Analysis of Multi-Reference Block Matching for Multi-View Video Coding

André Kaup Ulrich Fecker

Chair of Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg,
Cauerstraße 7, 91058 Erlangen, Germany, Email: {kaup,fecker}@LNT.de

Dedicated to Prof. Dr.-Ing. Heinz Gerhäuser on occasion of his 60th birthday

Abstract

Increasing interest can be observed in applications of image-based rendering, especially three-dimensional television (3D TV). In such systems, the original object or scene is recorded using a setup of several cameras. From the recorded video streams, arbitrary views can be interpolated, so that the user can navigate freely around the scene. However, the amount of data involved is huge, and efficient compression is needed. Several coding schemes have been proposed which exploit the spatial correlation between the views in addition to the temporal correlation between subsequent frames. In this paper, the block matching step of multi-view video coding is statistically analysed. Simulation results for several test sequences are presented. It is shown how much prediction gain can be achieved by introducing block matching across the camera views. The distribution of best matches among a set of possible references is investigated in detail, and from these results, conclusions can be drawn about the reasonable selection of reference frames in a practical multi-view video coding scheme.

1. INTRODUCTION

Traditionally, real objects are described by a three-dimensional model of their surface when they shall be represented in a computer. For being able to synthesise realistic views of the object, an accurate description of the optical properties of the surface is required. For objects with complex structures and reflective characteristics, this is hardly possible with the accuracy needed to render photorealistic views. Examples for such structures include natural objects such as fur or smoke, but also technical devices with a complex structure.

Image-based rendering is an approach to overcome these problems. The idea is to capture the so-called *light field* of a three-dimensional object or scene. For that, a large number of images is taken by multiple cameras from different positions. The acquired data can then be used to reproduce photorealistic images of the scene for any desired viewpoint and for any viewing angle. For viewpoints not coinciding with the original camera positions, intermediate views can be interpolated from the captured images. Ideally, no information about the geometry or the surface characteristics is needed [9], [5].

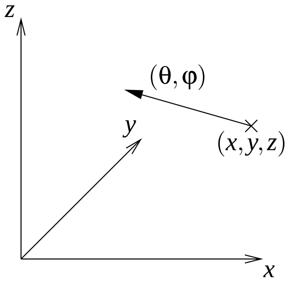

Figure 1. The plenoptic function

A light field can be represented by the *plenoptic function:*

$$\Phi\left(x, y, z, \theta, \varphi, \lambda, t\right) \tag{1}$$

This seven-dimensional function is defined as the light intensity depending on the viewpoint x, y, z, the viewing angle θ, φ, the wavelength λ and the time t (see Fig. 1).

The plenoptic function fully describes the radiance in the space around the desired object or scene. However, this function is far too general and too complex to be dealt with. Therefore, simplifications need to be introduced: Usually, the parameter λ is eliminated by introducing three colour channels, e. g. red, green and blue:

$$\Phi_{\mathrm{R,G,B}}\left(x, y, z, \theta, \varphi, t\right) \tag{2}$$

Furthermore, the function is in most cases not sampled throughout the entire space, but only on a surface surrounding the scene [20]. This configuration is then rather well applicable to a setup where the light emanating from a scene is captured by several cameras surrounding it.

If the recorded object or scene does not change over time, the parameter t can also be eliminated, and the light field is called a *static light field.* In this paper, the case of *dynamic light fields* is considered, where the light emitted by the object or scene varies over time. Dynamic light fields could find applications in medicine, where e. g. a surgeon could spatially examine a beating heart, or in a virtual webshop, where customers could view changing three-dimensional objects with complex surfaces from any desired viewpoint.

Another application, which has attracted increasing interest from the industry as well as from research institutes, is *three-dimensional television (3D TV).* In a 3D TV system, the scene is recorded with several cameras, transmitted and displayed on a 3D display. Such a system has e. g. been presented in [17], where the 3D experience is based on an array of multiple projectors. For *free viewpoint television (FTV)*, the scene is displayed on a conventional 2D display, but the user can choose his viewpoint and viewing angle freely [13].

To realise such systems, video streams from multiple cameras need to be recorded simultaneously. This leads to an enormous amount of data which must be stored or transmitted to the user. Therefore, efficient compression techniques are crucial for practical systems. For static light fields, where large numbers of still images are recorded, compression techniques were developed in [10]. In the case of dynamic light fields, when a *multi-view video* sequence is captured, the amount of data is even higher.

A straightforward method to compress multi-view video data is *simulcast coding*, which means that the video stream from each camera is coded separately using a normal video coder, for example the H.264/AVC standard [12]. However, this solution neglects the fact that the different camera views show similar content. While it excessively exploits the correlation between the different time steps of the sequence, it does not benefit from the correlation between the different views.

Recently, several coding schemes have been proposed for multi-view video coding. They share the common idea to perform prediction not only from temporally preceding frames but also from neighbouring camera views. In [3], the authors suggested a simple resorting scheme, where the frames from all cameras are interleaved into a new, single video stream. The resulting sequence can then be compressed by applying an off-the-shelf video coder, e. g. H.264/AVC. If the chosen number of reference frames is large enough, the coder can use frames from other cameras as well as temporally preceding frames for prediction. However, besides the resorting of the input data, the video coder has not been adapted to multi-view video coding in particular.

Within the scope of a Call for Proposals in MPEG, several schemes have been suggested which extend the H.264/AVC standard in such a way that it makes use of the correlation between the various camera perspectives by introducing prediction across different views. It was shown that H.264/AVC-based solutions featuring prediction between views can achieve a better visual quality at the same bit rate compared to simulcast coding. Alternatively, multi-view video data can be coded at a lower bit rate with the same quality level [6].

In the following, multi-view prediction is analysed in a general fashion. For that, the distribution of MSE-optimum matches among possible reference frames is evaluated. Furthermore, the possible prediction gain is calculated which an adapted and optimised multi-reference prediction scheme could achieve. Based on the statistical results, a selection strategy is derived for the reference frames in a practical multi-view video coder.

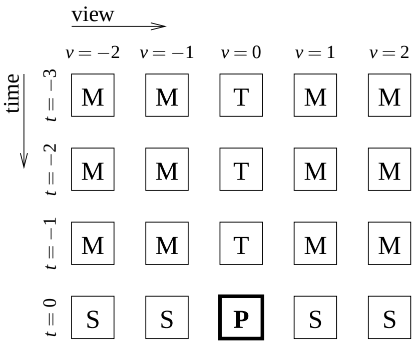

Figure 2. Prediction scheme assumed for the analysis

2. MULTI-REFERENCE PREDICTION

In this section, multi-reference prediction for multi-view sequences is introduced. We assume that a multi-view coder holds frames from a certain number of time steps for all camera views in its memory. This is shown in Fig. 2, where each rectangle corresponds to one frame currently stored in the memory. The vertical direction denotes the different time steps, while the horizontal direction corresponds to the different camera views. The view of the currently predicted image is denoted as $v = 0$, the preceding view as $v = -1$, the succeeding view as $v = 1$, and so on.

In the figure, the current frame to be predicted is marked as “P”-frame. The frame is divided into blocks, and for each of them, block matching is performed to find the reference block with the best match compared to the current block. For that, another frame (e. g. the temporally preceding frame) is chosen as a reference. Within a rectangular search area in the reference frame, the best matching block compared to the current block is searched. Each possible position in the search area can be specified by the displacement vector $\vec{d} = (d_1, d_2)^T$ between the current block and the reference block. As an optimisation criterion for finding the best reference, we use the minimum mean square error (MSE). It is defined as:

$$MSE\left(\vec{d}\right) = \frac{1}{N_1 N_2} \sum_{\vec{n} \in \mathcal{B}} \left[c\left(\vec{n}\right) - r\left(\vec{n} + \vec{d}\right)\right]^2, \qquad (3)$$

where $\mathcal{B}$ denotes a block of size $N_1 \times N_2$ and $\vec{n} = (n_1, n_2)^T \in \mathcal{B}$ a pixel position within that block. c denotes the values of the pixels in the current frame, r the pixels in the reference frame (see [15]).

The possible reference frames for the assumed prediction scheme can be divided into three categories, as indicated in Fig. 2:

- “T”-frames: temporally preceding frames from the same camera view as the currently predicted frame. These frames would also be available in a simulcast coding scheme, where each view is coded separately using a conventional video coder
- “S”-frames: frames from the same time step as the currently predicted frame, but from other camera views

Table 1. Parameters of the test sequences. For the Xmas sequence, only 10 of the original 101 views were used for complexity reasons.

Sequence	No. of views	No. of frames	Image Resolution	Total no. of blocks
Crowd	5	1 002	320×240	1 503 000
Flamenco1	8	624	320×240	1 497 600
Ballet	8	100	$1\,024 \times 768$	2 457 600
Breakdancers	8	100	$1\,024 \times 768$	2 457 600
Ballroom	8	250	640×480	2 400 000
Exit	8	250	640×480	2 400 000
Jungle	8	250	$1\,024 \times 768$	6 144 000
Uli	8	250	$1\,024 \times 768$	6 144 000
Xmas	10	101	640×480	1 212 000

Table 2. Results in absolute numbers

Sequence	Intra (I)	Temporal (T)	Spatial (S)	Mixed (M)
Crowd	10	1 242 162	150 142	110 686
Flamenco1	0	1 057 722	264 738	175 140
Ballet	0	1 990 258	199 443	267 899
Breakdancers	0	1 270 945	601 885	584 770
Ballroom	7	1 920 215	246 981	232 797
Exit	6	1 899 377	145 089	355 528
Jungle	0	5 821 375	99 284	223 341
Uli	0	5 400 084	189 514	554 402
Xmas	0	176 765	836 134	199 101

Table 3. Results in relative numbers

Sequence	Intra (I)	Temporal (T)	Spatial (S)	Mixed (M)
Crowd	0.00 %	82.65 %	9.99 %	7.36 %
Flamenco1	0.00 %	70.63 %	17.68 %	11.69 %
Ballet	0.00 %	80.98 %	8.12 %	10.90 %
Breakdancers	0.00 %	51.71 %	24.49 %	23.79 %
Ballroom	0.00 %	80.01 %	10.29 %	9.70 %
Exit	0.00 %	79.14 %	6.05 %	14.81 %
Jungle	0.00 %	94.75 %	1.62 %	3.64 %
Uli	0.00 %	87.89 %	3.08 %	9.02 %
Xmas	0.00 %	14.58 %	68.99 %	16.43 %

- “M”-frames: frames which do not fall in one of the other categories. As they have a temporal as well as a spatial offset to the currently predicted frame, they are referred to as “mixed” frames.

From all available frames, the frame delivering the overall minimum MSE is chosen as the optimal reference for the current block. If the minimum MSE is still larger than the energy of the current block, the current block is marked as intra coded (I), otherwise its prediction mode is marked as temporal (T), spatial (S) or mixed (M), respectively.

The described block matching scheme is performed for all blocks in all frames in all views of a multi-view sequence. The number of I-, T-, S- and M-blocks is counted and converted into percentages. These values give a rough feeling of the significance of the different references and the possible gain an optimised multi-view prediction scheme could achieve compared to simulcast coding, where only temporal prediction is possible [4].

It should however be noted that the resulting percentages need to be considered an upper bound, as not all the reference frames shown in Fig. 2 can be available for prediction in a practical coder due to causality reasons. For actually coding multi-view sequences, an appropriate, causal prediction scheme must be used. One example is the “group of group of pictures (GoGOP)” structure described in [8].

3. TEST SEQUENCES

For the analysis, different multi-view test data sets have been used, which have kindly been provided for the Multi-View Video Coding group within MPEG. They are listed together with their parameters in Table 1.

- “Crowd” and “Flamenco1” were generated by KDDI Corporation [7]. “Crowd” was recorded using a setup of five cameras which were aligned in the shape of a cross. For “Flamenco1”, eight cameras were used which were arranged in a horizontal line. This is also the arrangement of all the following test sequences. For both KDDI sequences, neighbouring cameras have a distance of 20 cm and the frame rate is 30 frames per second.
- “Ballet” and “Breakdancers” were provided by the Interactive Visual Media Group at Microsoft Research [21]. They each consist of eight camera views with a resolution of 1024×768 pixel and a frame rate of 15 frames per second.
- “Ballroom” and “Exit” originate from the Mitsubishi Electric Research Laboratories (MERL) [18]. They show eight views of the scenes at VGA resolution and 25 frames per second.
- “Jungle” and “Uli” were recorded by Fraunhofer HHI [1] with an arrangement of 8 cameras at a distance of about 20 cm. They have a resolution of 1024×768 pixel and a frame rate of 25 frames per second.
- “Xmas” is a very dense data set with 101 views at VGA resolution and was captured by Tanimoto Laboratory, Nagoya University [14]. The distance between neighbouring views is only 3 mm. This was achieved by using a single camera which was moved along the scene while recording. This leads to good calibration properties, but also to unnatural motion.

4. SIMULATION RESULTS

For the block matching tests, the block size was 16×16 pixel; the search range was set to 32. The resulting number of blocks for each sequence is shown in Table 1. Three pictures were used for prediction in the temporal direction, and all available views for prediction in the spatial direction. Furthermore, all “mixed” modes from all views in the last three time steps were searched.

The statistical results of the block matching test are shown in Table 2 using absolute numbers. In Table 3, the results are converted into percentages for better comparison.

For all tested sequences except the very dense Xmas sequence, the majority of the blocks is predicted best from temporally preceding frames, just like in simulcast coding. However, there is also a significant amount of blocks which is

Table 4. Mean square error (MSE) values after block matching using simulcast and multi-reference prediction

		Multi-Reference	
Sequence	**Simulcast**	**Prediction**	**Reduction**
Crowd	140.54	105.33	25.06 %
Flamenco1	50.53	23.50	53.49 %
Ballet	137.42	5.26	96.17 %
Breakdancers	79.76	10.79	86.47 %
Ballroom	85.71	26.15	69.50 %
Exit	46.00	14.42	68.64 %
Jungle	118.18	37.09	68.61 %
Uli	122.87	44.40	63.87 %
Xmas	263.69	2.95	98.88 %

Table 5. Peak-signal-to-noise ratio (PSNR) values after block matching using simulcast and multi-reference prediction

		Multi-Reference	Prediction
Sequence	**Simulcast**	**Prediction**	**Gain**
Crowd	26.65 dB	27.91 dB	1.25 dB
Flamenco1	31.10 dB	34.42 dB	3.32 dB
Ballet	26.75 dB	40.92 dB	14.17 dB
Breakdancers	29.11 dB	37.80 dB	8.69 dB
Ballroom	28.80 dB	33.96 dB	5.16 dB
Exit	31.50 dB	36.54 dB	5.04 dB
Jungle	27.41 dB	32.44 dB	5.03 dB
Uli	27.24 dB	31.66 dB	4.42 dB
Xmas	23.92 dB	43.43 dB	19.51 dB

better predicted from the spatial direction or from a mixed reference. For several sequences, the amount of “non-temporal” prediction is rather high (about 20 to 30 %), which leads to the assumption that for these sequences a significant coding gain could be achieved using a proper multi-view prediction scheme. For the Breakdancers sequence, this amount is even higher, and nearly half of the blocks is predicted best from spatial or mixed references. For two sequences (Jungle and Uli), in contrast, the amount is rather low.

Due to the high density of the Xmas sequence, the spatial correlation between the different views is very high, and the vast majority of blocks is predicted best from the spatial direction or from mixed references.

It can be seen that the amount of blocks which benefit from spatial and mixed prediction strongly depends on the parameters of the sequence. However, the results also depend on the content of the sequence, as even sequences having the same parameters show very different numbers. One example is given by the “Ballet” and “Breakdancers” sequence, which have the same parameters but show rather different results. Suboptimal camera calibration and different lighting between the views might have some effect here. It can therefore be desirable to compensate luminance and chrominance differences in a pre-filtering step. In [2], the authors suggested the usage of histogram matching to further improve the spatial prediction efficiency.

5. IMPROVEMENT IN MSE AND PSNR

In the last section, it could be shown that for a significant amount of blocks in the multi-view sequences, the remaining MSE can be reduced compared to simulcast, when prediction between different views is introduced. In this section, we analyse the quantity by which the MSE actually decreases when using multi-reference prediction. This is shown in Table 4 for the whole set of test sequences. The results are compared to simulcast, where only temporal (T) and intra (I) modes are available.

Table 6. Detailed results for the different prediction modes (Flamenco1 sequence). The probabilities for $|v| > 2$, which are very low, are not shown for clarity reasons.

$v =$	-2	-1	0	1	2
$t = 0$	0.88 %	7.30 %	0.00 %	7.04 %	0.93 %
$t = -1$	0.43 %	2.01 %	49.15 %	2.01 %	0.38 %
$t = -2$	0.26 %	0.92 %	12.19 %	0.97 %	0.26 %
$t = -3$	0.25 %	0.73 %	9.29 %	0.77 %	0.25 %

For nearly all sequences, the MSE is more than halved. That shows that the prediction performance is substantially better for multi-reference prediction than for simulcast. It should be noted that for the first time step in a sequence, I-frames are not longer necessary for each camera view when spatial prediction is introduced. This leads to a significant contribution in the MSE reduction.

As a more descriptive measure for the prediction performance, the PSNR after the block matching step is additionally given in Table 5. It can simply be calculated from the MSE value as:

$$PSNR \text{ [dB]} = 10 \log_{10} \frac{255^2}{MSE} \tag{4}$$

Assuming that the mean value for the residual error signal is zero, the difference between the PSNR values for multi-reference prediction and for simulcast is equivalent to the prediction gain G:

$$\begin{aligned} G \text{ [dB]} &= 10 \log_{10} \frac{\sigma^2_{\text{Simulcast}}}{\sigma^2_{\text{Multi-Reference}}} \\ &= 10 \log_{10} \frac{MSE_{\text{Simulcast}}}{MSE_{\text{Multi-Reference}}} \\ &= PSNR_{\text{Multi-Reference}} - PSNR_{\text{Simulcast}} , \end{aligned} \tag{5}$$

where σ^2 is the variance of the residual error signal [16], [11]. The values in Table 5 indicate that a prediction gain of several dB is possible by exploiting the correlation between the different views. For most sequences, the gain is about 3 to 5 dB.

6. DETAILED STATISTICS AND SELECTION OF PREDICTION MODES

In Table 6, the statistics for the different prediction modes are shown in detail. The results are exemplarily shown for the Flamenco1 sequence. The table contains the relative frequency of the possible references. The parameter t denotes the different time steps, where $t = 0$ ist the current time step, $t = -1$ the last time step, and so on. v denotes the different camera views in relation to the view of the currently predicted frame. Therefore, $v = -1$ means the view preceding the current frame, $v = 1$ the view succeeding the current

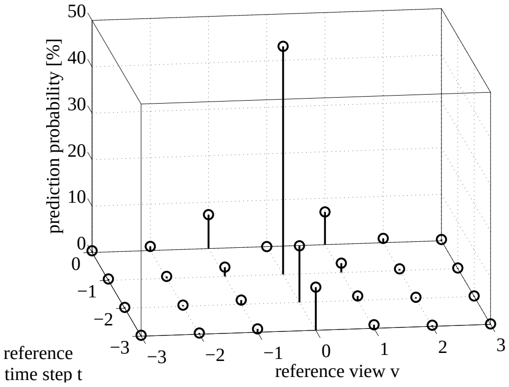

Figure 3. Detailed prediction statistics for the Flamenco1 sequence

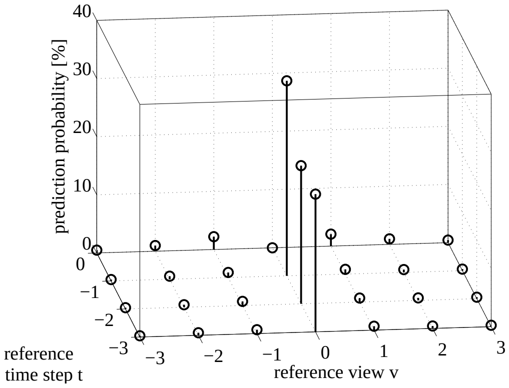

Figure 4. Detailed prediction statistics for the Ballet sequence

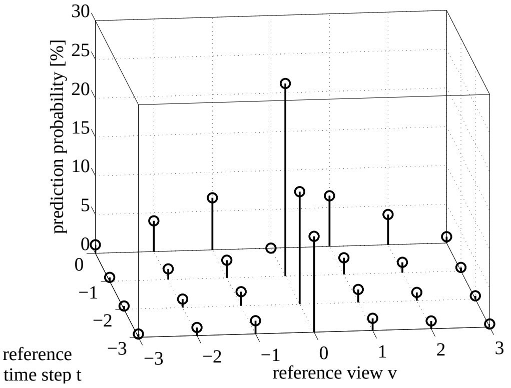

Figure 5. Detailed prediction statistics for the Breakdancers sequence

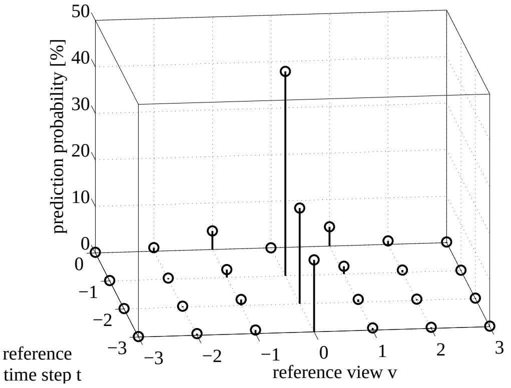

Figure 6. Detailed prediction statistics for the Ballroom sequence

frame, and so on (see also Fig. 2). For $t = 0$, $v = 0$, the probability for the intra (I) mode is listed. The plot in Fig. 3 further illustrates the results.

One can see that the most important references are the temporally preceding frames (where $v = 0$) and the immediately neighbouring frames from the same time step ($t = 0$, $v = \pm 1$). But also the frames originating from immediately neighbouring cameras and from the last time step ($t = -1$, $v = \pm 1$) are chosen with a significant likelihood (here: 2.01 %). Those frames are more likely than the frames with $t = 0$, $v = \pm 2$.

Fig. 4 to 6 show the detailed results for some of the remaining sequences. It can be seen that the mixed modes are generally chosen with small probabilities compared to the temporal and spatial modes. Although the overall probability for mixed mode prediction (see Table 3) is significant, the percentage for each of the single modes is small due to the large number of existing modes.

Searching all the modes which were included in the prediction scheme may well be far too time-consuming for a practical coder. Therefore, a suitable subset of the modes needs to be chosen which are then actually considered for prediction. The results of this analysis indicate that it might be suitable to search only temporal and spatial reference frames and to skip the mixed modes completely. If mixed mode prediction is applied, the modes with $t = -1$ and $v = \pm 1$ should be included in the search. Considering the spatial references, it is in most cases sufficient to search only the immediately neighbouring frames ($t = 0$, $v = \pm 1$). In very few cases, such as for the Breakdancers sequence, also the frames with $t = 0$, $v = \pm 2$, are useful. Only for the very dense Xmas sequence, which represents a rather untypical case, views with greater distances to the current view play an important role.

7. BLOCK MATCHING WITH TEMPORAL AND SPATIAL REFERENCES ONLY

In the last section, it was argued that in most cases, it might be suitable not to predict from mixed modes, as the probabilities for each of them are rather low. As a verification for this assumption, a simplified block matching scheme is analysed in this section, where only temporal and spatial references are considered. This new scheme is depicted in Fig. 7. The number of frames which have to be searched in the block matching step is highly reduced, and therefore, the complexity significantly decreases. For the Flamenco1 sequence, for example, the simulation time decreases by about two thirds of the original time.

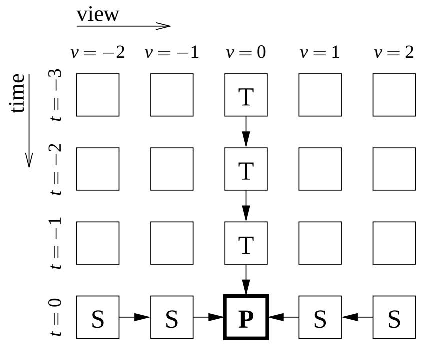

Figure 7. Modified prediction scheme with temporal and spatial references only

Table 7. Results in relative numbers when only temporal and spatial references are considerd

Sequence	Intra (I)	Temporal (T)	Spatial (S)
Crowd	0.00 %	86.20 %	13.80 %
Flamenco1	0.00 %	77.43 %	22.57 %
Ballet	0.00 %	87.20 %	12.80 %
Breakdancers	0.00 %	63.70 %	36.30 %
Ballroom	0.00 %	84.91 %	15.09 %
Exit	0.00 %	87.26 %	12.74 %
Jungle	0.00 %	96.13 %	3.87 %
Uli	0.00 %	91.67 %	8.33 %
Xmas	0.00 %	19.17 %	80.83 %

Table 7 shows the probabilities for temporal and spatial prediction. Comparing this table to the results with mixed mode prediction in Table 3, it can be noticed that the probability for temporal prediction has increased by a few percent. This means that the amount of non-temporal prediction and therefore the number of blocks which benefit compared to simulcast has reduced. However, the reduction is rather small compared to the benefit in terms of complexity which is achieved by omitting the large number of mixed references in the block matching step.

Table 8 and Table 9 contain the results in terms of MSE and PSNR after the block matching step. Let us compare the PSNR results to the PSNR results with mixed mode prediction (see also Table 9). Naturally, the PSNR is reduced without mixed mode prediction, but the reduction is rather small: between 0.04 dB for the Ballet sequence and 0.61 dB for the Ballroom sequence. For most of the sequences, the reduction is about 0.1 dB to 0.2 dB.

In Fig. 8 and 9, the prediction probabilities for the temporal and spatial prediction modes are shown in detail. The assumption that the temporal references are the most important ones still holds. For spatial prediction, the immediately neighbouring views still are of high significance, while the probabilities for the remaining views are rather low.

8. RATE-DISTORTION OPTIMISED SELECTION OF REFERENCE FRAMES

For a practical video coder, it is appropriate to assign short codewords to modes which are likely and longer codewords to modes which are not as likely. This has the effect that the modes with large probabilities are coded spending a smaller number of bits, and the more unlikely modes need a higher number of bits.

It is desirable to perform the mode selection in a rate-distortion optimised way [19]. That means that the prediction mode is chosen which minimises the value

$$J = D + \lambda R. \qquad (6)$$

The distortion D can e. g. be represented by the MSE value, while the length of the assigned codeword contributes to the rate R. In that case, the modes with larger percentages will be chosen with an even higher probability because they have an advantage in bit rate over the more unlikely modes. That is why the given considerations on mode selection will be even more distinct in a rate-distortion optimised coding scheme.

Table 8. Mean square error (MSE) values after block matching using simulcast and prediction with temporal and spatial references

Sequence	Simulcast	T/S-Prediction	Reduction
Crowd	140.54	109.69	21.95 %
Flamenco1	50.53	24.37	51.78 %
Ballet	137.42	5.31	96.14 %
Breakdancers	79.76	11.06	86.13 %
Ballroom	85.71	30.07	64.92 %
Exit	46.00	15.10	67.19 %
Jungle	118.18	37.76	68.05 %
Uli	122.87	45.54	62.94 %
Xmas	263.69	2.98	98.87 %

Table 9. Peak-signal-to-noise ratio (PSNR) values after block matching using simulcast and prediction with temporal and spatial references. For comparison, the PSNR loss compared to prediction with mixed modes (see Table 5) is also shown.

Sequence	Simulcast	T/S-Prediction	Pred. Gain	Loss without Mixed Modes
Crowd	26.65 dB	27.73 dB	1.08 dB	0.18 dB
Flamenco1	31.10 dB	34.26 dB	3.17 dB	0.16 dB
Ballet	26.75 dB	40.88 dB	14.13 dB	0.04 dB
Breakdancers	29.11 dB	37.69 dB	8.58 dB	0.11 dB
Ballroom	28.80 dB	33.35 dB	4.55 dB	0.61 dB
Exit	31.50 dB	36.34 dB	4.84 dB	0.20 dB
Jungle	27.41 dB	32.36 dB	4.95 dB	0.08 dB
Uli	27.24 dB	31.55 dB	4.31 dB	0.11 dB
Xmas	23.92 dB	43.38 dB	19.46 dB	0.05 dB

9. SUMMARY AND CONCLUSIONS

Multi-reference prediction, which can be applied in the coding of multi-view video sequences, was statistically analysed. Block matching was introduced where the video coder can not only use temporally preceding frames, but also references from other camera views to find the best matching block. The possible prediction modes were divided into temporal, spatial and mixed modes. Simulations were performed using several different multi-view test data sets. For each sequence, the number of temporal, spatial and mixed references was counted.

From the results, it could be seen that a significant amount of blocks was predicted best from spatial or mixed refer-

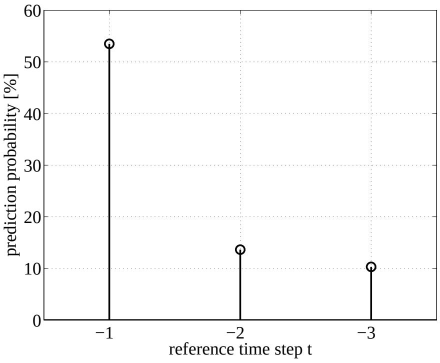

Figure 8. Probabilities for the different temporal prediction modes (Flamenco1 sequence)

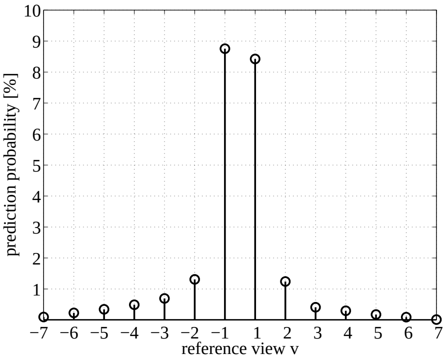

Figure 9. Probabilities for the different spatial prediction modes (Flamenco1 sequence)

ences. For these blocks, a gain can be achieved compared to simulcast coding, where each camera view is coded separately and only temporal prediction is possible. For most of the sequences, a significant gain in MSE and in PSNR — typically about 5 dB — was observed after the block matching step when compared to simulcast.

The probabilities for all different references were given. As they were low for the mixed modes, a simplified prediction scheme was introduced where only temporal and spatial prediction was considered. This leads to a large reduction in complexity, while the PSNR loss is only about 0.1 dB to 0.2 dB compared to the original scheme.

From the detailed statistics of the prediction modes, a conclusion could be drawn about the reasonable selection of reference frames in a practical coder. In most cases, it is sufficient to search the temporal references together with the immediately neighbouring frames from the same time step. The probability that one of the other reference frames delivers the best match is very low, and this effect will even be intensified when rate-distortion optimisation is performed. Therefore, the time needed for the block matching step in a practical multi-view video coder can be limited.

ACKNOWLEDGEMENT

This work was funded by the German Research Foundation (DFG) within the Collaborative Research Centre “Model-based analysis and visualisation of complex scenes and sensor data” under grant SFB 603/TP C8. Only the authors are responsible for the content.

REFERENCES

[1] Daase, J., Goelz, U., Kauff, P., Mueller, K., Schreer, O., Smolic, A., Tanger, R. and Wiegand, T., “Fraunhofer HHI Test Data Sets for MVC,” in: *ISO/IEC JTC1/SG29/WG11, Document MPEG2005/M11894*, Busan, Korea (2005).

[2] Fecker, U., Barkowsky, M. and Kaup, A., “Improving the Prediction Efficiency for Multi-View Video Coding Using Histogram Matching,” in: *Picture Coding Symposium (PCS 2006)*, Beijing, China (2006).

[3] Fecker, U. and Kaup, A., “H.264/AVC-Compatible Coding of Dynamic Light Fields Using Transposed Picture Ordering,” in: *Proc. 13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey (2005).

[4] Fecker, U. and Kaup, A., “Statistical Analysis of Multi-Reference Block Matching for Dynamic Light Field Coding,” in: *Proc. 10th International Fall Workshop Vision, Modeling, and Visualization (VMV 2005)*, pp. 445–452, Erlangen, Germany (2005).

[5] Gortler, S. J., Grzeszczuk, R., Szeliski, R. and Cohen, M. F., “The Lumigraph,” in: *Proc. SIGGRAPH 96*, pp. 43–54, New Orleans, Louisiana, USA (1996).

[6] ISO/IEC JTC1/SG29/WG11, “Subjective Test Results for the CfP on Multi-view Video Coding,” Document MPEG2005/N7779, Bangkok, Thailand (2006).

[7] Kawada, R., “KDDI Multiview Video Sequences for MPEG 3DAV Use,” in: *ISO/IEC JTC1/SG29/WG11, Document MPEG2005/M10533*, Munich, Germany (2004).

[8] Kimata, H., Kitahara, M., Kamikura, K. and Yashima, Y., “Multi-View Video Coding Using Reference Picture Selection for Free-Viewpoint Video Communication,” in: *Picture Coding Symposium (PCS 2004)*, San Francisco, CA, USA (2004).

[9] Levoy, M. and Hanrahan, P., “Light Field Rendering,” in: *Proc. SIGGRAPH 96*, pp. 31–42, New Orleans, Louisiana, USA (1996).

[10] Magnor, M., *Geometry-Adaptive Multi-View Coding Techniques for Image-based Rendering*, vol. 21 of *Berichte aus der Kommunikations- und Informationstechnik*, Shaker, Aachen, Germany (2001).

[11] Ohm, J.-R. and Lüke, H. D., *Signalübertragung: Grundlagen der digitalen und analogen Nachrichtenübertragungssysteme*, 8th ed., Springer, Berlin, Germany (2002).

[12] Sullivan, G. J., Topiwala, P. and Luthra, A., “The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions,” in: *SPIE Conference on Applications of Digital Image Processing XXVII*, Denver, CO, USA (2004).

[13] Tanimoto, M., “Free Viewpoint Television — FTV,” in: *Picture Coding Symposium (PCS 2004)*, San Francisco, CA, USA (2004).

[14] Tanimoto, M. and Fujii, T., “Test Sequence for Ray-Space Coding Experiments,” in: *ISO/IEC JTC1/SG29/WG11, Document MPEG2003/M10408*, Hawaii, USA (2003).

[15] Tekalp, A. M., *Digital Video Processing*, Prentice Hall PTR, Upper Saddle River, NJ, USA (1995).

[16] Vary, P., Heute, U. and Hess, W., *Digitale Sprachsignalverarbeitung*, B. G. Teubner, Stuttgart, Germany (1998).

[17] Vetro, A., Matusik, W., Pfister, H. and Xin, J., “Coding Approaches for End-to-End 3D TV Systems,” in: *Picture Coding Symposium (PCS 2004)*, San Francisco, CA, USA (2004).

[18] Vetro, A., McGuire, M., Matusik, W., Behrens, A., Lee, J. and Pfister, H., “Multiview Video Test Sequences from MERL,” in: *ISO/IEC JTC1/SG29/WG11, Document MPEG2005/M12077*, Busan, Korea (2005).

[19] Wiegand, T., *Multi-Frame Motion-Compensated Prediction for Video Transmission*, vol. 18 of *Berichte aus der Kommunikations- und Informationstechnik*, Shaker, Aachen, Germany (2000).

[20] Zhang, C. and Chen, T., “A survey on image-based rendering — representation, sampling and compression,” *Signal Processing: Image Communication* 19 (1), pp. 1–28 (2004).

[21] Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S. and Szeliski, R., “High-Quality Video View Interpolation Using a Layered Representation,” in: *ACM SIGGRAPH and ACM Transactions on Graphics*, pp. 600–608, Los Angeles, CA, USA (2005).