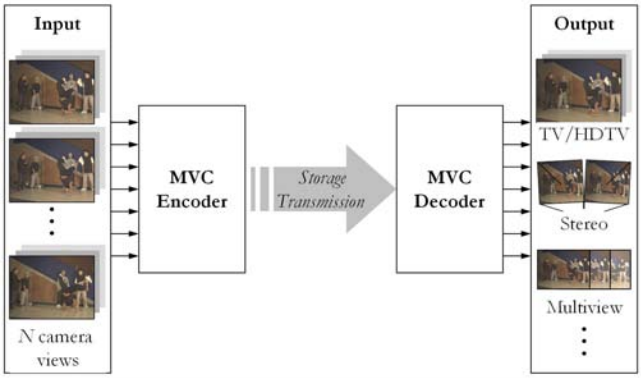

Figure 1 – Overall structure of a MVC system

where images are not only predicted from temporally neighboring images but also from corresponding images in adjacent views [3]-[6], referred to as multi-view video coding (MVC). The overall structure of MVC defining the interfaces is illustrated in Fig. 1. Basically the multi-view encoder receives *N* temporally synchronized video streams and generates one bit-stream. The multi-view decoder receives the bit-stream, decodes and outputs the *N* video signals.

Various researchers have reported their results in the field of multi-view video coding. Besides approaches for efficient compression, appropriate image correction methods [7] as well as an analysis of potential gains from combined temporal/inter-view prediction [8] have been presented. Further, scene geometry can be exploited to improve compression efficiency. Based on disparity or depth estimation view-interpolation or 3D warping can be performed as additional source for inter-view prediction [9][10]. 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 ISO/IEC JTC 1 Moving Picture Experts Group (MPEG) [11]. Currently a new MPEG standard for MVC is developed, which is scheduled to be finalized in early 2008.

2. PREDICTION STRUCTURES

In this section the configurations, properties and features of the different developed prediction structures for MVC are presented, starting from temporal prediction up to inter-view prediction over the complete multi-view sequence.

The main goal of MVC is to provide significantly increased compression efficiency compared to individually encoding all video signals. Therefore encoding all views using H.264/AVC [12] with the same test conditions was considered as the reference (anchors) for coding performance comparison. Encoding was done using typical settings and parameters with an *IBBP*... picture coding structure and the resulting decoded video signals serve as reference for objective and subjective evaluations.

2.1 Temporal Prediction

Encoding and decoding each view of a multi-view test data set separately can be done with any existing standard-conforming H.264/AVC codec. This would be a simple, but inefficient way to compress multi-view video sequences, due to not exploiting the inter-view statistical dependencies.

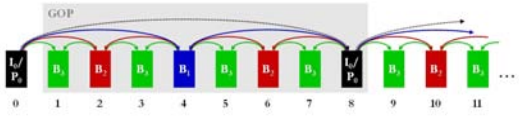

Figure 2 – Hierarchical reference picture structure for temporal prediction

Since the *IBBP*... structure used for anchor coding is not the most efficient temporal prediction structure possible with H.264/AVC, this section introduces the concept of *hierarchical B pictures* (see [13] for a detailed description). These types of prediction schemes benefit from the increased flexibility of H.264/AVC at picture/sequence level in comparison to former video coding standards through the availability of the multiple reference picture technique. A typical hierarchical prediction structure with three dyadic hierarchy stages is depicted in Fig. 2. The first picture of a video sequence is intra-coded as IDR picture and so-called key pictures (black in Fig. 2) are coded in regular intervals. A key picture and all pictures that are temporally located between the key picture and the previous key picture are considered to build a group of pictures (GOP), as illustrated in Fig. 2 for a GOP of eight pictures length.

The concept of hierarchical B pictures can easily be applied to multi-view video sequences as illustrated in Fig. 3 for a sequence with eight cameras and a GOP length of 8, where S_n denotes the individual view sequences and T_n the consecutive time-points. To allow synchronization and random access all the key pictures are coded in intra mode.

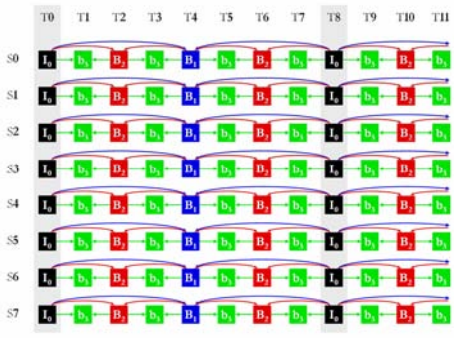

Figure 3 – Temporal prediction using hierarchical B pictures

Simulcast coding with hierarchical B pictures will be used as a reference to compare highly efficient temporal prediction structures with prediction structures that additionally use inter-view prediction.

2.2 Inter-view Prediction for Key Pictures

A universal property of video coding based on motion-compensated prediction is that coding pictures in *intra* mode, where no reference pictures are available for prediction, results in considerable higher bit rates than in *inter* prediction [12]. Consequently replacing intra-coded I pictures with inter-coded P or B pictures has the potential to achieve a substantial coding gain.

Adapting this approach to the multi-view video example of Fig. 3 leads to the prediction scheme in Fig. 4. The prediction structure of the first view *S0* remains to be temporal prediction and is called base view, as it is identical to the simulcast prediction structure with hierarchical B pictures for temporal prediction only. However, for the other views all intra-coded key pictures are replaced by inter-coded pictures using inter-view prediction. For the remaining pictures of each GOP the prediction structure does not change and remains temporal prediction with hierarchical B pictures. Furthermore synchronization and random access features are provided by still coding the key pictures of the base view in intra mode.

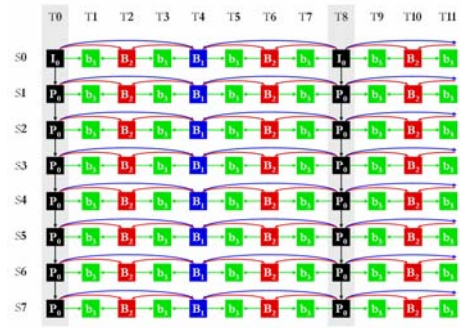

Figure 4 – IPP inter-view prediction for key pictures

Introducing this prediction scheme has a fundamental effect on the encoding and decoding process. As a consequence of using inter-view prediction the video sequences of individual views S_n can not be processed independently any more as they share reference pictures and rather have to be either interleaved into one bit-stream for sequential processing or signaled and stored in a shared buffer for parallel processing.

Fig. 5 presents alternative coding structures for multi-view video data with inter-view prediction for key pictures. It depicts the inter-view reference frame selection at the temporal level of key pictures, and again for a data set with eight linearly arranged camera views. The left figure represents the prediction structure of Fig. 4 which will be referred to as *KS_IPP* prediction mode. Since the base view position is not necessarily determined to be the first view, the middle figure illustrates a variation of the upper scheme, referred to as *KS_PIP* mode. This configuration, where the base view is one of the centre views, might benefit from the fact that not one but two of the inter-view predicted key pictures directly

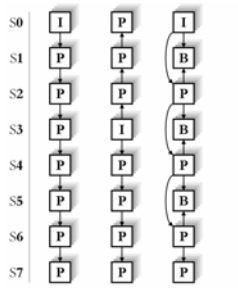

Figure 5 – Alternative structures for inter-view prediction

use the *I* picture as reference. The right figure presents a true alternative, because in addition to temporal prediction it uses B pictures also to inter-view prediction, called *KS_IBP* mode. Again, this prediction structure might have coding efficiency advantages over the other configurations, as the experimental results for hierarchical B pictures with temporal prediction indicate [13], but for the disadvantage of being more complex.

2.3 Inter-view Prediction for Non-Key Pictures

Analyses of temporal and inter-view prediction efficiency in MVV data sets indicate that using temporal and inter-view reference frames at the same time, has the potential to improve coding efficiency. In order to exploit all statistical dependencies within a multi-view test data set, inter-view prediction can be extended to non-key pictures.

Fig. 6 illustrates how the advantages of hierarchical B pictures can be combined with inter-view prediction, without any changes regarding the temporal prediction structure. Again, the example shows a multi-view sequence with eight cameras and a GOP length of 8. At key picture level and for view *S0* the prediction structure is identical to the right scheme in Fig. 5, but for all non-key pictures inter-view reference pictures are additionally used for prediction. According to the prediction structure of key pictures, prediction is extended by two inter-view reference frames for every second view. View *S7* demonstrates how inter-view prediction is realized with just one reference view, for example if using an *IPPP*... structure at key picture temporal level. In contrast to the prediction structures of Fig. 4, where the maximum number of reference pictures is two, now the non-key pictures have up to four references. Thus coding efficiency is improved, but at the cost of increased coding complexity.

Synchronization and random access features are still provided by coding the key pictures of the base view in intra mode. Furthermore backward compatibility is provided by these prediction structures, as the base view can be extracted and the resulting bit-stream is conforming to the H.264/AVC standard.

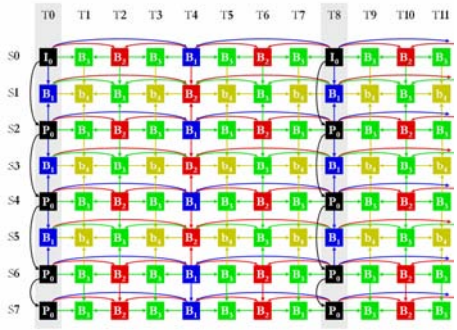

Figure 6 – Hierarchical inter-view prediction for all pictures

3. EXPERIMENTAL RESULTS

The prediction structures and coding schemes presented in this paper have been developed in the context of a MPEG/JVT standardization project for MVC, which defines most of the requirements as well as test data and evaluation conditions [14][15]. Therefore the next section explains how experiments with the presented MVC prediction structures can be implemented and how to configure them in order to achieve comparable results. In a second section the experimental results are presented and analyzed.

3.1 Setup of coding experiments and test conditions

The most important aspect regarding the coding experiments using the prediction structures presented in section 2 is that they can be carried out with a standard-conforming H.264/AVC encoder with an extended amount of memory for reference pictures. For that the multi-view video sequences are combined into one single uncompressed video stream as illustrated in Fig. 7, using a specific scan. This uncompressed video stream is used as the input of the H.264/AVC encoder software and the resulting bit-stream as the input of the decoder software respectively. Afterwards inverse reordering is applied to the decoded video stream in order to separate the individual view sequences.

The prediction structure itself is controlled by appropriate settings of the encoder's parameters for reference picture selection and memory management. Because this is pure encoder optimization, the resulting bit-stream is standard-conform and can be decoded by any standard H.264/AVC decoder. The only change this approach requires relative to a conforming H.264/AVC codec is the increase of the Decoded Picture Buffer (DPB) size to store all reference pictures necessary for prediction with the proposed structures, and a potentially larger number of output pictures per second than the currently allowed 172 frames in H.264/AVC. All the presented experiments and results are based on the eight test data sets together with the coding conditions specified by MPEG. Therefore the multi-view prediction schemes are adapted to the specific camera arrangements of the test data sets as well as the random access specifications by customizing the GOP length.

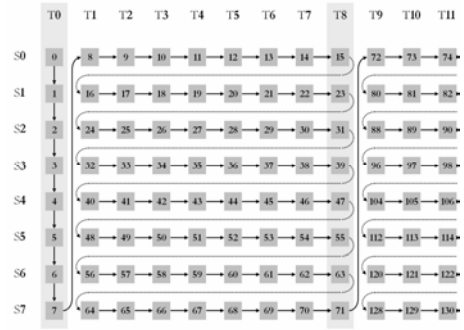

Figure 7 – Frame interleaving for compression with H.264/AVC

3.2 Objective Evaluation

Example results using the different prediction structures of section 2 are shown in Fig. 8. The PSNR-Y values are plotted over bit rate as the average over all views of a data set. In Fig. 8 *Anchor* refers to the reference *IBBP…* coding provided by MPEG, *Simulcast* to simulcast coding with hierarchical B pictures, *KS_IPP*/*KS_PIP*/*KS_IBP* to the three alternative multi-view structures for key picture inter-view prediction and *MVC* to multi-view coding with inter-view prediction for both key and non-key pictures. Apparently all those schemes using inter-view prediction outperform the ones using no inter-view prediction. However a good portion of the gain already originates from temporal prediction with hierarchical B pictures, the results show that exploiting inter-view statistical dependencies by multi-view prediction structures significantly improves compression performance for these two multi-view sequences.

In order to sum up the objective results of all the tested multi-view data sets, Fig. 9 presents the average PSNR improvements between each of the proposed prediction structures and *Anchor* coding, calculated from the difference of PSNR-Y values at three fixed bit rates. Depending on the specific sequence, coding improvements up to 3.2 dB are obtained, but basically depend on the temporal and inter-view correlations. These correlations are strongly influenced by attributes like temporal and spatial density as well as scene complexity. For example for the *Uli* test sequence almost no gain has been achieved, as the inter-view statistical dependencies are limited, or the encoder is not able to exploit them, because of too large disparities between the pictures of neighboring camera views that result in extremely large motion vectors, causing high bit rates.

Although the *MVC* prediction structure achieves the highest average coding gain of 1.6 dB, the comparison to the *KS_xxx* prediction structures (1.4 dB in average) shows that additionally predicting from inter-view references for non-key pictures does not always perform better, e.g. *Exit* and *Race1* in Fig. 9. In fact, this is not so much related to whether or not employing inter-view prediction for non-key pictures, but to having an inter-view prediction structure using hierarchical B pictures at key picture temporal level. For some sequences, prediction over two views and cascading the QPs according to prediction hierarchy levels turns out to be a disadvantage and if so, always both *KS_IBP* and *MVC* perform worse than *KS_IPP*/*KS_PIP*.

Regarding the quality distribution among the individual views, basically multi-view data sets with larger camera distance and higher scene complexity, e.g. the *Race1* sequence, show larger deviations, while sequences like *Rena* with very small camera distance show small deviations due to more similar content across all the views. Besides these data set dependent aspects, the quality distribution is also affected by the inter-view prediction structure. The corresponding experiments confirm, that using an inter-view prediction structure with hierarchical B pictures at key picture temporal level results in larger deviations. In addition to that, the aspect of coding complexity should be mentioned. Unsurprisingly, for the eight tested multi-view data sets the average encoding complexity for *MVC* is almost three times higher than for those structures that omit inter-view prediction for non-key pictures, as motion-compensated prediction from inter-view reference pictures is required.

Overall the presented results clearly indicate the pros and cons of this multi-view video coding approach. By additionally using inter-view reference pictures for disparity-compensated prediction an average gain of 1.5 dB is achieved, but for some test data sets, such as the *Uli* sequence, neither of the tested prediction structures resulted in a significant coding gain. One problem with multi-view video coding is illumination and color inconsistency between the camera sequences, which affects the exploitation of inter-view statistical dependencies. Usually such effects should be minimized by proper setting of the conditions, but a MVC algorithm should be able to cope with this as well,

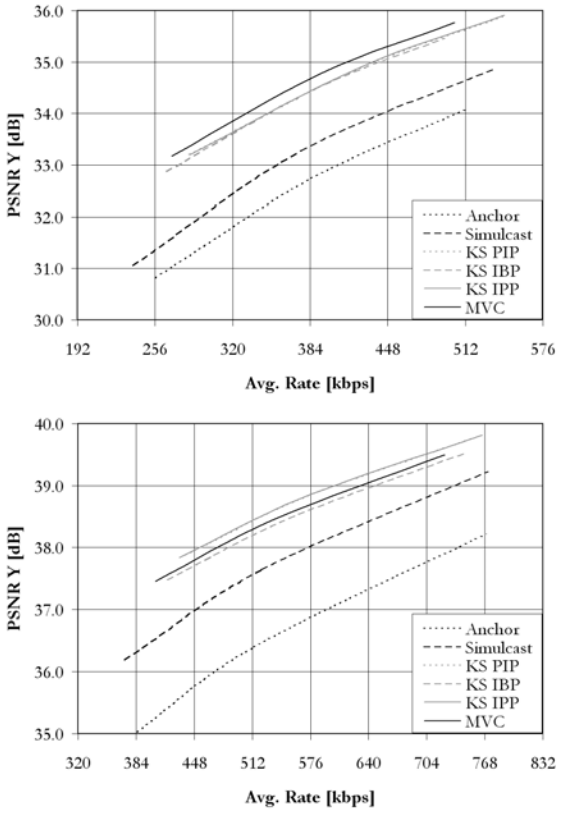

Figure 8 – PSNR results (top: *Ballroom*, bottom: *Race1* sequence)

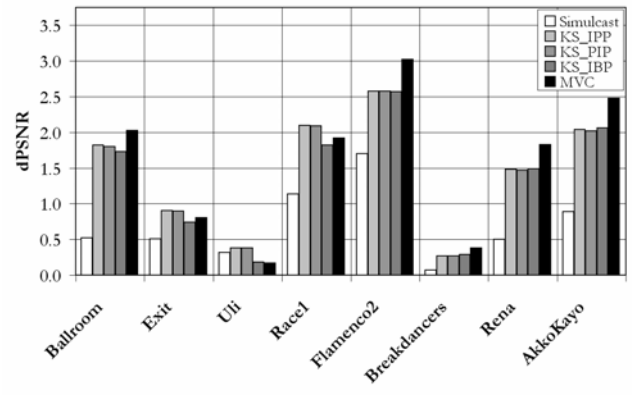

Figure 9 – Average coding gains

since perfect white-level and color balancing of the input can not be guaranteed. For MVC compensation of differences in illumination and color can be realized by modifying the prediction process of H.264/AVC on a block level [16][17], whereby additional coding gains of up to 0.6 dB are reported.

4. CONCLUSIONS

The presented prediction structures for multi-view video coding are based on the fact that multiple video streams, showing the same scene from different camera perspectives, show significant inter-view statistical dependencies. The combined temporal and inter-view prediction structures utilize inter-view prediction at different degrees, without losing the advantages of temporal prediction with hierarchical B pictures. The resulting multi-view prediction structures have the advantage of achieving significant coding gains and being highly flexible regarding their adaptation to all kinds of spatial and temporal setups at the same time. These prediction structures for multi-view video coding are very similar to H.264/AVC and require only minor syntax additions. Besides the presented sequential processing approach of the interleaved multi-view video sequences, parallel processing is supported as well. For this purpose multiple parallel encoder/decoder instances are combined in one framework that supports shared memory buffers and signaling for inter-view reference pictures.

Besides the problem with color and illumination inconsistencies the occurrence of large disparities between the different views of multi-view video sequences is an essential problem of multi-view video coding. The difficulties with large disparities concerning coding efficiency could be overcome by depth-based view interpolation prediction [9][10]. The idea is to estimate depth either at the encoder, which requires overhead for sending the depth, or the decoder(this may reduce estimation accuracy because only decoded signals are available), and to perform view interpolation or 3D warping for prediction. For example, if every other view is transmitted in a first step and depth information is available, it is possible to generate intermediate views from these data in a second step. Such an interpolated view might not be perfect for the whole image in terms of picture quality, but it might provide a useful additional source for prediction with significantly reduced disparity (ideally without any). Algorithms for both view interpolation prediction and illumination compensation are under investigation in MPEG and will most probably be included in the final MVC standard.

ACKNOWLEDGMENT

We would like to thank the Interactive Visual Media Group of Microsoft Research for providing the *Breakdancers* data set. The other test data have been provided to MPEG by Mitsubishi Electric Research Labs, KDDI Corp., Nagoya University, and Fraunhofer HHI.

This work is supported by European Commission Sixth Framework Program with grant No. 511568 (3DTV Network of Excellence Project).

REFERENCES

[1] L. Onural, A. Smolic, and T. Sikora, “An Overview of a New European Consortium: Integrated Three-Dimensional Television - Capture, Transmission and Display (3DTV)”, Proc. EWIMT04, European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, UK, November. 2004.

[2] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, “3D Video and Free Viewpoint Video – Technologies, Applications and MPEG Standards”, ICME 2006, IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006.

[3] 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.

[4] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, “Efficient Compression of Multi-view Video Exploiting Inter-view Dependencies Based on H.264/MPEG4-AVC”, ICME 2006, IEEE International Conference on Multimedia and Exposition, Toronto, Ontario, Canada, July 2006.

[5] K.-J. Oh and Y.-S. Ho, “Multi-view Video Coding based on the Lattice-like Pyramid GOP Structure”, Proc. PCS 2006, Picture Coding Symposium, Beijing, China, April 2006.

[6] X. Cheng, L. Sun, and S. Yang, “A Multi-view Video Coding Scheme Using Shared Key Frames for High Interactive Application ”, Proc. PCS 2006, Picture Coding Symposium, Beijing, China, April 2006.

[7] F. Shao, G. Jiang, M. Yu, and X. Chen, “A New Image Correction Method for Multiview Video System”, ICME 2006, IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006.

[8] A. Kaup and U. Fecker, “Analysis of Multi-Reference Block Matching for Multi-View Video Coding”, Proc. 7th Workshop Digital Broadcasting, pp. 33-39, Erlangen, Germany, Sep. 2006.

[9] E. Martinian, A. Behrens, J. Xin, and A. Vetro, “View Synthesis for Multiview Video Compression”, Proc. PCS 2006, Picture Coding Symposium, Beijing, China, April 2006.

[10] M. Kitahara, H. Kimata, S. Shimizu, K. Kamikura, Y. Yashimata, K. Yamamoto, T. Yendo, T. Fujii, and M. Tanimoto, “Multi-view Video Coding using View Interpolation and Reference Picture Selection”, ICME 2006, IEEE International Conference on Multimedia and Exposition, Toronto, Ontario, Canada, July 2006.

[11] A. Smolic, H. Kimata, and A. Vetro, “Development of MPEG Standards for 3D and Free Viewpoint Video”, Proc. SPIE Optics East, Three-Dimensional TV, Video, and Display IV, Boston, MA, USA, October 2005.

[12] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, July 2003, p. 560

[13] H. Schwarz, D. Marpe, and T. Wiegand, „Analysis of hierarchical B pictures and MCTF”, ICME 2006, IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006.

[14] ISO/IEC JTC1/SC29/WG11, “Requirements on Multi-view Video Coding v.4”, Doc. N7282, Poznan, Poland, July 2005.

[15] ISO/IEC JTC1/SC29/WG11, “Call for Proposals on Multi-view Video Coding”, Doc. N7327, Poznan, Poland, July 2005.

[16] J.-H. Kim, P.-L. Lai, A. Ortega, Y. Su, P. Yin, and C. Gomila, “Results of CE2 on Multi-view Video Coding”, Joint Video Team, Doc. JVT-T117, Klagenfurt, Austria, July 2006.

[17] Y.-L. Lee, J.-H. Hur, Y.-K. Lee, S.-H. Cho, H.J. Kwon, N.H. Hur, and J.W. Kim, “Results of CE2 on Multi-view Video Coding”, Joint Video Team, Doc. JVT-T110, Klagenfurt, Austria, July 2006.