Efficient Prediction Structures for Multi-view Video Coding

Philipp Merkle, *Member, IEEE*, Aljoscha Smolic, Karsten Müller, *Senior Member, IEEE*, and Thomas Wiegand, *Member, IEEE*

***Abstract*—An experimental analysis of multi-view video coding for various temporal and inter-view prediction structures is presented. The compression method is based on the multiple reference picture technique in the H.264/AVC video coding standard. The idea is to exploit the statistical dependencies from both temporal and inter-view reference pictures for motion-compensated prediction. The effectiveness of this approach is demonstrated by an experimental analysis of temporal versus inter-view prediction in terms of the Lagrange cost function. The results show that prediction with temporal reference pictures is highly efficient, but for 20% of a picture's blocks on average prediction with reference pictures from adjacent views is more efficient. Hierarchical B pictures are used as basic structure for temporal prediction. Their advantages 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.**

***Index Terms*—Multi-view Video Coding, 3D television, free viewpoint video, hierarchical B pictures, H.264/AVC.**

I. Introduction

THE deployment of 3D techniques is one of the most promising fields regarding the development of new applications for natural video scenes. Convergence of technologies from computer graphics, computer vision, multimedia and related fields together with rising interest in 3D television (3DTV) and free viewpoint video (FVV) lead to the promotion of these types of new media [1][2]. While 3DTV offers a 3D depth impression of the observed scenery, FVV allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics. Both technologies do not exclude each other and can be combined within a single system. Such systems are under investigation including the complete 3D processing chain, from capturing, representation, compression, transmission to interactive presentation.

Manuscript received December 15, 2006; revised July 10, 2007.

P. Merkle, A. Smolic, K. Müller, and T. Wiegand are with the Image Commmunication Group in the Image Processing Department, Fraunhofer Institute for Telecommunications–Heinrich-Hertz-Institut (HHI), 10587 Berlin, Germany (e-mail: {merkle, smolic, kmueller, wiegand}@hhi.de).

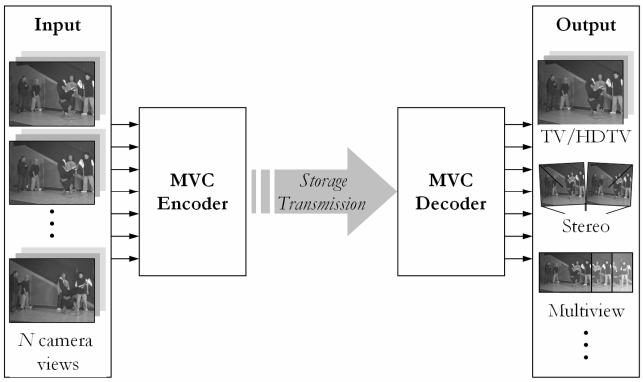

Fig. 1. Overall structure of a MVC system.

One common characteristic of many of these systems is that they use multiple camera views of the same scene, often referred to as multi-view video (MVV), which is implemented by simultaneously capturing the video streams of several cameras. Since this approach creates large amounts of data to be stored or transmitted to the user, efficient compression techniques are essential for realizing such applications. 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/AVC [3][4][5]. However, multi-view video 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, where images are not only predicted from temporally neighboring images but also from corresponding images in adjacent views, 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. A lot of approaches are based on predictive coding from reference pictures and therefore

closely related to classic video coding, but take advantage of multiple views of the same scene. This includes compression techniques [6]-[10] as well as appropriate image correction methods [11]. An analysis of potential gains from combined temporal/inter-view prediction has been presented in [12]. 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 [13][14].

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 (VCEG) or the ISO/IEC JTC 1 Moving Picture Experts Group (MPEG) [15]. In recent years, MPEG has been investigating the needs for standardization in the area of 3D and free viewpoint video in a group called 3DAV [16]. The results of an in-depth investigation of multi-view coding approaches were promising and MPEG decided to issue a "Call for Proposals" (CfP) for MVC technologies along with related requirements [17]-[22]. As a consequence of the various responses, currently a new standard for multi-view video coding is developed by the Joint Video Team (JVT) of VCEG and MPEG [23], which is scheduled to be finalized in early 2008.

This paper focuses on an experimental analysis of prediction structures for multi-view video coding. First, requirements, test data and conditions are described in section II. Section III investigates temporal versus inter-view correlations in multi-view video data. The prediction structures are presented in section IV, followed by the results of objective and subjective evaluation in section V. Finally, section VI concludes this paper.

II. Requirements and Conditions for MVC

The prediction structures and coding schemes presented in this paper have been developed and investigated in the context of the MPEG and later JVT standardization project for MVC. Therefore, most of the requirements as well as test data and evaluation conditions are defined by the MVC project [19] as presented in the next sections.

A. Requirements

The central requirement for any video coding standard is high compression efficiency. In the specific case of MVC this means a significant gain compared to independent compression of each view. Compression efficiency measures the trade-off between cost (in terms of bit rate) and benefit (in terms of video quality), i.e. the quality at a certain bit rate or the bit rate at a certain quality. However, compression efficiency is not the only factor under consideration for a video coding standard. Some requirements of a video coding standard may even be contradictory such as compression efficiency and low delay in some cases. Then a good trade-off has to be found. General requirements for video coding such as minimum resource consumption (memory, processing power), low delay, error robustness, or support of different pixel and color resolutions, are often applicable to all video coding standards [5].

Some requirements are specific to MVC as highlighted in the following. Temporal random access is a requirement for any video codec. For MVC also view random access becomes important. Both together ensure that any image can be accessed, decoded, and displayed. Random access can be provided by insertion of Intra-coded pictures that do not use any prediction from other pictures. Scalability is a desirable feature for some video coding standards. This means that a decoder can access a portion of a bit-stream in order to generate a low-quality video output. This may be a reduced temporal or spatial resolution, or a reduced video quality. For MVC, additionally view scalability is required. In this case a portion of the bit-stream can be accessed in order to output a limited number out of the N original views. Also backward compatibility is required for MVC. This means that one bit-stream corresponding to one view that is extracted from the MVC bit-stream shall be conforming to H.264/AVC. Quality consistency among views is also addressed. It should be possible to adjust the encoding for instance to provide approximately constant quality over all views. Parallel processing is required to enable efficient encoder implementation and resource management. Camera parameters (extrinsic and intrinsic) should be transmitted with the bit-stream in order to support intermediate view interpolation at the decoder.

B. Test Data and Test Conditions

The proper selection of test data and test conditions is crucial for the development of a video coding standard. The test data set must be representative for the targeted area of applications, and therefore cover a wide range of different content properties. For MVC eight different multi-view test data sets have been used with 5 to 16 camera views, including linear, arc, and array arrangements. Picture resolutions are either 640x480 or 1024x768 samples, and picture rates are 15, 25, and 30 Hz. The applications rather target high quality TV-type video than limited channel communication-type video. Therefore smaller resolutions like CIF or QCIF are not considered. The MVC test data set covers a wide range of different content types, indoor and outdoor scenes, fixed and moving camera systems, different complexities of motion and spatial detail. Fig. 2 shows some examples.

In order to perform comparative evaluations also the test conditions have to be specified. For each test sequence three bit rates have been chosen corresponding to low but acceptable, medium and high quality, depending on the properties and content of the particular sequence. These fixed bit rates allow a fair comparison of different approaches for multi-view video coding in objective and subjective tests as described below.

The main goal of MVC is to provide significantly increased compression efficiency compared to individually encoding all

Fig. 2. Examples of multi-view video test data with linear camera arrangement.

video signals. Therefore encoding all views using H.264/AVC with the same test conditions was considered as the reference for coding performance comparison. The resulting decoded video signals (anchors) serve as reference for objective and subjective comparison. Encoding was done using typical settings and parameters with an *IBBP*... picture coding structure as described in [24].

C. Evaluation

Evaluation of video coding algorithms can be done using objective and subjective measures. The most widely used objective measure is the peak-signal-to-noise-ratio (PSNR) of the luma signal which is given as

$$PSNR_Y = 10 \cdot \log_{10}\left(\frac{255^2}{MSE}\right) \tag{1}$$

with MSE being the mean squared error between the original and decoded video samples. Typically PSNR values are plotted over bit rate and allow then comparison of the compression efficiency of different algorithms (e.g. anchor encoding vs. a proposed MVC scheme). This can be done in the same way for MVC.

However, PSNR values do not always capture video quality as perceived by humans. Some types of distortions that result in low PSNR values do not affect the human perception in the same way. One example is a shift of the picture by one sample side wards. Therefore any video coding algorithm can finally only be judged in subjective evaluations. The formal MVC tests were conducted by MPEG using a Single Stimulus Impairment Scale (SSIS) test. This method has proven to deliver reliable results when used for evaluation of the visual quality of video codecs and is specified in ITU-R Rec. BT.500-11 [25]. In this subjective test, subjects are being shown the decoded video signal from a candidate codec. The subjects judge the quality of decoded video on a scale from bad to excellent. The votes of the subjects are statistically analyzed to quantify subjective quality. For statistical confidence, a large number of subjects need to be involved. Display conditions, viewing room conditions (including lighting and view distance), and execution of test sessions (order of presented video, display time, etc.) require careful design. In consequence such formal subjective tests require a tremendous effort.

III. Temporal and Inter-view Correlation

As denoted in the introduction, the main difference between classic video coding and multi-view video coding is the availability of multiple camera views of the same scene. As coding efficiency of hybrid video coding depends on the quality of the prediction signal to a great extent, a coding gain can be achieved for MVC by additional inter-view prediction. If there is no such gain, independently encoding each camera view with temporal prediction would already provide the best possible coding efficiency.

Therefore this section shows an analysis of the properties of temporal and inter-view correlation by investigating the statistical dependencies that can be exploited for prediction. Fig. 3 shows the eight possible first order spatial and temporal neighbor pictures of a picture in a MVV sequence with linear camera arrangement, where *S* indicates the cameras and *T* the subsequent time-points. The purpose of the analysis is to determine by which percentage a rate-distortion optimized encoder such as H.264/AVC would choose either one of these prediction modes, if all of them are available. In Fig. 3 the investigated prediction modes are indicated by different colors and indices of the reference pictures, where black indicates temporal, blue inter-view, and red and green mixed inter-view/temporal prediction.

Many of today's video codecs like H.264/AVC are based on predictive coding between one or more reference pictures and the currently encoded picture [26]. Motion estimation is conducted by minimizing a Lagrangian cost function [24]

$$J = D + \lambda \cdot R \tag{2}$$

This Lagrangian cost function J is the sum of rate R and distortion D, weighted by the Lagrange parameter λ. For each block S_i of a picture, the motion estimation algorithm chooses the motion vector m_i within a search range M in the reference picture that minimizes J.

$$m_i = \arg\min\{D(S_i, m) + \lambda \cdot R(S_i, m)\} \tag{3}$$

Here the distortion is calculated as the sum of squared errors between the current picture s and the previously decoded reference picture s'.

$$D(S_i, m) = \sum_{(x,y)\in B}\left[s(x,y,t) - s'(x - m_x, y - m_y, t - m_t)\right]^2 \tag{4}$$

The rate R is the number of bits to transmit all components of the motion vector.

To analyze the temporal and inter-view statistical dependencies of multi-view video sequences, every picture P

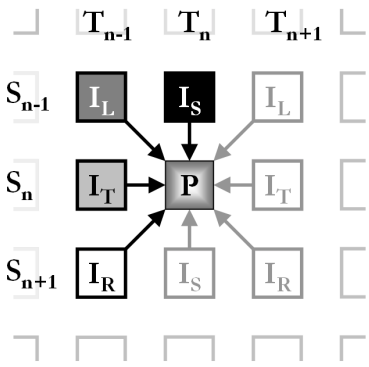

Fig. 3. Prediction with first order spatial and temporal neighbor images.

TABLE I
RESULTS OF THE TEMPORAL AND INTER-VIEW CORRELATION ANALYSIS

Sequence Name	T [%]	S [%]	R [%]	L [%]	ΔJ [%]
Ballroom	74.98	12.12	6.86	6.04	-5.65
Exit	76.96	8.66	7.49	6.90	-2.99
Uli	93.06	2.23	2.58	2.13	-0.52
Race1	96.64	1.35	1.06	0.96	-1.84
Breakdancers	57.95	19.30	12.15	10.60	-7.71

TABLE II
RESULTS OF THE SIMPLIFIEDTEMPORAL AND INTER-VIEW CORRELATION ANALYSIS

Sequence Name	T [%]	S [%]	ΔJ [%]
Ballroom	83.24	16.76	-4.07
Exit	86.42	13.58	-1.66
Uli	95.65	4.35	-0.33
Race1	98.26	1.74	-1.34
Breakdancers	70.93	29.07	-6.16

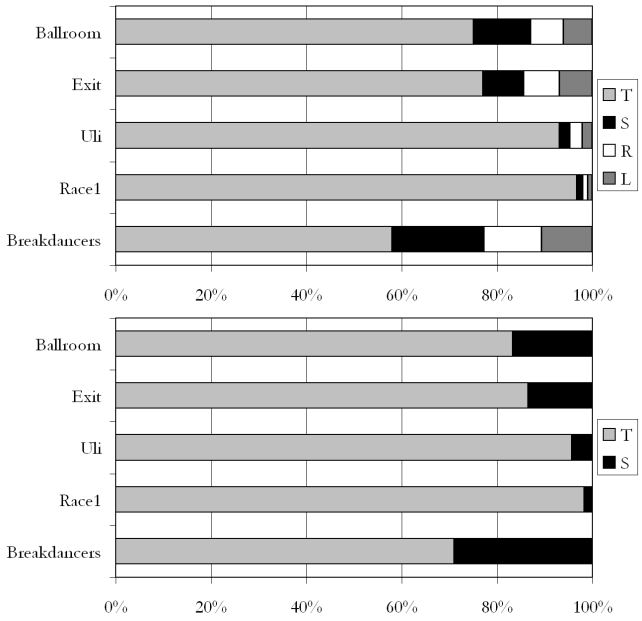

Fig. 4. Probabilities prediction modes in terms of providing the prediction signal with the lowest Lagrange cost factor. (top: temporal, inter-view and mixed prediction, bottom: simplified temporal vs. inter-view prediction)

of a multi-view data set is encoded with each of the four reference pictures $I_{T,S,R,L}$ for motion-compensated prediction, as depicted in Fig. 3.

If the same settings are used, the prediction mode associated with the smallest J can be identified for each block of the picture P. Table I and Fig. 4 show the results of this analysis over five multi-view data sets, where the percentage values give the portion of blocks (labeled as T, S, R, and L in Fig. 3) that are chosen for prediction with the reference picture of the corresponding mode in terms of the lowest Lagrange cost function value J. The last column in Table I shows the decrease of the average J in comparison to temporal prediction only. For this analysis an H.264/AVC encoder was used with the following settings: Disabled intra prediction-modes for the P pictures together with a fixed motion compensation block size of 16x16 samples, a search range of ±32 pixels and a Lagrange parameter of 29.5.

The results of this analysis over several multi-view data sets show, that often inter-view prediction is more efficient than temporal prediction for a significant number of blocks, although for all sequences temporal prediction is the most often chosen mode. A comparison between inter-view prediction only and combined inter-view/temporal prediction modes shows that inter-view prediction is more efficient than mixed inter-view/temporal prediction modes. These results indicate that coding gains can be expected with a prediction structure optimized for multi-view video, especially as the characteristics of the presented results did not change significantly, if the size of the search range and the Lagrange parameter were varied in supplementary experiments.

Since the percentages of the mixed inter-view/temporal prediction modes are relatively low, an additional analysis was carried out, which simply included temporal and inter-view prediction modes. The results of this analysis in Table II and Fig. 4 clearly indicate that prediction quality does not degrade much without using mixed mode reference pictures, but leads to a considerable complexity reduction concerning a real multi-view encoder. Especially the values for ΔJ decrease just slightly and thereby indicate the minor influence of mixed mode reference pictures on coding efficiency.

The fact that temporal prediction is on average the most efficient mode for all analyzed sequences results from the global displacement and distortion between the pictures of different camera perspectives. Especially, static image areas, like the background, cause a fundamental disadvantage for inter-view prediction. In general, the relationship between temporal and inter-view prediction strongly depends on the properties of the multi-view video sequence, particularly the temporal and spatial density on one and scene complexity on the other hand. Additional effects, that should be mentioned, are differences in brightness and color between the video streams of the individual cameras.

IV. 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.

A. Temporal Prediction

As mentioned previously, 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.

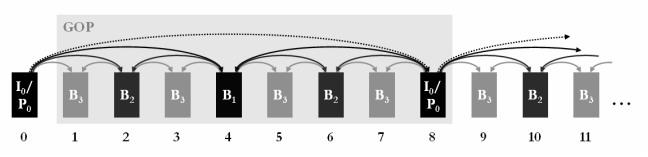

Fig. 5. Hierarchical reference picture structure for temporal prediction.

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 [27] 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 [3][24]. A typical hierarchical prediction structure with three stages of a dyadic hierarchy is depicted in Fig. 5. The first picture of a video sequence is intra-coded as IDR picture and so-called key pictures (black in Fig. 5) 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. 5 for a GOP of eight pictures.

The temporal hierarchy levels of the prediction structure are denoted by the indices in the figure and it has to be ensured that all pictures are predicted by using only pictures of the same or a higher temporal hierarchy level as references, to support several temporal scalability levels. As depicted in Fig. 5, the B pictures of such a hierarchical structure are typically predicted by using the two nearest pictures of the next higher temporal level as references. Regarding the coding order, this prediction structure implies the constraint that a picture's reference pictures have to be encoded before this picture itself is encoded, which especially affects the key pictures. Besides some minor changes to the encoder control, the new strategy of cascading the quantization parameter (QP) depending on the temporal hierarchy level should be mentioned, because it essentially contributes to the improved coding efficiency with hierarchical B pictures. For QP cascading, key pictures get assigned the highest fidelity with a successive decrease in fidelity when moving down the temporal hierarchy [27].

The concept of hierarchical B pictures can easily be applied to multi-view video sequences as illustrated in Fig. 6 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 key pictures are coded in intra mode. 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.

B. 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 [3]. Consequently replacing intra-coded (or *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. 6 leads to the prediction scheme in Fig. 7. The prediction structure of the first view *S0* remains the same and is called base view, as it is identical to the simulcast prediction structure with hierarchical B pictures for temporal prediction only. But 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 to be 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.

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

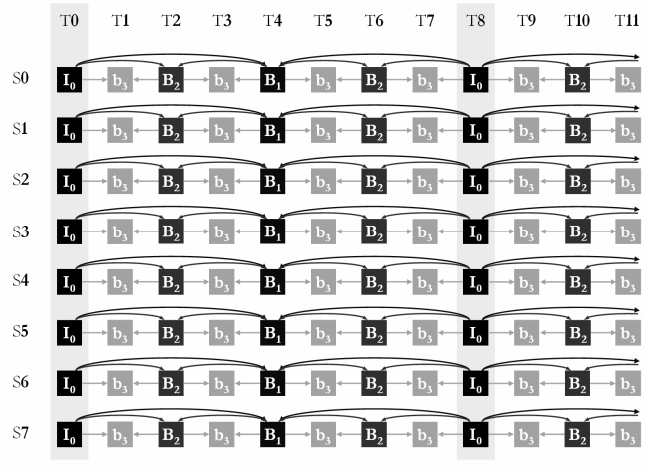

Fig. 6. Temporal Prediction using hierarchical B pictures.

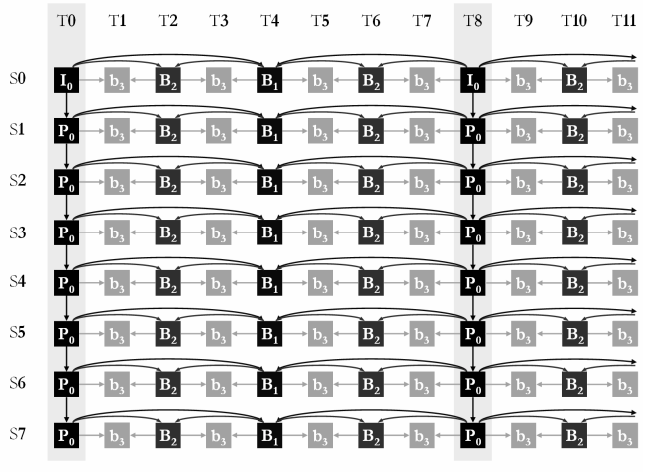

Fig. 7. Inter-view prediction for key pictures.

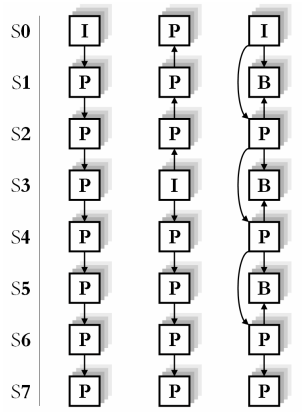

Fig. 8. Alternative structures of Inter-view prediction for key pictures.

views S_n cannot 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. 8 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. 7 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 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. This prediction structure might have coding efficiency advantages over the other configurations, at the disadvantage of being more complex.

A question arises with regards to the use of hierarchical B pictures for inter-view prediction. Please note that the QP cascading is very important for the efficiency of hierarchical B pictures. When hierarchical B pictures are applied in temporal direction, the varying fidelity between the pictures is not disturbing to the viewer as each picture is shown only a limited amount of time. However, if we would apply the hierarchical B picture approach including QP cascading to inter-view prediction, some views are coded with much lower fidelity than others. Since the viewer may be looking at one particular view for a long time, this may result in disturbing artifacts.

C. *Inter-view Prediction for Non-Key Pictures*

The analysis of temporal and inter-view prediction efficiency in Section III indicates 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.

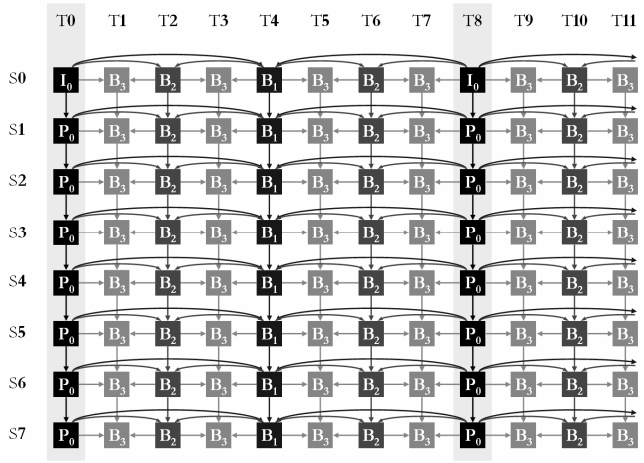

Fig. 9. Inter-view prediction for key and non-key pictures pictures.

Fig. 9 illustrates how the prediction structure of Fig. 7 can be extended to inter-view prediction with non-key pictures, 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 left scheme in Fig. 8, 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 one inter-view reference frame for every view, referred to as *AS_IPP* mode. Additionally the other alternative structures for inter-view prediction at key picture level presented in Fig. 8 can similarly be extended to non-key pictures, referred to as *AS_PIP* and *AS_IBP* accordingly. In contrast to the prediction structures of section IV.B, where the maximum number of reference pictures is two, now the non-key pictures have up to four references. Thus coding efficiency is improved at the cost of increased coding complexity. Furthermore synchronization and random access features are still provided by coding the key pictures of the base view in intra mode. Further 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 [3][4][5].

V. Experimental Results

This section presents the results of coding experiments with the prediction structures described in the previous section. As already denoted in section II, it is necessary to perform both objective and subjective evaluation of the coding quality for new MVC techniques. First of all the next section explains, how experiments with these MVC prediction structures can be implemented and how to configure them in order to achieve comparable results.

A. *Setup of Coding Experiments and Test Conditions*

The most important aspect regarding coding experiments using the prediction structures presented in section IV 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. 10, using a specific scan. This uncompressed video bit-stream is used as the input of the H.264/AVC encoder. The prediction structure itself is controlled by appropriate settings of the encoder's parameters for reference picture selection and memory management [28]. This kind of encoder configuration is well-established for hierarchical B pictures with temporal prediction.

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. In addition, the presented schemes for multi-view video coding simply require some high level syntax specification to signal that the bit-stream represents a multi-view sequence with N views. Then the decoder can set the Decoded Picture Buffer size appropriately, decode the bit-stream with existing tools, and knows how to invert the reordering in Fig. 10.

All presented experiments and results are based on the eight test data sets together with the coding conditions specified by MPEG (as described in section II). One task is to adapt the multi-view prediction schemes to the specific camera arrangements of the test data sets. The examples in section IV already demonstrated the inter-view prediction structures for one-dimensional linear or arched setups. Fig. 11 and Fig. 12 illustrate the customized inter view-prediction structures for a crosswise camera arrangement and for a 5x3 array of cameras respectively.

A second task is to adapt the prediction structures to the random access specifications. This can be done without any problems by customizing the GOP length, as the presented schemes provide all the features of hierarchical B pictures for temporal prediction. Fig. 13 illustrates possible settings for GOP lengths of 12 and 15 pictures.

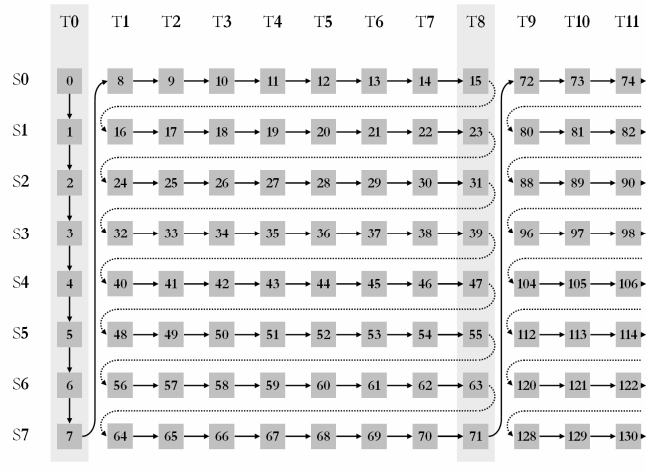

Fig. 10. Frame interleaving for compression with H.264/AVC.

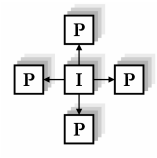

Fig. 11. Adapted prediction structure for a crosswise camera arrangement (*Flamenco2* test sequence).

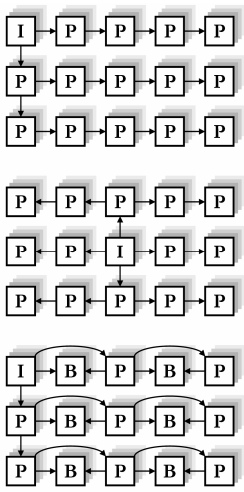

Fig. 12. Adapted prediction structures for a camera array (*AkkoKayo* test sequence).

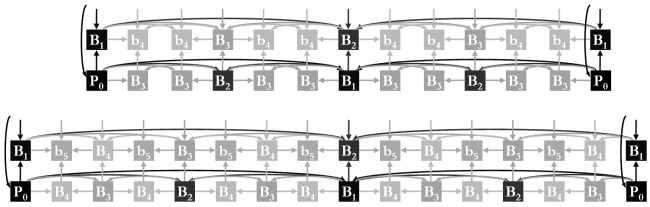

Fig. 13. Flexible length of multi-view GOPs (top: 12 pictures, bottom: 15 pictures).

B. Objective Evaluation

The experiments presented in this section are performed with H.264/AVC conforming encoder/decoder software, using typical settings for multi-view video coding (see [19] for details), like variable block size, a search range of ±96, CABAC enabled and rate control using Lagrangian techniques.

Example results using the different prediction structures of section IV are shown in Fig. 14. The $PSNR_Y$ values are plotted over bit-rate averaged over all views of a data set. In Fig. 14, *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 *AS_IPP/AS_IBP* 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

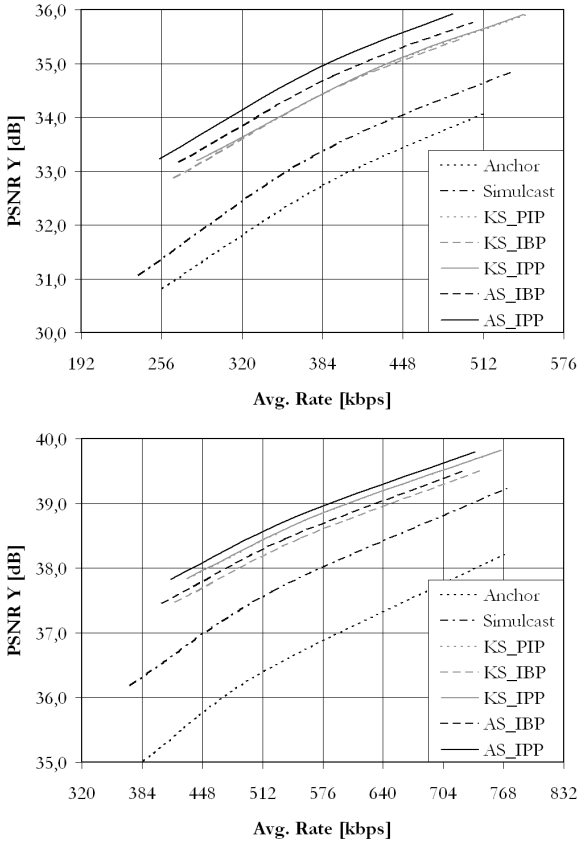

Fig. 14. PSNR results (top: *Ballroom*, bottom: *Race1* test sequence).

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. 15 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 strongly depend on the temporal and inter-view correlations.

As mentioned in section III, attributes like temporal and spatial density as well as scene complexity strongly influence these correlations. 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.

By comparing the coding gains in Fig. 15 with the analysis results in section III the following correlation can be found: The ΔJ values presented in section III only indicate the ratio between the gains for coding with and without inter-view prediction and are not directly related to the total coding gain.

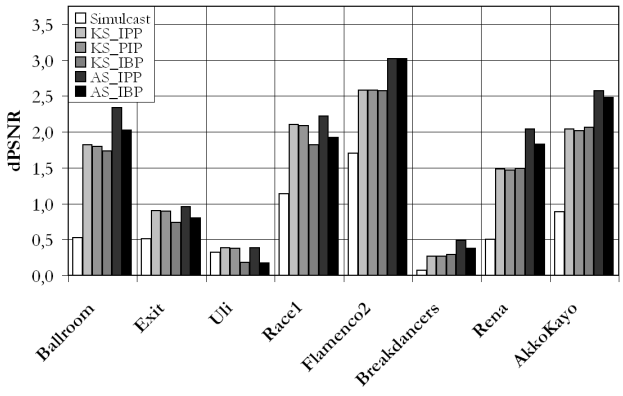

Fig. 15. Average coding gains.

Although the *AS_xxx* prediction structures achieve the highest average coding gain of 1.7 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. 15. 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 B pictures at key picture temporal level. For some sequences, prediction over two views and cascading the QPs according to prediction hierarchy levels (see section IV.A) turns out to be a disadvantage and if so, always both *KS_IBP* and *AS_IBP* perform worse than *KS_IPP*/*KS_PIP* and *AS_IPP*.

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, shows larger deviations, while sequences like *Rena* with very small camera distance show small deviations due to more similar content across all the views. But besides these data set dependent aspects, the quality distribution is also affected by the inter-view prediction structure. The corresponding experiments confirm, that using a *xx_IBP* multi-view structure results in larger deviations than for the corresponding *xx_IPP* structure without B pictures at key picture temporal level. 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 the *AS_xxx* prediction structures 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.

C. Subjective Evaluation

Normally PSNR already provides a very good indication about performance of a compression method. However, the final judgment can only be done by subjective tests, where humans evaluate visual quality. For this purpose MPEG has conducted exhaustive formal subjective tests to evaluate the performance of the responses to the CfP [22]. The tests were conducted at the Technical University of Munich.

From each of the 8 test data sets, 2 views were selected

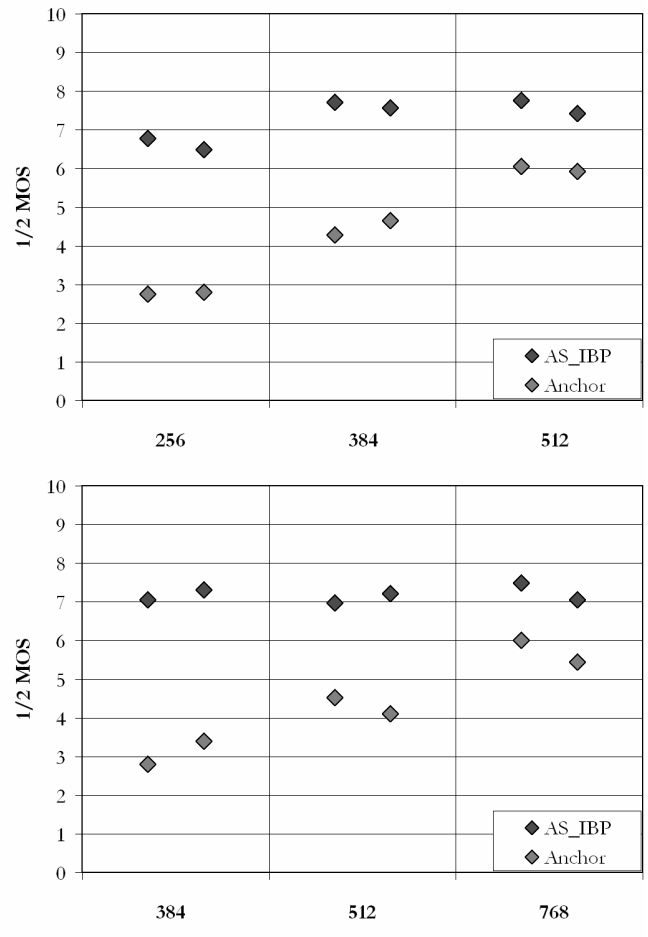

Fig. 16. Subjective results (top: *Ballroom*, bottom: *Race1* test sequence).

randomly for subjective evaluation for all 3 specified rate points. A method called Single Stimulus Multimedia (SSMM) was selected, which is a modified version of the Single Stimulus Impairment Scale (SSIS), and has proven efficiency and reliability in prior tests conducted by MPEG. In a Single Stimulus test, the test subjects have to rate the decoded video in the absence of an unimpaired reference. To minimize the influence of the videos shown previously each test case was shown twice. Showing all test cases twice also helped to verify that all subjects (20 "naïve viewers") were able to reproduce their votes. A modified mean opinion score (MOS) with values from 0-10 was used to capture all the votes. Basically the subjects judged the quality for each example individually by giving marks. Then the results were evaluated statistically.

Fig. 16 shows the MOS results of the presented *AS_IBP* prediction scheme in comparison to the MPEG anchors for 2 examples. These are the MOS values for the 2 randomly selected views at each bit-rate. Note that different bit-rates were selected for the different data sets. Obviously, the *AS_IBP* prediction structure outperforms the anchors significantly in terms of subjective quality.

Fig. 17 compares the average MOS values over all test data sets and selected views. These were averaged for the low, mid

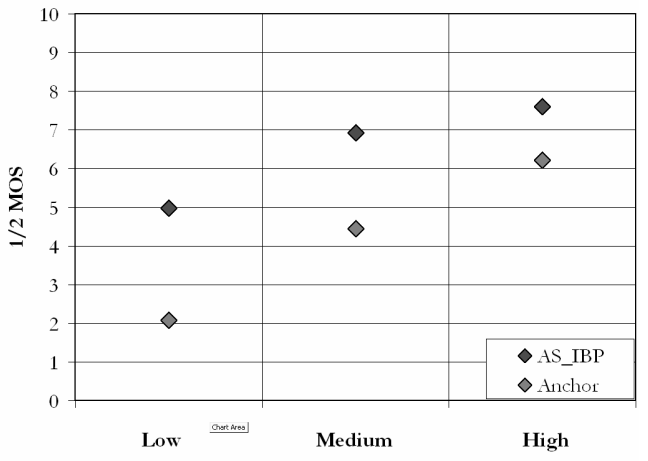

Fig. 17. Average subjective results over all test sequences.

and high bit-rate separately. Note that these bit-rates were different for all data sets. In summary, the presented MVC proposal outperforms the MPEG anchors significantly. The gain decreases slightly with higher bit-rates.

D. Influence of Camera Density

This section presents an experimental evaluation on the influence of camera density for MVC. In order to simulate different camera distances, the experiments are realized with the setups depicted in Fig. 18. Starting from the right-hand side in Fig. 18 with 9 camera views of a linearly arranged MVV data set, moving to the left in Fig. 18, camera views are removed by omitting every other camera view successively obtaining 5, 3, and 2, and 1 camera views. The experiments are carried out using the *Rena* sequence, which is the densest of the available MVV sequences, consisting of 16 linear arranged cameras with a 5 cm distance between the centers of two adjacent cameras. To obtain a regular camera distance refinement, we used 9 adjacent cameras. Since 16 camera views are available, the experiments depicted in Fig. 18 were repeated for each shifted set of 9 adjacent cameras. This approach minimizes irregular influences of individual camera views by averaging over the individual results of the repetitions, each shifted by one camera.

Since it is intended to analyze the influence of camera density on inter-view predictive coding, the structures of Fig. 18 are applied to every time instance of the MVV sequence without temporal prediction. The coding experiments presented in this section are performed with H.264/AVC conforming encoder/decoder software, using the typical settings for multi-view vide coding (see [19] for details), i.e. variable block size motion compensation, a search range of ±96, CABAC and rate control via Lagrangian techniques..

The experiments are realized by encoding the sequences applying the different inter-view prediction structures when varying over QP values. From every resulting bit-stream for one combination of prediction structure and QP value, the average bit-rate is computed. In order to have a reference bit-

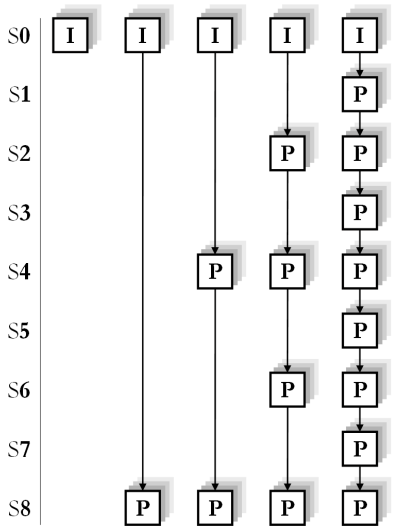

Fig. 18. Setup of coding experiments on camera density using 1, 2 3, 5, and 9 original cameras (from left to right).

rate for each of the 9 cameras involved in the coding experiments with inter-view prediction, the 1-camera method is applied to all of them and the result is averaged. Thus a pair of bit-rate values is obtained for each combination of prediction structure and QP value, one for encoding with inter-view prediction and one reference bit-rate for encoding without inter-view prediction.

For the analysis presented in Fig. 19, each bit-rate of an inter-view prediction case is divided by its reference bit-rate. The resulting relative rates that are given as percentage increase make it possible to sum up all results in one diagram, starting at 100% for coding each view independently. Consequently the results for different QP values can directly be compared with each other. While the top diagram in Fig. 19 shows the relative rate for all cameras, the bottom diagram shows the relative rate per camera and therefore the derivate of the functions in Fig. 19 top. Both diagrams contain the curves for four different QP values (dashed and solid), representing different reconstruction qualities, and the straight line for the reference method (dotted). According to the structures in Fig. 18 the achieved rates are plotted over the number of cameras used for inter-view prediction. As aforementioned the results are averaged over the individual shifted repetitions of 9 adjacent cameras along the 16 cameras of the *Rena* sequence.

The data rate for inter-view prediction increases less than *N*-times the number of cameras, which results in a coding gain due to inter-view prediction. This coding gain increases with both parameters: decreasing camera distance and decreasing reconstruction quality. This characteristic is best highlighted in the bottom diagram. As expected, the gain strongly depends on the reconstruction quality, namely a larger QP value leads to a larger coding gain with respect to the reference method for any camera distance. For example the bit rate reduction for coding with QP42 results in about twice the amount for coding with QP30, i.e. 32% vs. 59% for a camera distance of

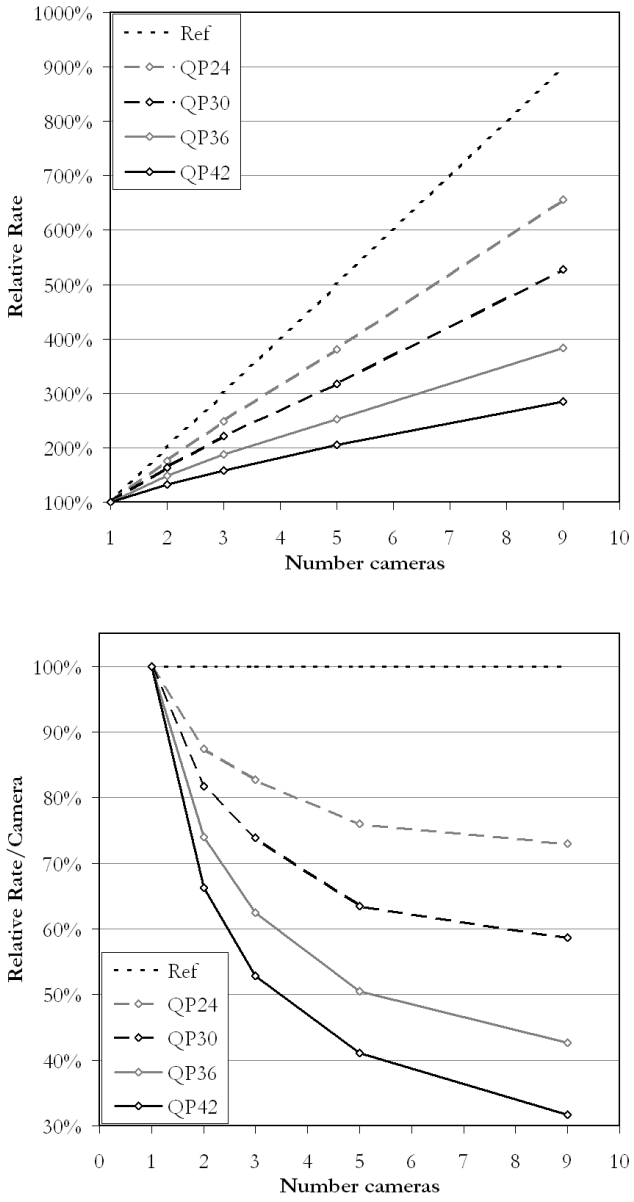

Fig. 19. Results of coding experiments on camera density with *Rena* test sequence in terms of average rate (top) and average per camera rate (bottom) relative to the one camera case.

1. Since the relative rates per camera in the bottom diagram tend to saturate with an increasing number of cameras, the relation between camera distance and coding gain has a logarithmic superimposed by a linear characteristic. For example the relative rate per camera for coding with QP24 tends to saturate at a value of about 70%.

Our experiments indicate that if the camera distance is decreased to a very small practical value (like 5cm), the rate per camera tends to saturate at a certain value and the saturation value itself only depends on the reconstruction quality. This implicates, that even the densest practical camera settings require a data rate linearly proportional to the number of cameras, if MVC with inter-view prediction is applied the way described in this paper.

VI. Conclusion

The presented prediction structures for multi-view video coding are based on the fact that multiple video bit-streams, showing the same scene from different camera perspectives, show significant inter-view statistical dependencies. The corresponding evaluation pointed out, that these correlations can be exploited for efficient coding of multi-view video data. The combined temporal and inter-view prediction structures are based on the idea that inter-view prediction is supported 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 very minor syntax changes. 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.

Regarding coding efficiency, the reported results clearly indicate the assets and drawbacks of this multi-view video coding approach. Coding gains up to 3.2 dB and an average gain of 1.5 dB could be achieved by additionally using inter-view reference pictures for disparity-compensated prediction.

However for some test sequences neither of the tested prediction structures resulted in an improved coding efficiency, addressing two basic problems of multi-view video coding. One problem is large disparities between the different views of multi-view video sequences. The problem of large disparities could be solved by depth-based view interpolation prediction [13][14]. The idea is to estimate depth either at the encoder (this 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 instance, if every other view is transmitted first and depth information is available, it is possible to generate prediction for the views in-between the available views. 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). In terms of varying the camera distance, it was found, that even for the densest practical camera setting a non-vanishing data rate per camera was obtained, leading to a primarily linear dependency of total data rate vs. number of cameras. Thus, methods beyond pure image-based coding, as shown in this article, are required to further improve coding efficiency.

The second problem of multi-view video coding is illumination and color inconsistencies across views that also affect the exploitation of inter-view statistical dependencies, for instance with the *Uli* test sequence. Usually such effects should be minimized by proper setting of the conditions; however, an MVC algorithm should also be able to cope with this as well, since perfect white-level and color balancing of the input can not be guaranteed. Also, the illumination (spotlights, shadows, etc.) varies largely over the multi-view images due to the lighting conditions. For MVC, compensation of differences in illumination and color is realized by modifying the prediction process of H.264/AVC on a block level [29][30], whereby additional coding gains of up to 0.6 dB are reported. 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 25.-26. 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] ITU-T Rec. & ISO/IEC 14496-10 AVC, “Advanced Video Coding for Generic Audiovisual Services,” version 3, 2005.

[4] 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

[5] G. Sullivan and T. Wiegand, “Video Compression - From Concepts to the H.264/AVC Standard”, Proc. of the IEEE, Special Issue on Advances in Video Coding and Delivery, Vol. 93, No. 1, January 2005, p. 18

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Y. Yang, G. Jiang, M. Yu, F. Li, and Y. Kim, “Hyper-Space Based Multiview Video Coding Scheme for Free Viewpoint Television”, Proc. PCS 2006, Picture Coding Symposium, Beijing, China, April 2006.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] A. Smolic and D. Mc Cutchen, “3DAV Exploration of Video-Based Rendering Technology in MPEG”, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Immersive Communications, Vol. 14, No. 9, pp. 348-356, March 2004.

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

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

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

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

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

[22] ISO/IEC JTC1/SC29/WG11, “Subjective test results for the CfP on Multi--view Video Coding”, Doc. N7779, Bankok, Thailand, January 2006.

[23] A. Vetro, Y. Su, H. Kimata, and A. Smolic, “Joint Draft 1.0 on Multiview Video Coding”, Joint Video Team, Doc. JVT-U209, Hangzhou, China, October 2006.

[24] 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.

[25] ITU-R BT.500-11, “Methodology for the subjective assessment of the quality of television pictures”, 2002.

[26] T. Wiegand, X. Zhang, and B. Girod, "Long-Term Memory Motion-Compensated Prediction", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 1, pp. 70-84, February 1999.

[27] 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.

[28] H. Schwarz, M. Wien, and J. Vieron, „JSVM Software Manual”, Joint Video Team, Doc. JVT-S070, Geneva, Swizerland, April 2006.

[29] 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.

[30] 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.

[31] O. Schreer, P. Kauff, and T. Sikora, „3D Videocommunication”, 1st edition, John Wiley & Sons, Chichester, July 2005.

[32] E. Lamboray, S. Wümlin, M. Waschbüch, M. Gross, and H. Pfister, “Unconstrained Free-Viewpoint Video Coding”, Proceedings of the IEEE International Conference on Image Processing (ICIP) 2004, Singapore, October 24-27, 2004.

[33] A. Smolic, K. Mueller, P. Merkle, T. Rein, P. Eisert, and T. Wiegand, “Representation, Coding, and Rendering of 3D Video Objects with MPEG-4”, MMSP 2004, IEEE International Workshop on Multimedia Signal Processing, Siena, Italy, September 29.-October 1. 2004.

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

[35] M. Tanimoto, “Free Viewpoint Television - FTV”, Proc. PCS 2004, Picture Coding Symposium, San Francisco, CA, USA, December 15.-17. 2004.

[36] T. Fujii and M. Tanimoto, “Free-Viewpoint TV System Based on Ray-Space Representation”, SPIE ITCom Vol. 4864-22, pp.175-189 (2002).

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

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

[39] ISO/IEC JTC1/SC29/WG11, "ISO/IEC 14496-16/PDAM1", Doc. N6544, Redmont, WA, USA, July 2004.
[40] W. Matusik and H. Pfister, "3D TV: A Scalable System for Real-Time Acquisition, Transmission and Autostereoscopic Display of Dynamic Scenes", ACM Transactions on Graphics (TOG) SIGGRAPH, ISSN: 0730-0301, Vol. 23, Issue 3, pp. 814-824, August 2004.
[41] M. Magnor and B. Girod, "Data Compression for Light-Field Rendering", IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 3, pp. 338-343, Apr. 2000.
[42] J. Li, H.Y. Shum, and Y.Q. Zhang, "On the Compression of Image Based Rendering Scene", Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[43] P. Eisert, E. Steinbach, and B. Girod, "Automatic Reconstruction of Stationary 3-D Objects from Multiple Uncalibrated Camera Views", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 2, pp. 261-277, March 2000.
[44] T. Matsuyama and T. Takai, "Generation, Visualization, and Editing of 3D Video", Proc. Symposium on 3D Data Processing Visualization and Transmission, pp.234-245, Padova, Italy, June 2002.
[45] W.H. Leung and T. Chen, "Compression with Mosaic Prediction for Image-Based rendering Applications", Proc. ICME2000, IEEE International Conference on Multimedia and Expo, New York, NY, USA, July 2000.
[46] J. Li, H.Y. Shum, and Y.Q. Zhang, "On the Compression of Image Based Rendering Scene", Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[47] T. Pintaric, U. Neumann, and A. Rizzo, "Immersive Panoramic Video", Proceedings of the 8th ACM International Conference on Multimedia, pp. 493.494, October 2000.
[48] C. Zhang, and J. Li, "Compression of Lumigraph with Multiple Reference Frame (MRF) Prediction and Just-in-time Rendering", Proc. DCC2000, IEEE Data Compression Conference, Snowbird, Utah, USA, March 2000.
[49] S. Würmlin, E. Lamboray, O. Staadt, and M. Gross, "3D Video Recorder: A System for Recording, Processing and Playing Three-Dimensional Video", Computer Graphics Forum 22 (2), Blackwell Publishing Ltd, Oxford, U.K., pp. 181-193, 2003.
[50] S. Würmlin, E. Lamboray, and M. Gross, "3D video fragments: dynamic point samples for real-time free-viewpoint video", Computers and Graphics 28 (1), Special Issue on Coding, Compression and Streaming Techniques for 3D and Multimedia Data, pp. 3-14, Elsevier Ltd, 2004.
[51] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, and I. Sexton, "An Evolutionary and Optimised Approach on 3D-TV", Proc. of IBC 2002, Int. Broadcast Convention, Amsterdam, Netherlands, Sept. 2002.

Philipp Merkle received the Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin, Germany, in 2006.

He joined the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, Germany, as a student in 2003 and became a Research Assistant in 2006. He has been involved in several projects focused on multiview video coding, free viewpoint video, 3D scene reconstruction and augmented reality.

His research interests include representation and coding of multiview video scenes, free viewpoint video, 2D and 3D video-based rendering. He has been involved in ISO standardization activities where he contributed to the development of the MPEG-4 multiview video coding standard.

Aljoscha Smolic received the Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin, Germany in 1996, and the Dr.-Ing. degree in electrical and information engineering from Aachen University of Technology (RWTH), Germany, in 2001.

He joined the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, Germany in 1994 and is employed as Scientific Project Manager since 2001. He has been involved in several national and international research projects, where he conducted research in various fields of video processing, video coding, computer vision and computer graphics and published more than 80 referred papers in these fields. In this context he has been involved in ISO standardization activities where he contributed to the development of the multimedia standards MPEG-4 and MPEG-7. In current projects he is responsible for research in free viewpoint and 3D video processing and coding, augmented reality, 3D reconstruction and video-based rendering. Since 2003 he is Adjunct Professor at the Technical University of Berlin and teaches Multimedia Communications and Statistical Communications Theory.

Dr. Smolić received the "Rudolf-Urtlel-Award" of the German Society for Technology in TV and Cinema (FKTG) for his dissertation in 2002. He is Area Editor for Signal Processing: Image Communication and Guest Editor for IEEE Transactions on CSVT and IEEE Signal Processing Magazine. He is Committee Member of several conferences, including ICIP, ICME, and EUSIPCO. He chaired the MPEG ad hoc group on 3DAV pioneering standards for 3D video. Currently he is editor of the Multi-view Video Coding (MVC) standard and co-chairs a working group of the JVT for MVC.

Karsten Müller (M'98-SM'07) received the Dr.-Ing. degree in Electrical Engineering and Dipl.-Ing. degree from the Technical University of Berlin, Germany, in 2006 and 1997 respectively.

He has been with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Berlin, since 1997, where he is currently project manager. His research interests are mainly in the field of representation, coding and reconstruction of 3D scenes in Free Viewpoint Video scenarios and Coding, Multi-view applications and combined 2D/3D similarity analysis. He has been involved in MPEG activities, where he contributed to visual MPEG-7 descriptors and MPEG-4.

Thomas Wiegand (M'05) is the head of the Image Communication Group in the Image Processing Department of the Fraunhofer Institute for Telecommunications - Heinrich Hertz Institute Berlin, Germany. He received the Dipl.-Ing. degree in Electrical Engineering from the Technical University of Hamburg-Harburg, Germany, in 1995 and the Dr.- Ing. degree from the University of Erlangen-Nuremberg, Germany, in 2000. His research interests include video processing and coding, multimedia transmission, semantic image representation, as well as computer vision and graphics.

From 1993 to 1994, he was a Visiting Researcher at Kobe University, Japan. In 1995, he was a Visiting Scholar at the University of California at Santa Barbara, USA. From 1997 to 1998, he was a Visiting Researcher at Stanford University, USA and served as a consultant to 8x8, Inc., Santa Clara, CA, USA. He is currently a member of the technical advisory boards of the

two start-up companies Layered Media, Inc., Rochelle Park, NJ, USA and Stream Processors, Inc., Sunnyvale, CA, USA.

Since 1995, he is an active participant in standardization for multimedia with successful submissions to ITU-T VCEG, ISO/IEC MPEG, 3GPP, DVB, and IETF. In October 2000, he was appointed as the Associated Rapporteur of ITU-T VCEG. In December 2001, he was appointed as the Associated Rapporteur / Co-Chair of the JVT. In February 2002, he was appointed as the Editor of the H.264/AVC video coding standard and its extensions (FREXt and SVC). In January 2005, he was appointed as Associated Chair of MPEG Video.

In 1998, he received the SPIE VCIP Best Student Paper Award. In 2004, he received the Fraunhofer Award for outstanding scientific achievements in solving application related problems and the ITG Award of the German Society for Information Technology. Since January 2006, he is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology.