J. Vis. Commun. Image R. 19 (2008) 589–599

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Side information estimation and new symmetric schemes for multi-view distributed video coding ☆

Thomas Maugey *, Béatrice Pesquet-Popescu

Signal and Image Processing Department, Ecole Nationale Supérieure des Télécommunications (ENST), 46 rue Barrault, 75634 Paris Cedex 13, France

ARTICLE INFO

Article history:
Received 30 November 2007
Accepted 8 September 2008
Available online 19 September 2008

Keywords:
Multi-view
Distributed video coding
Side information
Fusion
Rate-distortion analysis
Wyner-Ziv
Motion interpolation
Long-term estimation

ABSTRACT

This paper deals with distributed video coding (DVC) for multi-view sequences. DVC of multi-view sequences is a recent field of research, with huge potential impact in applications such as videosurveillance, real-time event streaming from multiple cameras, and, in general, immersive communications. It raises however several problems, and in this paper we tackle two of them. Based on the principles of Wyner–Ziv (WZ) coding, in multi-view DVC many estimations can be generated in order to create the side information (SI) at the decoder. It has been shown that the quality of the SI strongly influences the global coding performances. Therefore, this paper proposes to study the contribution of multiple SI estimations (in the temporal and view directions) to the global performances. Moreover, we propose new symmetric schemes for longer group of pictures (GOP) in multi-view DVC and show that we can further exploit the long-term correlations using a new kind of estimation, called diagonal. For such schemes, several decoding strategies may be envisaged. We perform a theoretical study of the temporal and inter-view dependencies, and confirm by experiments the conclusion about the best decoding strategy.

1. Introduction

A new paradigm in video coding is the distributed video coding (DVC) which allows to move the computation complexity from the encoder to the decoder. This can be interesting in many applications, whenever a compression is due to be done with a light hardware. This new coding paradigm is based on two results of information theory appeared in the 70's [1,2], which show that with two correlated sources encoded independently and decoded jointly one can achieve the same performances as with the two sources encoded and decoded jointly. Therefore, the correlation between the frames is not exploited anymore at the encoder but only at the decoder. Theoretically, the performances of such a method can attain the ones of classical hybrid codecs (H.263, MPEG-4, H.264, etc.). In practice, for the moment, the performances are still below even though, in most cases, DVC performs better than Intra coding [3–5]. In this paper, we propose to study the distributed video coding of multi-view sequences. This recent field of research is very interesting for a wide range of applications, such as videosurveillance, real-time event streaming from multiple cameras, and immersive communications in general. We propose to tackle two of the numerous problems raised by DVC of multi-view sequences. In Wyner–Ziv coding, one or several estimations are generated at the decoder in order to build the side information (SI). It is proven that the general coding performances highly depend on the quality of this SI. We thus propose to study the contribution of multiple existing estimations for building the SI, and propose a new kind of estimation exploiting further dependencies. Moreover, we propose new symmetric schemes, in which many decoding strategies may be conceivable. We thus perform a theoretical study based on temporal and inter-view dependencies in order to propose the best decoding strategy.

After presenting a short summary of the theoretical distributed source coding framework in Section 2, we expose the main features of DVC implementation, the problems and their existing answers in Section 3. In Section 4, we introduce a new model for rate-distortion analysis of video sequences. In Section 5, we analyze the ideal estimation achieved by side information in a multi-view setting and we then propose new symmetric schemes with longer GOPs, based on the previous proposed model. This is followed by experimental results in Section 6, and finally the conclusion and future work are drawn in Section 7.

2. Theoretical results

2.1. Slepian and Wolf

In 1973, Slepian and Wolf [1] studied the performances of the transmission of two correlated sources X and Y in many cases,

☆ Part of this work has been funded by the French ANR project no. ANR-FI-071215-01-01 (ESSOR).

* Corresponding author.
E-mail addresses: maugey@telecom-paristech.fr (T. Maugey), pesquet@telecom-paristech.fr (B. Pesquet-Popescu).

1047-3203/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2008.09.002

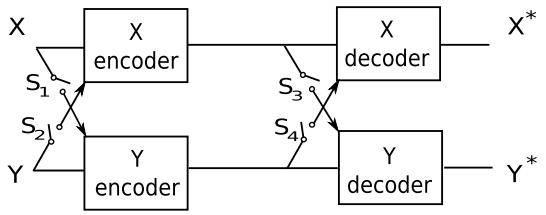

Fig. 1. Sixteen cases of encoding-decoding, depending on the ON/OFF position of the connections S_1 to S_4.

summarized in Fig. 1. The connections S_1 to S_4 can be open or closed, depending on whether the information is known or not by the encoder or decoder. In each of the sixteen possible cases, Slepian and Wolf found the performances of the coding scheme, and plot the admissible rate region $\mathscr{R}$. Among these sixteen cases, two can be highlighted: the one which corresponds to the classical coding scheme (Fig. 2 left) and the one which corresponds to the distributed coding scheme (Fig. 2 right). These results bring the fundamental and surprising conclusion that the performances can be the same if the two sources are encoded jointly (S_1 and S_2 closed) or not (S_1 and S_2 open). As described in Section 2.2, Wyner and Ziv extended the previous results to the case of lossy transmission.

2.2. Wyner and Ziv

In similar conditions (one source X to transmit with side information Y at the encoder and/or the decoder) Wyner and Ziv [2] studied the rate-distortion function in the cases of classical coding, $R_{X|Y}(d)$, and distributed coding, $R^*(d)$. Consider the following obvious inequality:

$$R_{X|Y}(d) \leqslant R^*(d) \tag{1}$$

They found the expression of the rate-distortion function and showed that the inequality (1) was usually strict but in some particular cases, like for example for Gaussian sources, the equality could be proven:

$$R_{X|Y}(d) = R^*(d) \tag{2}$$

meaning that the distributed source coding can achieve the same performances as classical coding schemes for lossy transmission, too.

The rate-distortion function $R^*(d)$ was studied from an experimental point of view in a practical case of DVC in [6,7]. Under a state-space model using Kalman filtering, the rate-distortion function is determined and is used to theoretically study the efficiency of DVC.

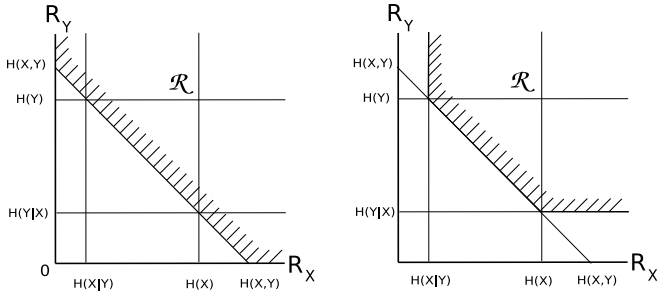

Fig. 2. Admissible rate regions $\mathscr{R}$ for: (left) classical coding with joint encoding and joint decoding, (right) distributed coding with separated encoding and joint decoding.

In our work, we will also make a rate-distortion analysis, involving the temporal and view correlations in a multi-view scheme, in order to find the best decoding strategy for a new multi-view scheme. This is described in Section 4.

3. Distributed video coding

The above theoretical results have been brought into practice in video coding much later [5,8], when the above principles have been applied to the case of mono-source and multi-source video encoding. We summarize in this section the main ideas behind the two settings.

3.1. Mono-source case

From one video, one first has to create two correlated sources. The frames are thus separated in two groups: the key frames (KFs) and the Wyner–Ziv frames (WZFs). The Groups of Pictures (GOPs) have generally the following structure [3]: one KF, n WZF. Here, n is often fixed, but some papers [9] propose methods in which n can vary in order to find the optimal value.

The KFs are encoded and decoded with a classical Intra codec. At the decoder, the decoded KF contribute to generate an estimation, called side information (SI), of the WZF in the middle of the GOP. The WZF processing is different. First a spatial transformation is applied (very often, a 4×4 DCT or integer DCT). This operation is optional and not performed in all practical DVC schemes, but the results in the DCT-domain are often better than in the pixel-domain [3,10]. Then the frame is quantized. This step is the link between the Slepian–Wolf and Wyner–Ziv theories, because it is where the loss appears. Afterwards, the frame is channel encoded using performant channel codes such as turbo codes [3] or LDPC [11]. At the output of this channel coding, only the parity bits are sent. That is the key idea of practical distributed coding. It is assumed that the estimation error at the decoder (between the generated SI and the original frame) can be assimilated to a channel error, so that the parity bits, introducing the redundancy necessary for correcting channel errors, correct in fact the estimation error through a channel decoder and a reconstruction step. This channel decoding is done iteratively by estimating the error probability of the decoded frame. If that probability is too high, the decoder asks more parity bits to the buffer, through a backward channel. This process is done until the error probability is low enough. Another solution is to estimate at the encoder the number of parity bits to send [12]. Finally, an inverse transform is applied to the reconstructed frame. The adopted scheme is illustrated in Fig. 3 and is based on a turbo coder with a backward channel. The experiments in Section 6 are run with this coding scheme.

Analyzing this encoding/decoding process, one can see that the complexity was shifted from the encoder to the decoder. Indeed, H264 intra encoder is lighter than any inter encoder, and WZ encoding is even more light because the coding process only consist of a transformation, a quantization and a very fast channel encoding. This last step is less complex than the entropy coding used by intra source encoders. The reconstruction is, on the contrary, much more complex, relying on an iterative channel decoding.

3.2. DVC of multi-view sequences

For multi-view sequences, the coding process is very similar to the mono-source case. The frames of each view are also separated in KFs and WZFs. The former ones are Intra coded, while the latter ones are Wyner–Ziv coded. Having many cameras brings a new estimation direction, in addition to the temporal one: the view

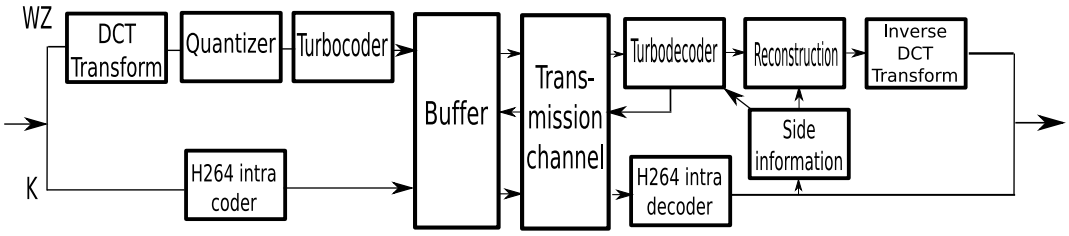

Fig. 3. Adopted DVC scheme.

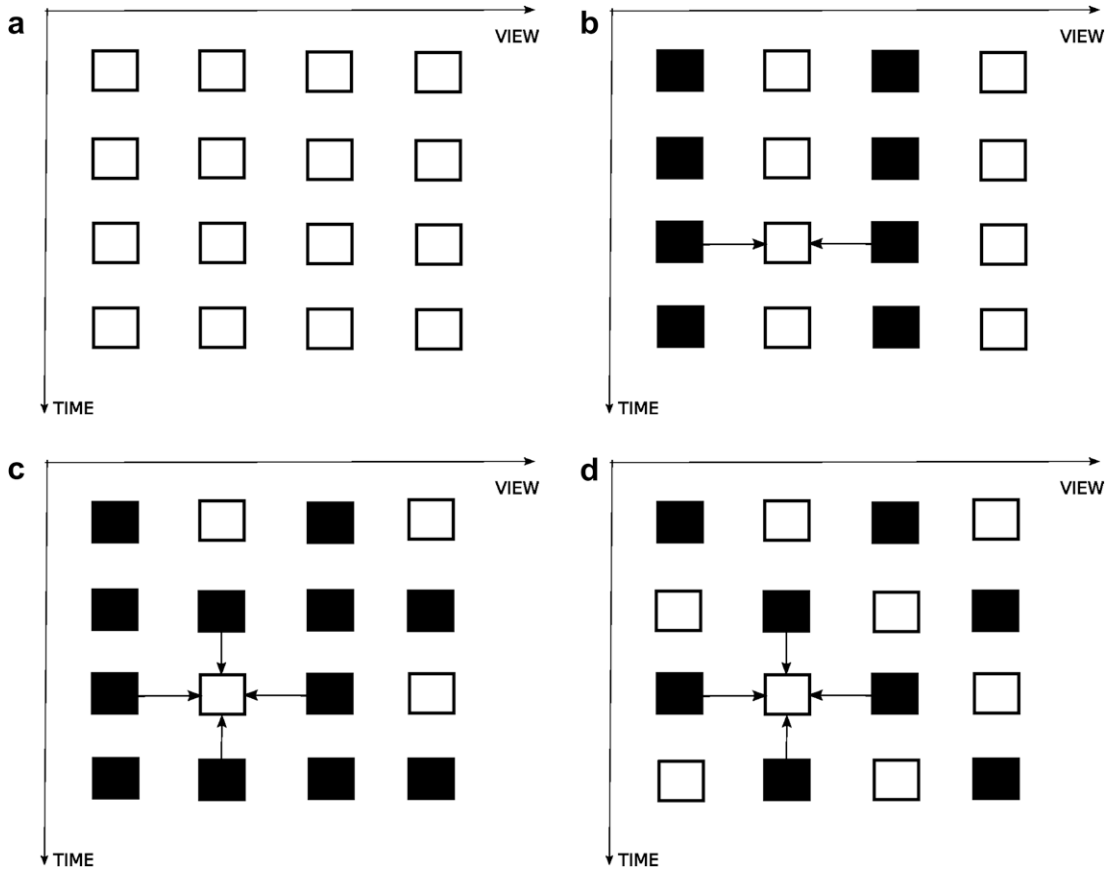

Fig. 4. Frame disposition in the time-view space for different schemes. (a) Time-view space representation. (b) The asymmetric scheme. (c) The hybrid 1/2 scheme. (d) The symmetric 1/2 scheme. KFs are in black, WZFs in white. Arrows indicate the directions for generating the SI.

axis. As illustrated in Fig. 4(a), this raises new problems and brings additional challenges for multi-view DVC. Indeed, though this figure is a simple representation of the disposition of frames in the time-view space, we notice that choosing the repartition of the WZFs and KFs is more complicated than in the mono-view case, where the only degree of freedom is the length of the GOP. The side information is not anymore generated only along the temporal direction, but also between the different views. So for one WZF, many estimations can be generated, and another problem is the method to generate an estimation between different views. A last issue, which will not be tackled in this work, is the fusion of these estimations to create a unique side information.

First, we summarize the different problems, and see what are the solutions proposed in the literature.

3.2.1. The scheme

The first problem to solve is to set the positions of the KFs and the WZFs. The way of generating side information also depends on this disposition. We can think about many strategies, but in the literature the solutions are not so numerous and can be classified in three categories. Before enumerating the three kinds of schemes, it is important to classify the cameras. There are:

- *Key cameras:* all of their frames are KFs. They can be encoded with an Intra coder but also with an Inter coder, involving only frames from other Key cameras. Anyway, these cameras need to be more powerful.
- *Wyner–Ziv cameras:* all of their frames are WZFs. The side-information for them is built based on the KFs of the other cameras. These cameras can be less powerful.
- *Hybrid cameras:* their frames can be KFs or WZFs. The side-information is built thanks to the KFs of the other cameras and also thanks to their own KFs. The advantage of using this type of cameras is that the problem becomes symmetric, all the cameras in the system can be identical.

Using all these types of cameras, many possible settings are conceivable. The following shows the one present in the literature. In Fig. 4(b–d), the KFs are in black and the WZFs in white. Three main schemes appear:

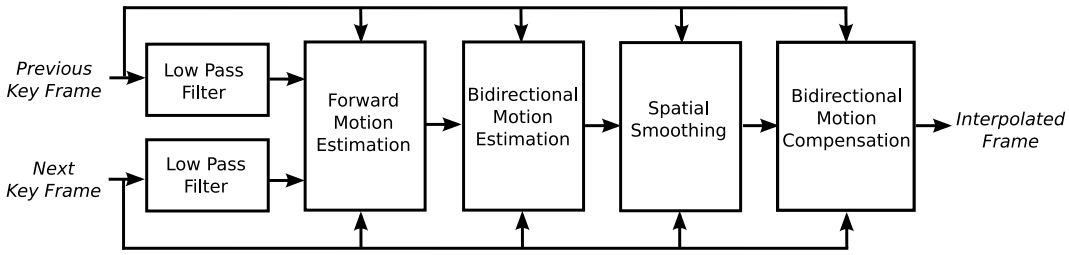

Fig. 5. Frame interpolation architecture proposed in [4,16].

[◆] The *asymmetric scheme* (AS): the type of cameras alternates between Key and Wyner–Ziv, as shown in Fig. 4(b). Then the side-information is built using the closest frames in the view direction. This principle is used for example in [13].
[◆] The *hybrid 1/2 scheme* (Hyb2): one camera over two is a Key camera and between them, there are hybrid cameras. The scheme $K-H-K-H-\ldots$ is illustrated in Fig. 4(c). In this case, the side-information can be estimated in the temporal and in the view direction. Then the problem of fusion of the two estimations appears. This scheme was proposed for example in [13,14].
[◆] The *symmetric 1/2 scheme* (Sym2): the cameras are all hybrid with one KF for one WZF. This $H-H-H-H$ scheme is presented in Fig. 4(d). There is a shift in the role of the cameras, the KFs and the WZFs being placed on a quincunx grid in the time-view axes. The side-information for each WZF can be then computed in the view direction and in the time direction. This case also has to cope with the fusion problem. It was proposed in [3,15].

3.2.2. Side information generation

[◆] Temporal estimation: in the literature, many solutions are proposed for the interpolation between two frames. Most of them are motion estimation based. The underlying assumptions are the ones made in inter codecs when estimating the B frames: the motion between successive frames is supposed uniform and the motion vectors between the WZF and the KF can be deducted from the motion vectors between two KF. The performances of distributed coding highly depend on the quality of side information. That is why we have to give a particular attention to these interpolation methods. In our codec we use the estimator proposed by Ascenso et al. in [4,16–18], shown in Fig. 5. First, the KFs are low-pass filtered, then a block-based motion estimation is performed between them. Then, the motion vectors obtained in the previous step are refined by using a bidirectional motion estimation scheme. The vector field is then smoothed with a median filter.
[◆] Inter-view estimation: many solutions have been proposed in the literature [19,13]. For the moment we use the above temporal interpolation method also for the inter-view estimation. It is not the best disparity estimation method, but the decoder is simpler with one single estimation method.

3.2.3. The fusion

The last problem to solve is the fusion. When many estimations are generated for one WZF, how can be created a unique side information? The purpose is to estimate an image I which is a matrix of $n \times m$ pixels. Let us assume that we have several estimations $\hat{I}_i$ of I, $i \in \{1,\ldots,s\}$. Let E_i be the error matrices:

$$\forall i \in \{1,\ldots,s\}, E_i(k,l) = |I(k,l) - \hat{I}_i(k,l)|, \quad \forall k \in \{0,\ldots,n-1\}, \quad \forall l \in \{0,\ldots,m-1\} \tag{3}$$

The purpose of the fusion is to find for each pixel the best estimation. We denote the final estimation by $\hat{I}$. Ideally, it is obtained as:

$$\forall k \in \{0,\ldots,n-1\}, \quad \forall l \in \{0,\ldots,m-1\}, \qquad \hat{I}(k,l) = \hat{I}_i(k,l) \tag{4}$$

such as:

$$E_i(k,l) = \min_{j \in \{1,\ldots,s\}} (E_j(k,l))$$

The difficulty in practice is to estimate the matrices E_i, because the original frame is not available at the decoder. In order to estimate the error matrices, different methods have been proposed in the literature [13,14,20] and they can be classified in two categories:

[◆] A first category tries to replace the original frame I in Eq. (3). For instance, the method proposed in [13] uses the previous or next frames instead of the original frame. Other methods have been proposed in [14], using the frames available from the closest cameras.
[◆] The second category of methods estimates directly the error matrices by exploiting a certain quality criterion. For example, in [13], the criterion can be based on the fact that the estimation is not good when the motion is too quick, so we can set a threshold for the motion vectors in order to locate where the motion estimation failed. In [20], the proposed methods compute the mask of the previously decoded WZF and directly uses this mask or a motion compensated mask to estimate the current mask.

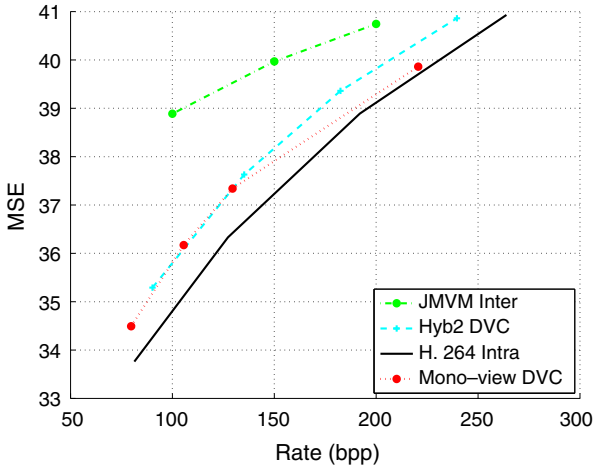

Fig. 6. Comparison of the rate-distortion performances for the "Ballet" sequence of an intra codec (H.264), an inter codec (JMVM), and a distributed codec (Hyb2).

As for the mono-source case, the performances of multi-view DVC do not reach the ones of classical inter codecs such as JMVM [21]. In Fig. 6, we present experimental results run with the "Ballet" sequence, for H.264 intra-frame codec (all cameras are intra, and all the frames are intra), JMVM inter codec, mono-view distributed video codec applied independently on each view (the mono-view DVC curve represents an average of the seven single camera curves), and Hyb2 multi-view distributed video codec presented above. One can see that the multi-view DVC performs better than intra coding but is not efficient enough to reach the inter coding performances. One can also remark that it is worth taking into account the inter-view dependencies in DVC, but only at medium and high bitrate. Indeed, at lower bitrate, the quality of the KFs is too low to allow a good estimation of the disparity (using the block-based method presented above).

4. Rate-distortion model

In this section, a rate-distortion model for video sequences is introduced. Based on classical assumptions and on fundamental results of information theory [22–24], our model yields an interesting expression of frame estimation error which is separated in two independent terms and permits to make a recursive analysis of error propagation in closed-loop predictive schemes.

We set the problem as illustrated in Fig. 7. The frames A and G are the original KFs used to generate the SI, $\widetilde{A}$ and $\widetilde{G}$ are their quantized versions. We denote by $\bar{A}$ and $\bar{G}$ the quantized motion or disparity compensed KFs. We assume that the generated SI is the linear combination between $\bar{A}$ and $\bar{G}$. The weighting coefficients in this combination, $k_A \in [0,1]$ and $k_G \in [0,1]$, depend on the distance (in number of frames) to the KFs A and G, and are such that $k_A + k_G = 1$. If d_{KF} is the distance between the two KFs A and G, and d_A (respectively, d_G) the distance between the estimated WZF B and the KF A (respectively, G), then the weights are:

$$k_A = \frac{d_{KF} - d_A}{d_{KF}}, \quad k_G = \frac{d_{KF} - d_G}{d_{KF}} \tag{5}$$

We denote by e_B the estimation error for the B frame.

First, let us calculate the variance of e_B. In this calculation we notice that for a vector $\boldsymbol{p} = (n,m)$ corresponding to the pixel in line n and column m, the motion compensated frame reads $\bar{A}(\boldsymbol{p}) = \widetilde{A}(\boldsymbol{p} - d\boldsymbol{p})$ (the same remark can be done for $\bar{G}(\boldsymbol{p})$) where $d\boldsymbol{p}$ is the motion vector associated with the position $\boldsymbol{p}$ in frame $\bar{A}$. We assume that the motion estimation error and the quantization error are independent. Then we have:

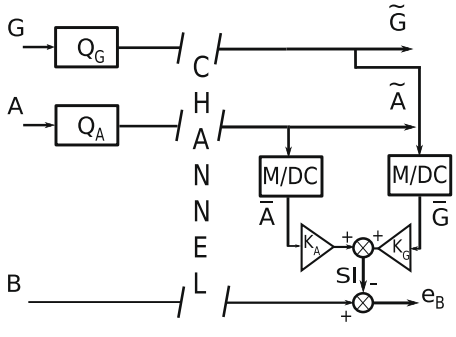

Fig. 7. Problem statement: A, G are the KFs, $\widetilde{A}$ and $\widetilde{G}$ their quantized versions, and $\bar{A}$ and $\bar{G}$ are the motion/disparity compensated (M/D C) frames, generating the SI for the frame B.

$$\begin{aligned}\sigma_{e_B}^2 &= \mathrm{E}\left[\left(B(\boldsymbol{p}) - k_A\bar{A}(\boldsymbol{p}) - k_G\bar{G}(\boldsymbol{p})\right)^2\right] \\ &= \mathrm{E}\left[\left(B(\boldsymbol{p}) - k_A\widetilde{A}(\boldsymbol{p} - d_1\boldsymbol{p}) - k_G\widetilde{G}(\boldsymbol{p} - d_2\boldsymbol{p})\right)^2\right].\end{aligned}$$

Writing again this expression as:

$$\sigma_{e_B}^2 = \mathrm{E}\left[\left(\underbrace{B(\boldsymbol{p}) - k_A A(\boldsymbol{p} - d_1\boldsymbol{p}) - k_G G(\boldsymbol{p} - d_2\boldsymbol{p})}_{\text{motion estimation error}} + \underbrace{k_A(A(\boldsymbol{p} - d_1\boldsymbol{p}) - \widetilde{A}(\boldsymbol{p} - d_1\boldsymbol{p})) + k_G(G(\boldsymbol{p} - d_2\boldsymbol{p}) - \widetilde{G}(\boldsymbol{p} - d_2\boldsymbol{p}))}_{\text{quantization error}}\right)^2\right] \tag{6}$$

one can remark that the first term corresponds to the motion estimation error, while the second term represents the quantization error of the reference frames. At this step, we introduce M_{d_A,d_G} which is the variance of the motion estimation error (with the non-quantized frames). This error depends on d_A and d_G, which are the distances from the current frame B to the two frames used to generate the SI. We also denote by D_A and D_G the two quantization distortions of the reference frames which are supposed, without loss of generality, to be independent. Then, the previous relations yield:

$$\begin{aligned}\sigma_{e_B}^2 = {} & \underbrace{\mathrm{E}\left[(B(\boldsymbol{p}) - k_A A(\boldsymbol{p} - d_1\boldsymbol{p}) - k_G G(\boldsymbol{p} - d_2\boldsymbol{p}))^2\right]}_{M_{d_A,d_G}} + k_A^2 \\ & \times \underbrace{\mathrm{E}\left[(A(\boldsymbol{p} - d_1\boldsymbol{p}) - \widetilde{A}(\boldsymbol{p} - d_1\boldsymbol{p}))^2\right]}_{D_A} + k_G^2 \\ & \times \underbrace{\mathrm{E}\left[(G(\boldsymbol{p} - d_2\boldsymbol{p}) - \widetilde{G}(\boldsymbol{p} - d_2\boldsymbol{p}))^2\right]}_{D_G}\end{aligned}$$

$$\sigma_{e_B}^2 = M_{d_A,d_G} + k_A^2 D_A + k_G^2 D_G \tag{7}$$

We notice that the estimation error $\sigma_{e_B}^2$ is divided into two independent terms: M_{d_A,d_G}, the motion estimation error computed with perfect reference frames and $k_A^2 D_A + k_G^2 D_G$, the quantization distortions of the reference frames. Based on this equation, it is easy to make a recursive analysis of the estimation error. A classical result of information theory gives the general expression of the rate-distortion function of a frame X, under the assumption of high bitrate [22,23]: $D_X = \alpha_X \sigma_X^2 2^{-2R_X}$, where R_X is the allocated rate in bits per pixel, σ_X^2 the spatial variance of the frame X and α_X is a constant depending on the source distribution. Then, if a frame F_0^0 is estimated from two reference frames F_0^1 and F_1^1 (respectively, at a distance of d_0^1 and d_1^1), we can write:

$$D_{F_0^0} = \alpha_{F_0^0}\sigma_{e_{F_0^0}}^2 2^{-2R_{F_0^0}} = \alpha_{F_0^0}\left(M_{d_0^1,d_1^1} + k_{F_0^0}^2 D_{F_0^1} + k_{F_1^1}^2 D_{F_1^1}\right)2^{-2R_{F_0^0}} \tag{8}$$

If we assume that, without loss of generality, F_0^1 was itself previously estimated from the two frames F_0^2 and F_1^2 and that the corresponding distances are d_0^2 and d_1^2, we can write:

$$D_{F_0^0} = \alpha_{F_0^0}\left(M_{d_0^1,d_1^1} + k_{F_0^1}^2\alpha_{F_0^1}\left(M_{d_0^2,d_1^2} + k_{F_0^2}^2 D_{F_0^2} + k_{F_1^2}^2 D_{F_1^2}\right)2^{-2R_{F_0^1}} + k_{F_1^1}^2 D_{F_1^1}\right)2^{-2R_{F_0^0}} \tag{9}$$

More generally, if we assume that the frame F_i^j is estimated from the two reference frames $F_{i(j+1)}^{j+1}$ and $F_{i(j+1)+1}^{j+1}$, we have to replace $D_{F_i^j}$ at iteration j, $i \in \{0,\ldots,2^j - 1\}$ by

$$D_{F_i^j} = \alpha_{F_i^j}\left(M_{d_{i(j+1)}^{j+1},d_{i(j+1)+1}^{j+1}} + k_{F_{i(j+1)}^{j+1}}^2 D_{F_{i(j+1)}^{j+1}} + k_{F_{i(j+1)+1}^{j+1}}^2 D_{F_{i(j+1)+1}^{j+1}}\right)2^{-2R_{F_i^j}} \tag{10}$$

This recursive analysis is done by repeating this process for all frames, until the last frames are KFs, not estimated from other frames. In that case, we have the corresponding distortion $D_{KF} = \alpha\sigma_{KF}^2 2^{-2R_{KF}}$, where R_{KF} is the allocated rate in bits per pixel, σ_{KF}^2 the spatial variance of the original key frame and α is a constant depending on the source distribution. We can finally have an expression of the rate-distortion function where all the parameters (estimation error: $M_{x,y}$ and quantization constant: α) can be experimentally estimated.

5. Proposed solutions for multi-view DVC

After drawing the different problems involved in multi-view distributed video coding, we propose some new solutions for two of them based on the theoretical study of Section 4. In Section 5.1, long term estimations for SI are proposed, while in Section 5.2 we introduce a new symmetric scheme and use the theoretical analysis proposed in Sec 4 to study the SI estimation efficiency for different situations.

5.1. Long-term estimation for side information

In the SI generation step, many estimations may be computed. For us, the first question was the real contribution of having all these estimations to build a single better SI. In [13,15,20], some authors have already shown that when an ideal fusion of temporal and inter-view estimations is done, the performance is about 1 dB better than when the side information is only build with a temporal estimation. However, the existing work by now does not consider more than two estimations. We investigate here the contribution of exploiting further dependencies present in other estimations. At this point, we do not take into account the real fusion issue, and only make an ideal fusion. Thus, the error matrices are not estimated but simply computed using the original frame. Of course, this cannot be done in a real implementation and that does not solve the problem of fusion. But this approach will provide not just an upper bound of the performances one can obtain with these schemes but a more meaningful indication of the contribution one can expect from the long-term estimations and answer many questions: Is it useful to search better new estimations? Is it useful to search fusion methods able to create a unique side information from many estimations?

We adopted the existing symmetric scheme, Sym2, and computed at the decoder the temporal, the inter-view and what we call the "diagonal" estimations. The diagonal estimations are the two represented in Fig. 8. Actually, as presented in Fig. 8, four diagonal estimations can be generated, but we only use the ones corresponding to a longer distance along the temporal axis. This is because we use a temporal estimator also for disparity estimation, not really efficient for the other diagonal estimations. In Section 6, thanks to our approach, we present the advantages of exploiting these long-term dependencies. Note that these diagonal estimations introduce some additional complexity at the decoder, related to the motion estimations between KFs. This is however much less complex (almost 10 times faster) than the existing channel decoding operations. In the following, we will study the Sym2 scheme, in which three methods are compared. The first, called Sym2 T in the sequel, only uses the temporal estimation. The second, called Sym2 T+V, is the ideal fusion between the temporal and the inter-view estimation. The last, called Sym2 T+V+D, is the ideal fusion between the Sym2 T+V and the two diagonal estimations introduced above.

5.2. New symmetric scheme

5.2.1. Symmetric 1/4 scheme (Sym4)

Based on the analysis of the dependency between the number of estimations and the quality of the side information, we propose a new symmetric scheme. Our first goal is to preserve the symmetric nature of the schemes. We notice that in the mono-view distributed video coding the length of the GOP can be more than 2. Why not also decrease the number of KF in the multi-view distributed case? This is why we propose a scheme called *symmetric* 1/4 (Sym4) in Fig. 9. This scheme, if its performances prove to be acceptable, has the advantage of being even less complex at the encoder, and this is one of the main goals of distributed coding. However, the decoder complexity is increased, since the number of WZFs which need to be channel decoded has grown.

We did not consider a scheme similar to the one used for hierarchical B frames (in multi-view source coding [21]), with I frames obtained only by a dyadic subsampling of the video sequence, since we wanted to fully exploit the correlations in both temporal and view directions for each WZF. Indeed, in the JMVM approach, the first motion/disparity compensated interpolations are done in a single direction (temporal or view).

5.2.2. Decoding strategy

With this new symmetric scheme, many ways of decoding are conceivable. In this section we propose a theoretical study, in order to choose the one having the best rate-distortion (RD) performances. Based on the recursive rate-distortion analysis introduced in Section 4, we will first study the mono-dimensional case, and

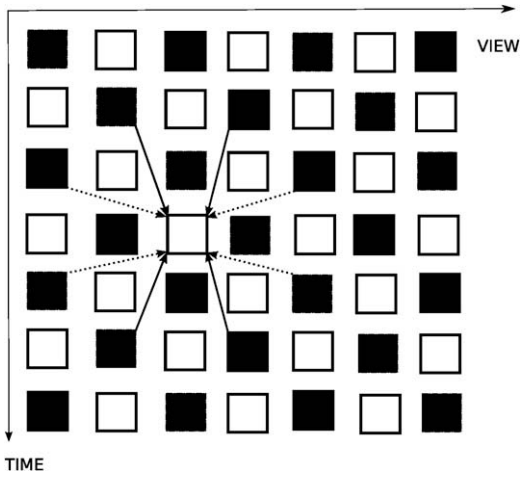

Fig. 8. Diagonal estimations used (full line) and potentially useful but not used (dashed line) for the generation of SI. KF are in black, WZF in white.

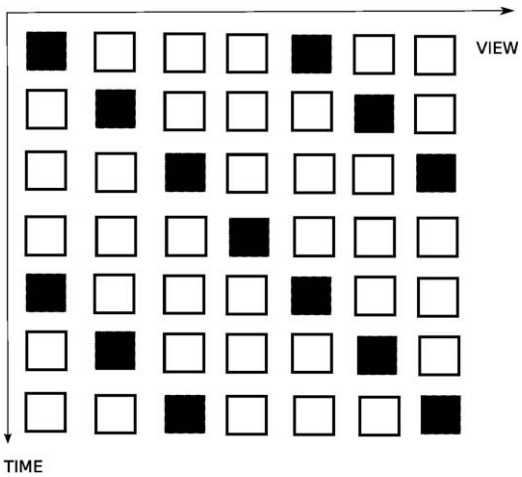

Fig. 9. Symmetric 1/4 scheme (Sym4). KF are in black, WZF in white.

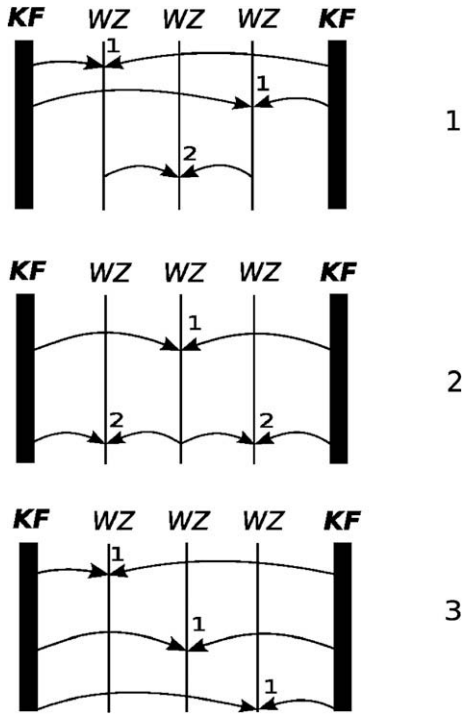

Fig. 10. Three decoding strategies for Sym4. The numbers indicate the temporal order of estimating the SI for the different WZFs.

then we will extend the found conclusions for multi-dimensional (temporal and view) conditions.

In one dimension, corresponding to the view or time axis in the Sym4 scheme, three decoding strategies may be envisaged, as illustrated in Fig. 10. In the first strategy, the two WZFs closest to the KFs are first decoded and thanks to them, the SI of the middle WZF is then interpolated. In the second strategy, very similar in spirit with the "hierarchical B frames" [25], the middle WZF is first decoded and then it is used to generate the SI necessary for decoding the two other WZFs. In the third strategy, all the WZFs are simultaneously decoded, thanks to the SI generated from the two KFs.

In order to choose the best decoding strategy, let us study the theoretical dependencies between frames in the three situations. Based on the RD model introduced in Section 4, and with the notations in Fig. 10, let us calculate the rate-distortion function for each of the three strategies, and compare them. We call the middle WZF, WZ_m, and the two others are called lateral frames, WZ_l. We do not make the difference between the two WZ_l, because the three decoding strategies give an identical role to both lateral WZFs. Denoting by D_l and D_m (resp. by R_l and R_m) the variances of the estimation errors (resp. the rates) of the frames WZ_l and WZ_m, let us calculate the total distortion: $D = 2D_l + D_m$. We denote by D_K the distortion of a KF.

Strategy 1: Following the temporal order for estimating the SI illustrated in Fig. 10, we can first write the distortion of the lateral frames generated by two KFs at a distance of 1 and 3. With the notation of Section 4, the coefficient k_A and k_G are then $\frac{3}{4}$ and $\frac{1}{4}$. Eq. 10 leads to:

$$D_l = \alpha\sigma_l^2 2^{-2R_l} = \alpha\left(M_{1,3} + \left(\frac{3}{4}\right)^2 D_K + \left(\frac{1}{4}\right)^2 D_K\right)2^{-2R_l} = \alpha\sigma_l^2 2^{-2R_l}$$
$$= \alpha\left(M_{1,3} + \frac{5}{8}D_K\right)2^{-2R_l}$$

Following the same methods, the distortion of the middle frame, after reconstructing the lateral WZFs, is:

$$D_m = \alpha\sigma_m^2 2^{-2R_m} = \alpha\left(M_{1,1} + \left(\frac{1}{2}\right)^2 D_l + \left(\frac{1}{2}\right)^2 D_l\right)2^{-2R_m}$$
$$= \alpha\left(M_{1,1} + \frac{1}{2}D_l\right)2^{-2R_m}$$
$$= \alpha M_{1,1}2^{-2R_m} + \alpha^2\frac{1}{2}\left(M_{1,3} + \frac{5}{8}D_K\right)2^{-2(R_m+R_l)}$$

Strategy 2: Again according to the temporal estimation order in Fig. 10, the distortion of the middle frame is:

$$D_m = \alpha\sigma_m^2 2^{-2R_m} = \alpha\left(M_{2,2} + \frac{1}{2}D_K\right)2^{-2R_m}$$

Then the distortion of the lateral frames reads:

$$D_l = \alpha\sigma_l^2 2^{-2R_l} = \alpha\left(M_{1,1} + \frac{1}{4}D_m + \frac{1}{4}D_K\right)2^{-2R_l}$$
$$= \alpha(M_{1,1} + \frac{1}{4}D_K)2^{-2R_l} + \alpha^2\frac{1}{4}\left(M_{2,2} + \frac{1}{2}D_K\right)2^{-2(R_m+R_l)}$$

Strategy 3: We start by estimating the distortion of the middle frame:

$$D_m = \alpha\sigma_m^2 2^{-2R_m} = \alpha\left(M_{2,2} + \frac{1}{2}D_K\right)2^{-2R_m}$$

Then, the distortion of the lateral frames is:

$$D_l = \alpha\sigma_l^2 2^{-2R_l} = \alpha\left(M_{1,3} + \frac{5}{4}D_K\right)2^{-2R_l}$$

Then, it is possible to compute the total distortion of the WZFs for each strategy:

$$D_1 = D_m + 2D_l \tag{11}$$

In order to plot these three rate-distortion functions, we have to estimate the quantities: σ_K^2, $M_{1,3}$, $M_{1,1}$ and $M_{2,2}$ using the maximum number of frames in each direction in order to have the best estimation of these coefficients (100 frames of the first camera for temporal coefficients and four times 8 frames at the same temporal instant for the view coefficients). Fig. 11 presents these coefficients estimated on two multi-view test sequences, in the time direction and in the view direction. Three remarks can be made at this point: first, as expected, the motion/disparity pre-

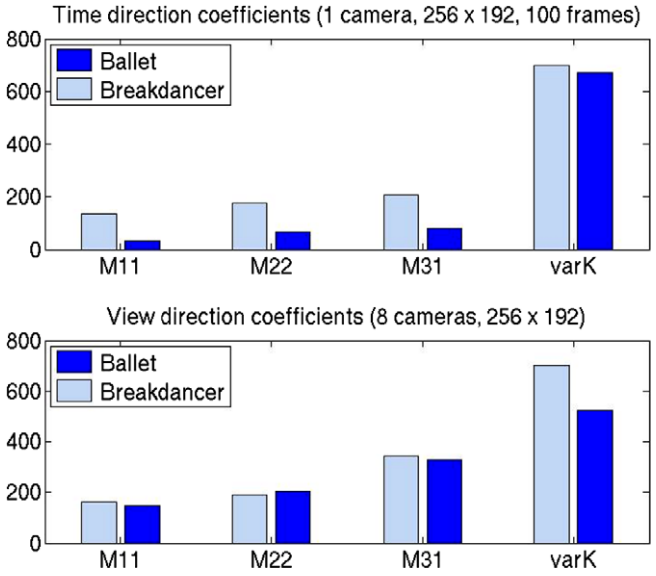

Fig. 11. Values of the different dependency coefficients for "Ballet" and "Breakdancer" sequences.

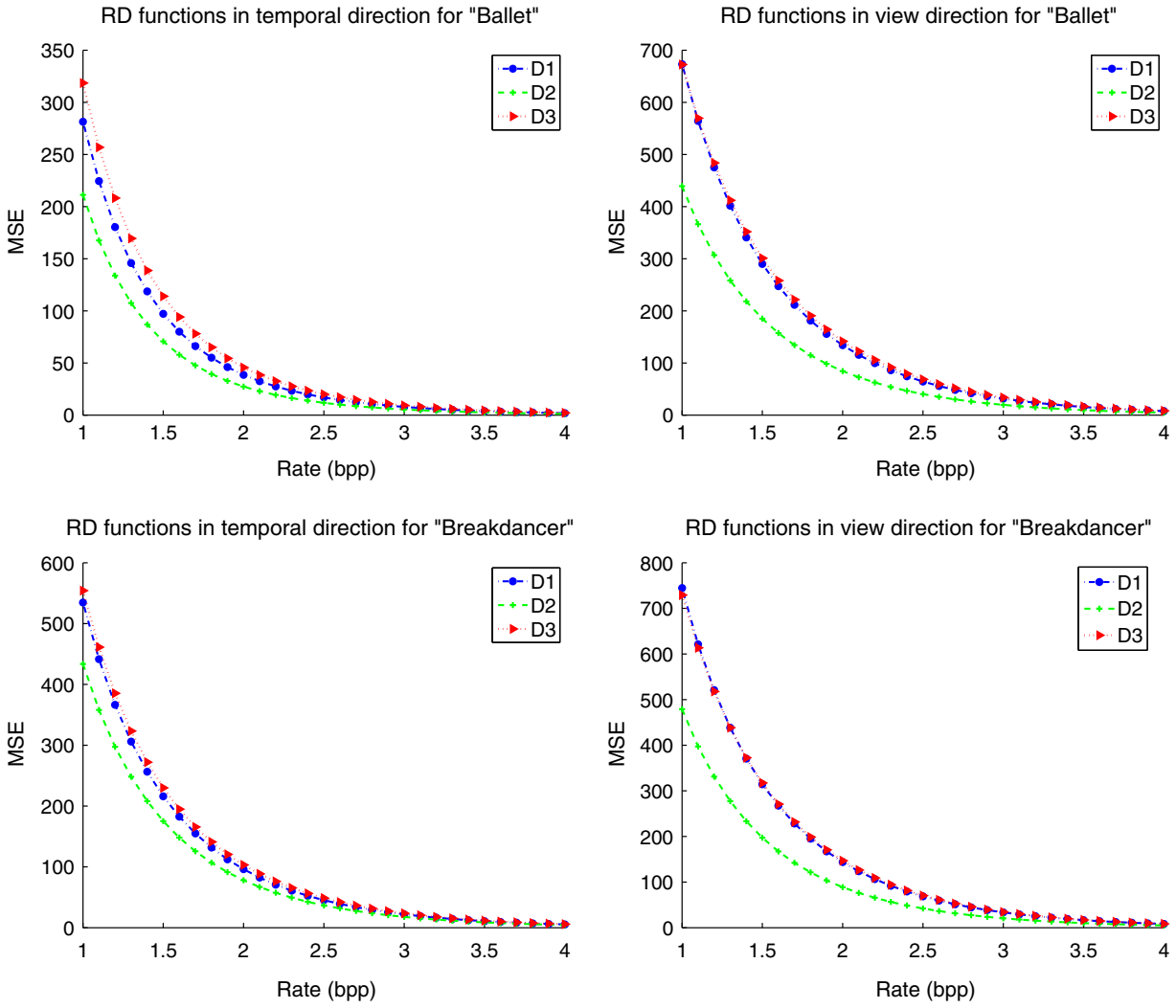

Fig. 12. Rate-distortion functions for the test sequences "Ballet" and "Breakdancer" (8 cameras, 256 × 192, 15 fps per view). D_1, D_2 and D_3 are the distortions corresponding to the three estimation strategies.

diction errors, as well as the quantization errors, are much lower than the variance of the KFs. Secondly, we notice that the estimation error is lower when the maximum distance (i.e., the distance to the furthest frame) is small. Indeed, $M_{1,1} < M_{2,2} < M_{3,1}$. Finally, the estimation errors are more important for "Breakdancer" sequence than for "Ballet" sequence. We can thus expect worse results for this sequence and in general, estimating these prediction errors gives a good idea about the coding performances that may be expected for a given sequence. The estimation of α_i coefficients is based on a detailed rate-distortion analysis presented in [26]. For the KFs, the hypotheses are: Gaussian distribution, high bitrate, while for WZFs we considered a Laplacian distribution. Note also that, in this reference, we deduce rate-distortion models for theoretical sources and low bitrates. However, these are less practical to exploit, so here we keep with the classical high bitrate rate-distortion model. The α_i coefficients can also be estimated from the real RD functions of the KFs or WZFs, not necessarily be computed based on a theoretical model of the source.

Using these estimated values, we plot the different rate-distortion functions for the two test sequences, "Ballet" and "Breakdancer" in temporal and view directions. Fig. 12 shows the experimental results and one can see that the best strategy is the second one.

We have thus the best solution for the one-dimensional problem. Fig. 13 shows the proposed two-dimensional solution corresponding to the previous analysis. Indeed, separately in the view direction and in the temporal direction, the best decoding strategy is the second one. We have kept in mind the ideal fusion approach in order to compare this scheme with the Sym2 T+V+D. Fig. 13 presents the decoding strategy, and the different estimations made for each WZF. For the first WZF to decode, we make the fusion between three estimations (temporal, inter-view and diagonal). For the second, we compute the fusion of temporal and view estimations.

6. Experimental results

In this section we test the proposed approaches. We use again the two multi-view test sequences: "Breakdancer" and "Ballet". For reasons of computation complexity, we reduce the spatial resolution to 256 × 192 after a low-pass filtering as it is done in [20]. For both, the time resolution is 15 fps and we used the 8 cameras with the first 20 frames per view.[1] The results are

[1] The number of frames used for experiments is different in Section 5.2.2 and here. This is not a problem because in Section 5.2.2 the experiments were done in order to confirm the RD model for the whole video sequence.

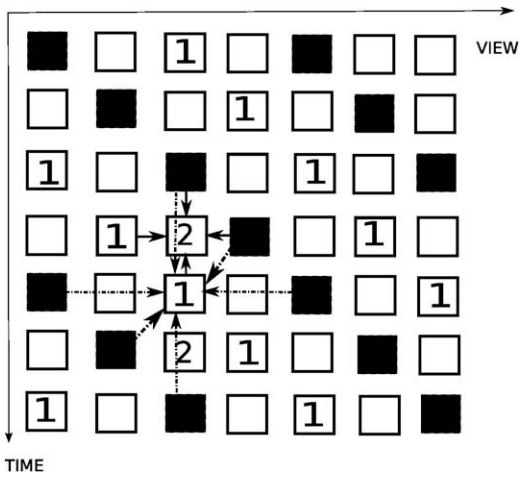

Fig. 13. Decoding strategy for Sym4.

presented through rate-distortion performance. The rates presented are the total rates (WZF + KF) per camera (because the schemes used are symmetric) for the luminance component (as usual for WZ coding).

6.1. Influence of the diagonal estimations on the side information quality

As presented below, the first experiments are made in order to show the contributions of further KF. In Fig. 14 is presented the quality of the SI generation when using original KFs for PSNR estimation and the three estimation strategies. We can clearly notice that diagonal estimations increase the performances by more than 0.5 dB. This is also visible in Fig. 15. One can see in the first three images the difference of the SI generated with Sym2 T, Sym2 T+V, Sym2 T+V+D. In the remaining of the experiments, the KFs are lossy encoded and the quantization parameters are adjusted in order to obtain a similar visual quality along the sequence.

It is well-known that the quality of the side information has a direct impact on the final rate-distortion performance. In order to verify this affirmation we have made the experiments to assess the relation between the PSNR of the SI using different types of estimation and the global quality of the decoded sequence. We compare these three ways of coding with the Intra coding and the Hyb2 scheme described in Fig. 4(c). The Intra codec is H264 intra. For the Hyb2, columns of Key cameras and Hybrid cameras alternate in the two-dimensional time-view space. In each of the Hybrid cameras, the WZFs are estimated using the ideal fusion of temporal

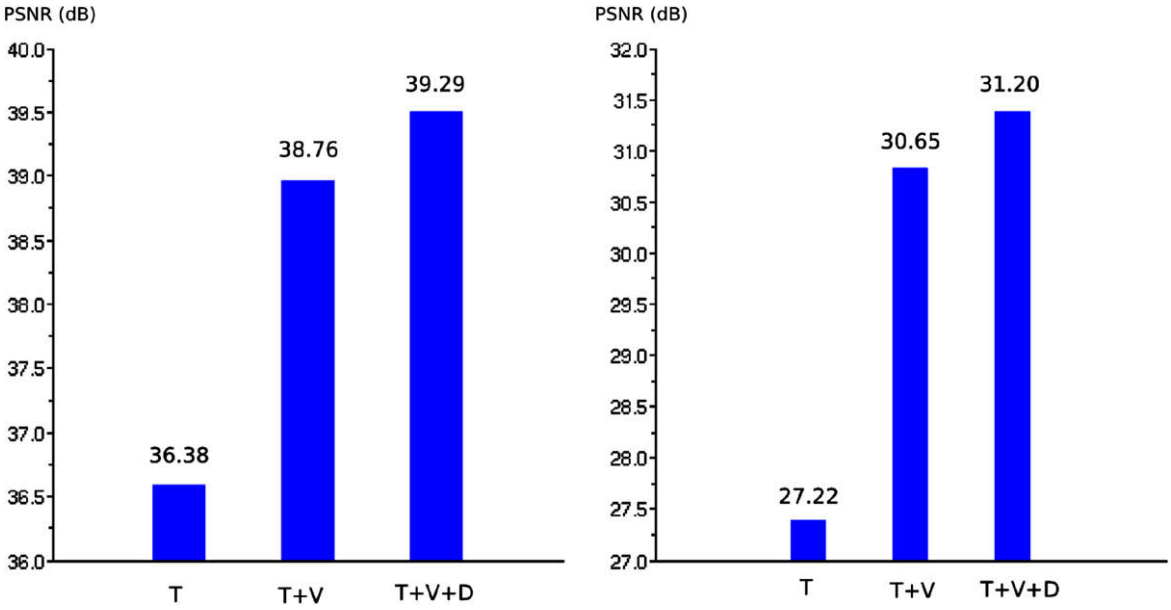

Fig. 14. PSNR of SI for " Ballet" (left) and for " Breakdancer" (Right) with Sym2 and different strategies for SI generation (T, temporal; T+V, ideal temporal and view fusion; T+V+D, temporal, view and diagonal estimations). All the estimations are generated with lossless KFs.

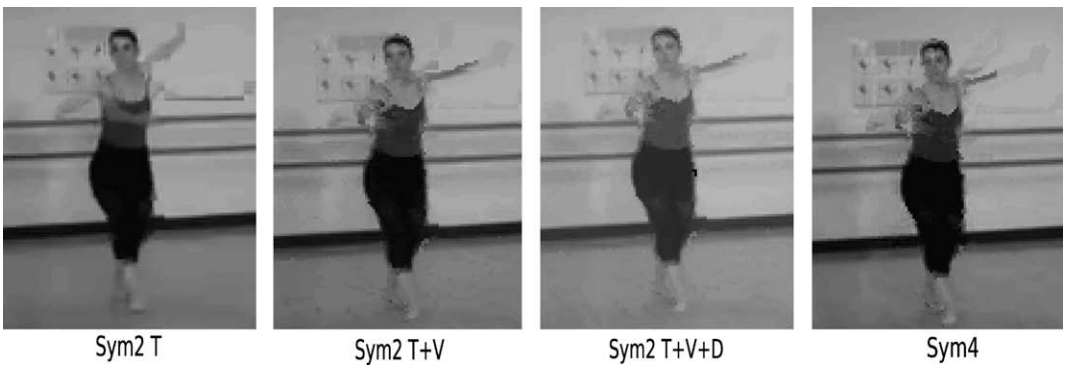

Fig. 15. Visual comparison of the generated SI, depending on schemes, for the " Ballet" sequence, 256 × 192, view 2, temporal instant 3.

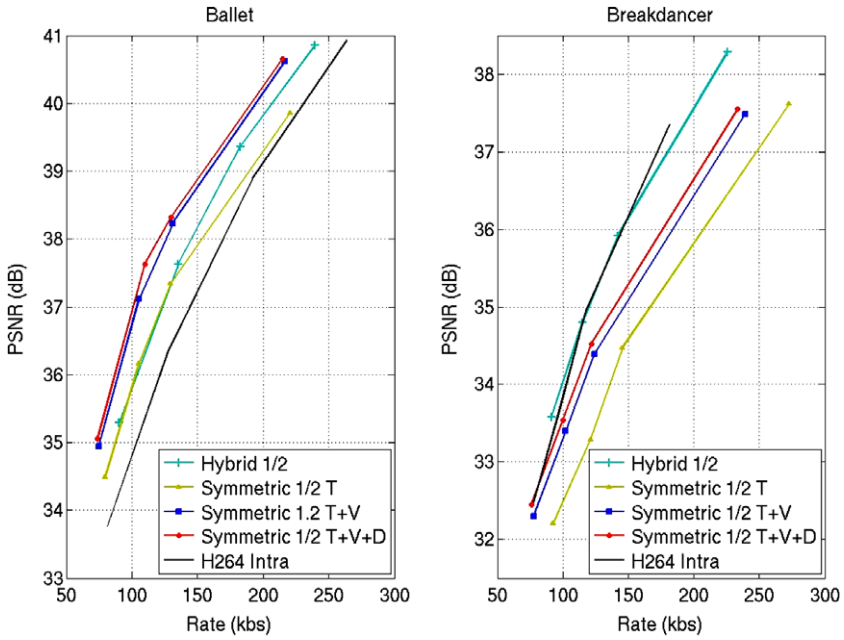

Fig. 16. Comparison of the RD performance for "Ballet" and "Breakdancer" (8 cameras, 256 × 192, 15 fps per view) for different SI estimation strategies.

Table 1
Complexity comparison (computation time in seconds per frame) at the encoder and at the decoder, for different schemes

Scheme	Encoder	Decoder
H.264 Intra	0.25	0.03
Hyb2 T+V	0.21	5.11
Sym2 T	0.13	13.39
Sym2 T+V	0.13	14.45
Sym2 T+V+D	0.13	15.08
Sym4	0.07	17.46

and view predictions. The rate presented is, as in the other schemes, the average of the different rates per camera. Fig. 16 shows the results and, as expected, the diagonal estimations improve the performances. The improvement is not of 0.5 dB anymore as in Fig. 14, because the estimations are now generated with lossy KFs, but the contribution of diagonal estimations is still clearly visible. It is therefore interesting to further investigate the contribution of diagonal estimations, using a real fusion strategy (as in [20]).

Indeed, we present in Table 1, the computational complexity (time in seconds per frame), at the encoder and at the decoder for different schemes. This was measured on a "Intel Core 2 Duo" machine, 2.66 GHz, under linux, for "Breakdancer" sequence, on 5 views and 5 frames per view. The reported results are average computation times per frame. The experimental results confirm that computing more estimations (at the decoder) does not affect the encoding complexity. At the decoder, while the motion estimation complexity is negligible compared to the turbodecoding com-

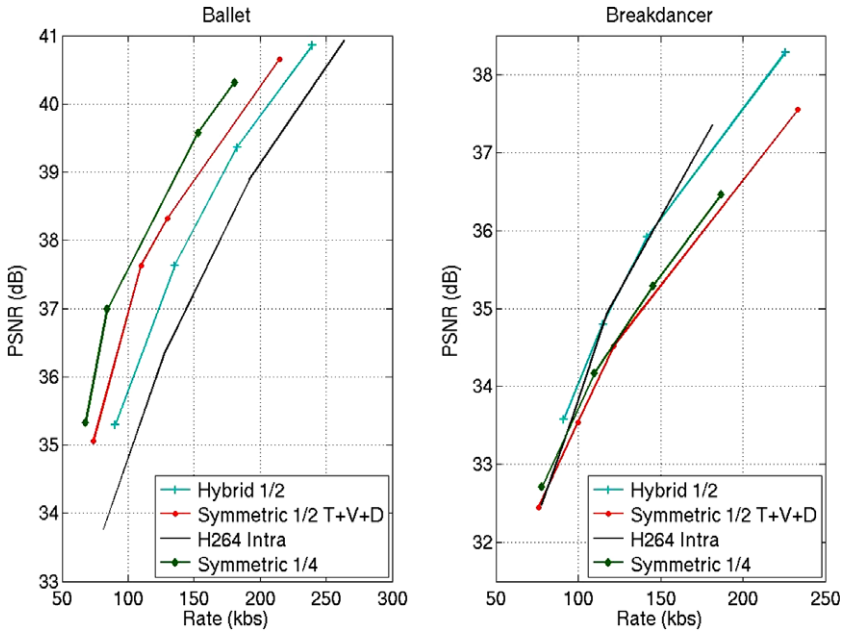

Fig. 17. Comparison of the RD performance for "Ballet" and "Breakdancer" (8 cameras, 256 × 192, 15 fps per view) for different coding schemes.

plexity, the computation of diagonal estimations does not increase sensibly the decoding time. Note that the number of diagonal estimations is smaller than that of the temporal and view estimations (for example, for 5 views and 5 frames we have 4 diagonal estimations and 16 time or view estimations).

6.2. The symmetric 1/4 scheme (Sym4)

In this section, the test results concerning the new Sym4 scheme are shown. First, one can see the visual difference between Sym2 and Sym4, for the SI quality in Fig. 15. The SI quality is relatively equivalent between Sym2 T+V+D and Sym4. In experiments shown in Fig. 17, we compare the Sym4 with the Sym2 T+V+D of the previous experiments and with the Hyb2 (as presented in Section 6.1). We notice that, when the performance of Sym2 is better than the Intra coding, the Sym4 is better than both Sym2 and Hyb2. This is normal, since the Intra frames are replaced by WZFs, using lower bit rates. However, for "Breakdancer" sequence, the coding efficiency is lower for the WZFs than for the Intra frames, and thus replacing KFs by WZFs degrades the performances. This explains why for this sequence Sym4 has lower performance than Hyb2, but we notice that Sym4 is better than Sym2 T+V+D. The results are interesting because they show the potential of Sym4. Moreover, as presented in Table 1, the encoding complexity of Sym4 represents only 50% of the Sym2 complexity and only 30% of the Intra configuration complexity.

7. Conclusion and future work

In this paper we first proposed several symmetric schemes for multi-view DVC. In this framework, we investigated the potential advantages of building several estimations for the side information (SI). A new way of long-term prediction for the SI, called "diagonal", was also proposed. For one of the proposed schemes, having the distance between KFs of 4, we have provided a theoretical analysis of several decoding strategies, based on a rate-distortion model involving a first term only depending on the motion/disparity fields in the sequence, and a second term only depending on the quantization. We have confirmed it by experimental results, both for the temporal and inter-view prediction. This scheme allows to increase the performances of the distributed coding for high-correlated sequences, while further reducing the complexity at the encoder. However, the decoding complexity is slightly increased, so in practice, a trade-off between performances and complexity has to be considered.

Future work will focus on the study of different techniques to construct the SI given many reference frames. Moreover, the results we have shown need to be confirmed in schemes involving performant strategies for real fusion.

References

[1] D. Slepian, J.K. Wolf, Noiseless coding of correlated information sources, IEEE Trans. Inform. Theory 19 (1973) 471–480.
[2] A. Wyner, J. Ziv, The rate-distortion function for source coding with side information at the receiver, IEEE Trans. Inform. Theory 22 (1976) 1–11.
[3] B. Girod, A. Aaron, S. Rane, D. Rebollo-Monedero, Distributed video coding, Proc. IEEE 93(71) (2005) pp. 71–83.
[4] J. Ascenso, C. Brites, F. Pereira, Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding, in: EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services, Smolenice, Slovak Republic, 2005.
[5] R. Puri, K. Ramchandran, PRISM: a video coding architecture based on distributed compression principles, Tech. Rep. UCB/ERL M03/6, EECS Department, University of California, Berkeley, 2003.
[6] Z. Li, L. Liu, E.J. Delp, Rate distortion analysis of motion side estimation in Wyner–Ziv video coding, in: IEEE Transactions on Image Processing, vol. 16(1), January 2007, pp. 98–113.
[7] M. Tagliasacchi, L. Frigerio, S. Tubaro, Rate-distortion analysis of motion-compensated interpolation at the decoder in distributed video coding, in: IEEE Signal Processing Letters, vol. 14, 2007, pp. 625–628.
[8] A. Aaron, R. Zhang, B. Girod, Wyner–Ziv coding of motion video, in: Proceedings of Asilomar Conference on Signals and Systems, Pacific Grove, CA, November 2002.
[9] J. Ascenso, C. Brites, F. Pereira, Content adaptive Wyner–Ziv video coding driven by motion activity, in: IEEE ICIP, Atlanta, GA, USA, 2006, pp. 605–608.
[10] A. Aaron, S. Rane, E. Setton, B. Girod, Transform-domain Wyner–Ziv codec for video, in: SPIE Visual Communications and Image Processing Conference, vol. 5308, San Jose, CA, 2004, pp. 520–528.
[11] Q. Xu, Z. Xiong, Layered Wyner–Ziv video coding, IEEE Trans. Image Process. 15 (12) (2006) 3791–3803.
[12] M. Morbee, J. Prades-Nebot, A. Pizurica, W. Philips, Rate allocation algorithm for pixel-domain distributed video coding without feedback channel, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, vol. 1, 15–20 April, 2007, pp. I-521–I-524.
[13] M. Ouaret, F. Dufaux, T. Ebrahimi, Fusion-based multiview distributed video coding, in: ACM International Workshop on Video Surveillance and Sensor Networks, Santa Barbara, CA, USA, 2006.
[14] X. Artigas, E. Angeli, L. Torres, Side information generation for multiview distributed video coding using a fusion approach, in: 7th Nordic Signal Processing Symposium, Iceland, 2006.
[15] X. Guo, Y. Lu, F. Wu, W. Gao, S. Li, Distributed multi-view video coding, SPIE-IST Electronic Imaging, SPIE, vol. 6077, San Jose, CA, USA, 2006, pp. 15–19.
[16] C. Brites, J. Ascenso, F. Pereira, Improving transform domain Wyner–Ziv video coding performance, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, 2006.
[17] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, J. Ostermann, Distributed monoview and multiview video coding: basics, problems and recent advances, IEEE Signal Processing Magazine, Special Issue on Signal Processing for Multiterminal Communication Systems.
[18] X. Artigas, F. Tarres, L. Torres, Comparison of different side information generation methods for multiview distributed video coding, in: International Conference on Signal Processing and Multimedia Applications SIGMAP, Barcelona, Spain, July 2007.
[19] L. Zhan-Wei, A. Ping, L. Su-Xing, Z. Zhao-Yang, Arbitrary view generation based on DIBR, in: International Symposium on Intelligent Signal Processing and Communication Systems, 2007, ISPACS 2007, November 28, 2007–December 1, 2007, pp. 168–171.
[20] J. Areia, J. Ascenso, C. Brites, F. Pereira, Wyner–Ziv stereo video coding using a side information fusion approach, in: IEEE International Workshop on Multimedia Signal Processing, Chania, Greece, 2007, pp. 453–456.
[21] ISO/IEC MPEG & ITU-T VCEG, Joint multiview video model (JMVM), Marrakech, Morocco, January 13–19, 2007.
[22] T. Berger, Rate Distortion Theory: A Mathematical Basis for Data Compression, Englewood Cliffs, 1971.
[23] T.M. Cover, J.A. Thomas, Elements of Information Theory, second edition, Hardcover, July 2006.
[24] S. Hemami, Distortion analyses for temporal scalability coding techniques, in: International Conference on Image Processing, 1999, ICIP 99, vol. 3, Kobe, Japan, 1999, pp. 349–353.
[25] T. Wiegand, G. Sullivan, G. Bjntegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 560–576.
[26] A. Fraysse, B. Pesquet-Popescu, J. Pesquet, On the uniform quantization of a class of sparse source, IEEE Trans. Inform. Theory (submitted for publication).