EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

# Novel solutions for side information generation and fusion in multiview DVC

Giovanni Petrazzuoli, Marco Cagnazzo* and Béatrice Pesquet-Popescu

**Abstract**

One of the key problems in distributed video coding is the generation of side information. This task consists of producing an estimate of an image with some neighboring ones, such as those taken by the same camera at different time instants, or, in the case of multiview setups, images taken at the same time instant by different cameras. If both estimates are available, a further problem arises, which is how to merge them in order to create a single side information. This problem is very relevant since a good estimate of the unknown image will require only a few bits to be corrected. Considering a multiview distributed video-coding setup, we propose a novel technique for inter-view interpolation based on occlusion prediction, a new fusion technique from multiple estimates, and finally an adaptive validation step for switching among the three possible side information images: temporal, inter-view, and fusion. We provide a comprehensive set of experimental results, which indicate bit rate reductions of more than 9% in average; moreover, we observe much more consistent results with respect to state-of-the-art techniques.

**Keywords:** Distributed video coding; Side information; Occlusion detection; Inter-view interpolation; Fusion

## 1 Introduction

Even if the distributed source coding theory is more than 30 years old [1,2], it is only in the last 10 years that practical distributed video coding (DVC) systems have been proposed [3]. Since then, this topic has gathered much attention from the research community [4,5] because several applications could benefit from efficient DVC, as for example interactive multiview video streaming [6-8].

In a typical DVC system, the encoder is not able to jointly encode all the input images. There can be several reasons: there are multiple sources that cannot communicate, as in the classical DVC paradigm; the joint encoding process is too computational expensive, such as in the case of light-weight sensor networks; or, even if the encoder knows all the input images, it does not know which ones have been requested by the user, such as in the case of interactive multiview video streaming. In all these cases, for these unknown images, the encoder only sends some parity bits of an error-correcting code. These bits will be used by the decoder to correct a suitable estimate of the current image, which is produced using the available information such as images from the same camera at different time instants, or, for multiview systems, images from different cameras at the same instant. This estimate is called side information (SI). It is clear that the more the SI is similar to the actual image, the less bits will be requested to correct it [9]. Therefore, side information generation is a crucial step for an efficient DVC system.

In this paper, we provide a system for SI generation in multiview DVC made up of three main components: temporal interpolation, inter-view interpolation, and adaptive fusion. We implement them within a system based on the architecture proposed by Aaron et al. [3], which is also at the basis of the popular distributed coding for video services (DISCOVER) [10] and VISNET I/II [11] DVC systems. In such a scheme (shown in Figure 1), the input images are split into key frames (KFs) and Wyner-Ziv frames (WZFs).

The KFs are encoded with a still image coding technique; at the decoder, they are used to produce an estimate of the current WZF, i.e., the SI. This estimate is then corrected by requesting a suitable number of bits from a channel code. Therefore, the encoder need not to know the SI to produce these bits: the encoding of KFs and WZFs is actually distributed. The Wyner-Ziv coder usually works in the DCT domain: the WZFs are transformed

*Correspondence: cagnazzo@telecom-paristech.fr
Department of Image and Signal Processing, Institut TELECOM, TELECOM-ParisTech 46 rue Barrault, F-75634 Paris, France

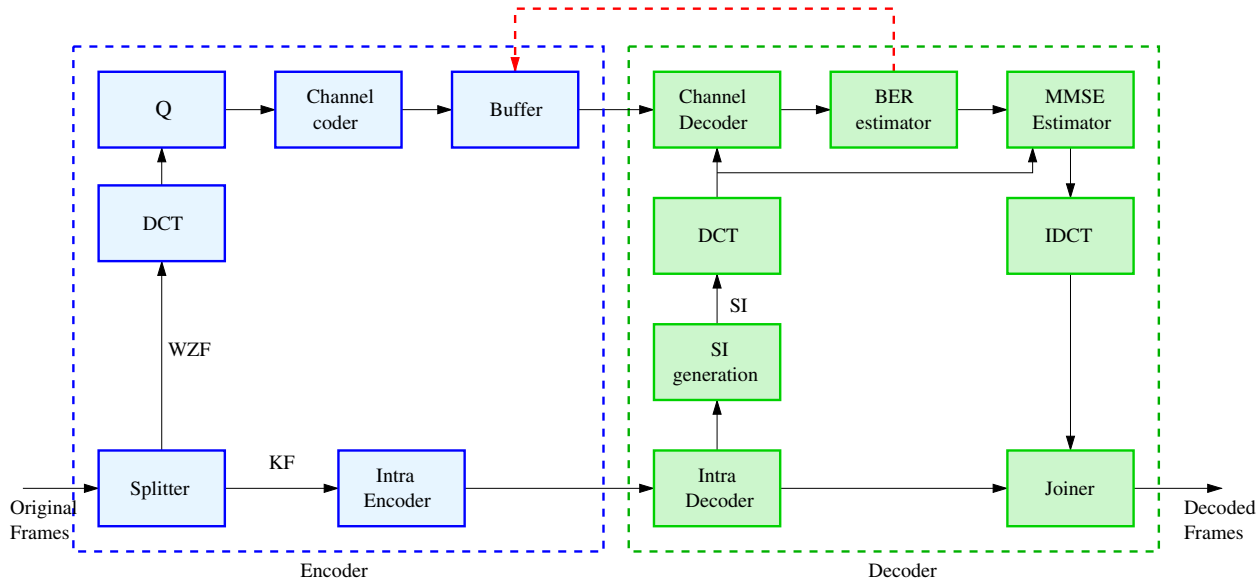


**Figure 1 The reference DVC system [3].**

and quantized, and the channel code produces the parity bits for the coefficients' bit planes. At the decoder side, the SI is in turn transformed, and the received parity bits are used to correct the DCT bands. Further bits are requested until the estimated bit error rate drops below a given threshold, then each coefficient value is reconstructed using a minimum mean square error estimator, and the IDCT returns the decoded Wyner-Ziv frame.

The technique for SI generation should take into account as much as possible the *a priori* information about the DVC system. In particular, for a multiview DVC system, different techniques exist for the case of temporal interpolation (SI generated from images of the same camera at different instants) or inter-view interpolation (SI generated from images of different cameras at the same time). When both estimates are available, a further problem arises, which is called fusion: for each pixel, should we use the temporal estimate, the inter-view estimate, or a combination of both? In this paper, we propose new solutions to these problems. The first contribution is a novel solution for the inter-view SI generation, based on occlusion avoidance. For the temporal interpolation, we consider a method we proposed for monoview video, and we adopt it in the context of multiview. We then provide a new method for SI fusion, based on two contributions: an occlusion detection method and an adaptive decision algorithm based on the rate of the channel code. The latter allows to overcome the performance inconsistency that has been often observed in previous fusion algorithms. Finally, we consider a multiview DVC codec that takes advantage of all these improvements, and we compare it with state-of-the-art systems. An extensive experimental validation fully supports the proposed techniques.

The rest of the paper is organized as follows. In Section 2, we describe the state of the art for interpolation and fusion. Then, two sections (Sections 3 and 4) are devoted to the new inter-view interpolation technique and the fusion algorithms. Experimental results (Section 5) are then followed by the conclusion and the outline of future work (Section 6).

## 2 State of the art for side information generation

Let us introduce some notations. For temporal SI generation, we consider monoview video sequences (or a single view from a multiview setup). Therefore, we only need a single (temporal) index to designate images: with $I_t$, we refer to the $t$th image of a video sequence. When we deal with multiview systems, we need a second index for the view point. We note as $I_{t,k}$ the image taken at the instant $t$ by the $k$th camera. In both cases, the corresponding SI is indicated with a hat. When the indexes are not necessary, we will simplify the notation referring to the temporal interpolation as $\widehat{I}_T$, to the inter-view interpolation as $\widehat{I}_V$ and to their fusion as $\widetilde{I}$.

In this section, we will describe the motion interpolation algorithm proposed within the DISCOVER project that will be the reference method for comparison with our proposed algorithms. Then, we will provide a state of the art about techniques improving the linear interpolation of DISCOVER both in time and in the view domain. Finally, a state of the art on fusion techniques is given.

### 2.1 Side information generation by temporal interpolation

#### *2.1.1 The DISCOVER algorithm*

The most popular method for temporal image interpolation is the one proposed within the DISCOVER project [10]. It consists of the following steps. First, the two KFs,[a] say $I_{t-1}$ and $I_{t+1}$, are low-pass filtered in order to smooth out the noise. Then a block-matching motion estimation from $I_{t+1}$ to $I_{t-1}$ is performed. The resulting motion vector field $\mathbf{v}(\cdot)$ is split in a couple of fields $\mathbf{u}(\cdot)$ and $\mathbf{w}(\cdot)$ (pointing from $t$ respectively to $t-1$ and $t+1$). For the block of the WZF centered in $\mathbf{p}$, this is obtained by looking for the trajectory closest to $\mathbf{p}$: in the hypothesis of linear motion, the block centered in $\mathbf{q}$ at time $t+1$, is centered in $\mathbf{q}+\frac{1}{2}\mathbf{v}(\mathbf{q})$ at time $t$. Then, we select the position $\mathbf{q}^*$, such that

$$\mathbf{q}^*(\mathbf{p}) = \arg\min_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \left\| \mathbf{q} + \frac{1}{2}\mathbf{v}(\mathbf{q}) - \mathbf{p} \right\|^2 \quad (1)$$

where $\mathcal{N}(\mathbf{p})$ is the set of the block center positions near $\mathbf{p}$. The vectors $\mathbf{v}$ and $\mathbf{w}$ are defined as $\mathbf{u}(\mathbf{p}) = \frac{1}{2}\mathbf{v}(\mathbf{q}^*(\mathbf{p}))$ and $\mathbf{w}(\mathbf{p}) = -\frac{1}{2}\mathbf{v}(\mathbf{q}^*(\mathbf{p}))$. Next, a bidirectional motion refinement is performed. Finally, a weighted median filter is applied on the resulting motion vector fields in order to regularize them. After this process, one ends up with two motion vector fields that are used to compensate the previous and the next key frame. The average of the resulting images is the side information $\hat{I}_t$. We observe that, even if DISCOVER uses a simple motion model (linear trajectories), it gives very good results in some cases and is so popular that it will be used as reference to validate the proposed SI generation technique.

#### *2.1.2 Other techniques*

Several algorithms have been proposed in order to improve the DISCOVER performance. Huang et al. [12] have proposed to improve the forward motion estimation using also the chroma components of the decoded KFs, as already proposed for wavelet-based video coding in [13]. They further proposed an adaptive weighted overlapped block motion compensation: each compensated block is weighted by the inverse of MSE between the forward and backward compensated blocks. In the work of Macchiavello et al. [14], the authors have slightly modified the DISCOVER algorithm: without shifting vectors, the splitting is performed. The blank areas are filled with a block matching between these partially blank blocks and the adjacent KFs.

Instead, Ascenso et al. [15] have proposed to add a constraint during the motion estimation from $t+1$ to $t-1$: all the motion vectors have to cross the center of a block in the WZFs. Kubasov et al. [16] have proposed to replace block matching technique for motion estimation by mesh-based motion-compensated interpolation, in order to take into account more complex motion than simple translation. Mys et al. [17] have proposed to use the SKIP mode (like the one used in H.264/AVC) in DVC. Since there are several blocks that do not move between $I_{t-1}$ and $I_{t+1}$, their SI does not need to be corrected, similar to what is done in the PRISM codec [5,18]. In [19,20], it is proposed to use dense vector field techniques such as the Cafforio-Rocca algorithm and a total variation-based algorithm, for motion estimation and interpolation. The results show an average rate reduction up to 5.9%.

In the two works of Ye et al. [21,22], a partially decoded WZF (using the parity bits of the DC band) allows the detection of suspicious vectors and to refine the motion vector fields. Finally, an optimal motion compensation mode selection is proposed: the previous, the next frame, the bidirectional motion-compensated average of the previous and the next frame are used for the SI by evaluating which of them gives the smallest matching error. Abou-El-Ailah et al. [23,24] have proposed to fuse global and local motion estimation for constructing the SI: for global motion estimation, the parameters that model the global motion (translational, affine, or perspective) are estimated by SIFT features and sent to the decoder. Local motion estimation is obtained by DISCOVER algorithm. These two SI are fused during the decoding process using also the partially decoded WZF. Martins et al. [25] have proposed an iterative refinement as in [21,26] for each band but only for some selected blocks. Let $Y$ be the initial SI and let $R^b$ be the partially decoded WZF up to band $b$. The blocks for which the MAD between $Y$ and $R^b$ is larger than a given threshold are refined by searching among the neighboring blocks the one that minimizes the MAD. This will be used as new SI from the correction of the next band. This refinement is performed for each band.

Hash-controlled motion estimation techniques have also been proposed: some additional information, called hash signature, is sent to the decoder [27,28], for example, some blocks of the original WZF. In this context, Verbist et al. [29] have proposed a probabilistic model for SI construction. An overlapped block motion estimation is performed from $I_{t-1}$ to $I_{t+1}$, and a collection of candidate predictor values is available for each pixel. Then, by supposing that the noise between the side information and the real WZF can be modeled as Laplacian, the best candidate can be chosen by maximum likelihood estimation.

Recently, we proposed a temporal interpolation technique based on high-order motion interpolation (HOMI) [30-32] that outperforms the DISCOVER RD performance. In this algorithm, the basic idea is to use four frames, $I_{t-3}$, $I_{t-1}$, $I_{t+1}$, and $I_{t+3}$, to estimate the current WZF $I_t$. At first, the DISCOVER motion interpolation algorithm is performed on the frames $I_{t-1}$ and $I_{t+1}$. Then, the obtained vectors are lengthened to the frames $I_{t-3}$ and

$I_{t+3}$ through a motion estimation criterion in order to estimate the trajectory of each block among the frames. The obtained four positions are interpolated in order to estimate the position of that block in the frame $I_t$. Finally, the motion vectors for that block are obtained as the difference between the position in $t$ and the positions in the next and the previous image. This algorithm will be used in the following temporal interpolation in our proposed algorithms.

Another framework for DVC developed in recent years is VISNET II codec [11]: the motion interpolation step for SI generation is the same as of DISCOVER. The main modifications at the decoder side are the iterative reconstruction of the WZF for each DCT band and an adaptive deblocking filter applied to the decoded WZFs. At the encoder, a CRC is added in order to improve robustness. A comprehensive review and classification of techniques proposed for temporal SI generation can be found in the work by Brites et al. [33].

### 2.2 Side information generation by inter-view interpolation

In multiview DVC systems, three types of camera are commonly considered: pure key cameras for which all the frames are KFs, pure Wyner-Ziv cameras where all the frames are WZFs, and hybrid cameras for which each second frame is a WZF and the others are KFs. These cameras can be arranged in a variety of configurations, but mainly three schemes have been considered in the literature [34]:

- *Asymmetric scheme* (see Figure 2a). Pure key cameras (black) and pure WZ cameras (white) are alternated.
- *Hybrid 1/2 scheme* (see Figure 2b). Every second camera is a pure key one, and the other (gray) is hybrid.
- *Symmetric 1/2 scheme* (see Figure 2c). All the cameras are hybrid, and the KFs and the WZFs are placed on a quincunx grid in the time-view axes.

In all these schemes, for a generic WZF $I_{t,n}$, at least two KFs are available, $I_{t,n-1}$ and $I_{t,n+1}$, i.e., two images taken at the same temporal instant $k$ by two neighboring cameras. Moreover, except for the first scheme, two other images are available, namely $I_{t-1,n}$ and $I_{t+1,n}$, the previous and next frames from the same camera. All these images should be used in order to generate a side information as reliable as possible. This is commonly achieved by three steps: a temporal estimation, an inter-view estimation, and finally a fusion.

In this section, inter-view estimation is considered; it can be seen as a form of image-based rendering (IBR) process. Under this perspective, this problem has been long studied [35], long before the rise of DVC systems. There is a simple approach in using the motion interpolation algorithms among neighboring views: the concept of motion is replaced by that of disparity. The results are acceptable as far as the views are rectified with a viewing axis perpendicular to the baseline [36]. This approach is simple but has the drawback of not taking into account the specificity of inter-view estimation and cannot assure the best performance. Therefore, in the literature, there are several approaches based on increasingly complex models of the multicamera setup. For example, Guo et al. [37] propose a global affine motion model with six parameters for inter-view interpolation. However, global motion models cannot accurately describe the sudden variation in the disparity field associated to object borders. Occlusions are neither taken into account.

Artigas et al. [38] have proposed a method based on depth maps. Indeed, depth maps along with camera parameters allow to create a virtual view point, namely that of the WZ camera. However, in our work, we consider

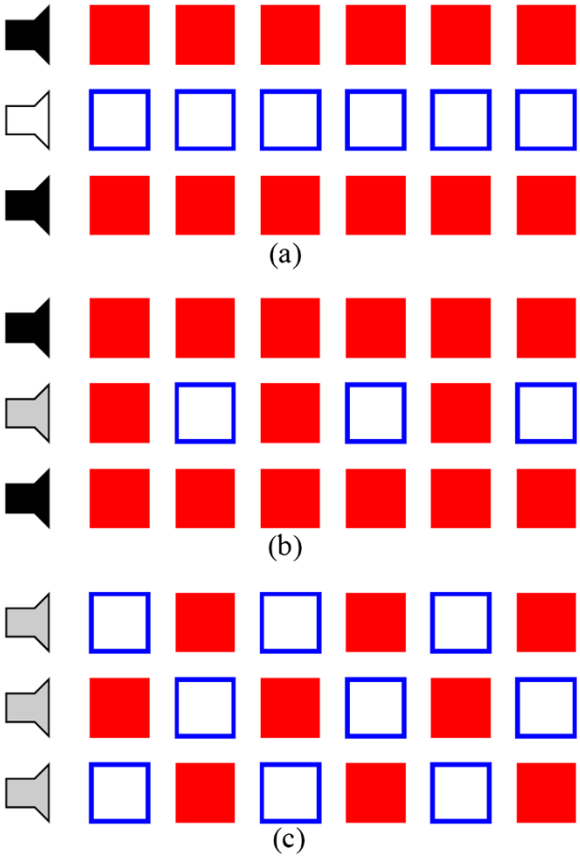


**Figure 2 Configurations for multiview DVC. (a)** Asymmetric scheme, **(b)** hybrid 1/2 scheme, and **(c)** symmetric 1/2 scheme. The filled frames are KFs, and the blank ones are WZFs. For simplicity, only three cameras per scheme are shown.

a different problem where depth information is not directly available.

A similar approach is used by Ouaret et al. [39] for a hybrid 1/2 scheme with three cameras. In this case, the homography matrices that needed to map the KF into the WZ camera are estimated using a MMSE criterion. However, local object motion can create outliers, and this invalidates the estimation of the homography matrices. Moreover, this technique can suffer from distortion caused by camera lens.

Areia et al. [40] have proposed a very simple method for the hybrid 1/2 scheme with two cameras (stereoscopic video): the disparity field is computed over the last two decoded frames, and then it is directly applied to the current KF to generate the side information. The authors have recognized that this approach may not achieve the best performance, but it has the advantage of being very simple and of reusing the same algorithms as that of the temporal interpolation step.

### 2.3 Fusion techniques

When both temporal and inter-view interpolations are available, such as in the case of a hybrid or a symmetric 1/2 scheme (Figure 2), they have to be combined in order to create a single side information. In the context of DVC, this operation is referred to as fusion.

A common reference is the ideal fusion [40]: for each pixel $\mathbf{p}$, the value of the fused image $\widetilde{I}(\mathbf{p})$ is selected as the temporal estimate $\widehat{I}_T(\mathbf{p})$ value or the inter-view estimate $\widehat{I}_V(\mathbf{p})$ value, according to the error with respect to the actual WZF. Of course, this method cannot be used in practice and serves only as a theoretical bound. Practical methods for image fusion are often based on the differences between the interpolated images and the KFs. Let us refer to the absolute difference between the temporal interpolation and the forward key frame (respectively, to the backward key frame) as $e_{\mathrm{T}}^{\mathrm{F}}$ (respectively, to $e_{\mathrm{T}}^{\mathrm{B}}$). Likewise, the difference between the inter-view interpolation and the left and right views are referred to as $e_{\mathrm{V}}^{\mathrm{L}}$ and $e_{\mathrm{V}}^{\mathrm{R}}$. In [39], for each pixel $\mathbf{p}$, the inter-view interpolation is chosen if $e_{\mathrm{V}}^{\mathrm{L}}(\mathbf{p}) < e_{\mathrm{T}}^{\mathrm{B}}(\mathbf{p})$ and $e_{\mathrm{V}}^{\mathrm{R}}(\mathbf{p}) < e_{\mathrm{T}}^{\mathrm{F}}(\mathbf{p})$. Otherwise, the temporal interpolation is used. In [41], an encoder-driven fusion is proposed. Each WZF is compared with respect to (w.r.t.) the previous and the forward frame. A binary mask is set to 0 or 1 if the value of each pixel of the WZF is closest to the co-located pixel of the previous or the forward frame, respectively. This mask is sent to the decoder by JBIG. At the decoder side, the two SI (temporal and inter-view) are compared w.r.t. the previous frame or the forward one according to the mask sent by the encoder. For each pixel, the closest SI is chosen for decoding. In [42], two fusion techniques are proposed: temporal motion (TM) interpolation projection fusion (IPF) and spatial view (SV) IPF. In TM-IPF, a block matching between the temporal estimate $\widehat{I}_T$ and the two adjacent KFs in time domain is performed. For the pixels where the two compensation errors are larger than a threshold, inter-view interpolation is used because temporal interpolation is supposed to have failed. Otherwise, the temporal interpolation is kept. SV-IPF is the counterpart of TM-IPF: $\widehat{I}_V$ is kept if its error relative to KFs is small; otherwise, $\widehat{I}_T$ is used. Guo et al. in [37] computed the absolute difference between the two motion-compensated frames in time domain. For pixels where this difference is larger than a threshold and if the norm of the two motion vectors is larger than a threshold, inter-view interpolation is chosen; otherwise, temporal interpolation is used. In [43], the decision on which SI has to be used is taken by observing the difference of the two SI on the neighboring, already decoded pixels. The correlation between the temporal/inter-view SI and the real WZF is assumed to be stationary, so they establish a criterion based on the error of the neighboring, already decoded pixels.

In [44], the fusion based on two different error images, referred to as $E_T$ and $E_V$. $E_T$, is the absolute difference between the motion-compensated backward reference and the motion-compensated forward reference. Likewise, disparity compensation on the left and right images is used to create $E_V$. Two efficient fusion techniques are proposed: in the binary fusion, the pixel value in $\mathbf{p}$ is selected from the inter-view interpolation if $E_V(\mathbf{p}) < E_T(\mathbf{p})$; otherwise, it is selected from the temporal interpolation; in the linear fusion, the coefficient $\alpha = \frac{E_T(\mathbf{p})}{E_T(\mathbf{p})+E_V(\mathbf{p})}$ is computed and $\widetilde{I}(\mathbf{p})$ is defined as a weighted average of $\widehat{I}_V(\mathbf{p})$ and $\widehat{I}_T(\mathbf{p})$ using, respectively, $\alpha$ and $1-\alpha$ as weights.

In [45], a new correlation model for the temporal and inter-view side information is proposed: the smaller the difference between the two SIs, the smaller is the variance of the probability density function (PDF) that models the correlation noise. Finally, a reconstruction with two SIs is performed such as that in [46]. Multihypothesis techniques are used by [47,48] for fusing different SIs. Three SIs are generated (by block-based techniques, by optical flow, and by overlapped block motion compensation), and three parallel decoders are used in order to decode them. They choose the SI of the decoder that firstly converges. This technique has been extended also for the WZ coding of depth maps.

Recently, support vector machine was proposed by Dufaux [49] for fusion, which is considered as a classification problem with two classes. The features are extracted from the four KFs surrounding the WZF that has to be estimated. This technique is better than the previous methods even though it surpasses binary and linear fusion just slightly. Since, in addition, the latter methods are much simpler, we use them as reference for fusion performance.

## 3 Inter-view interpolation with priority to large disparity

We introduce a novel inter-view interpolation algorithm for multiview DVC for the case of aligned cameras or rectified sequences. In this case, the inter-view estimation is commonly produced by adapting the motion-compensated temporal interpolation algorithms since the object 'trajectories' along views are practically straight lines. As a consequence, the DISCOVER algorithm works fairly well in this case, and considering higher order interpolation does not lead to significant gains. However, it does not take into account occlusions, so we can try to improve upon it by addressing this issue.

We took inspiration from the paper by Daribo and Pesquet-Popescu [50], where, in the context of multiview-plus-depth video coding, the original depth information was used to change the patch priority in the inpainting of occlusion areas in synthesized views. The key idea for our algorithm is quite similar: whenever two objects (actually, two blocks of pixels) of the reference KFs have estimated trajectories both passing near the current position, we take into account the object depth, instead of simply selecting the trajectory closer to the current position, as DISCOVER would do (see Equation 1). The foreground objects, having smaller depths, occlude the background and should be preferred by the SI generation algorithm. Therefore we modify Equation 1 by adding a penalization for small disparities (which translates into large depths). We obtain the following equation:

$$\mathbf{q}^*(\mathbf{p}) = \arg\min_{\mathbf{q}} \left( \left\| \mathbf{q} + \frac{1}{2}\mathbf{v}(\mathbf{q}) - \mathbf{p} \right\|^2 - \gamma \, \|\mathbf{v}(\mathbf{q})\|^2 \right) \tag{2}$$

where the penalization parameter $\gamma > 0$ is to be chosen experimentally. In conclusion, the proposed method consists in replacing Equation 1 of DISCOVER with Equation 2; the other steps of DISCOVER stay unaltered. We call this method 'interpolation with priority to large disparity' (IPLD).

In Figure 3, we show a schematic example where IPLD allows the selection of the right trajectory. In this figure, the $y$ axis of frames is orthogonal to the drawing plane. After the first disparity estimation, the disparities[b] for blocks centered in $\mathbf{q}_1$ and $\mathbf{q}_2$ are respectively $\mathbf{v}(\mathbf{q}_1)$ and $\mathbf{v}(\mathbf{q}_2)$. Both blocks $B_1$ and $B_2$ could be interpolated in position $\mathbf{p}$ in the WZF. DISCOVER would simply select the block whose trajectory is closer to $\mathbf{p}$, $B_1$ in this case. On the contrary, we should select the block that occludes the other, $B_2$ in this case. In order to estimate the occlusion, we penalize blocks with small disparities since they probably belong to the background. Thus, provided that a suitable value for the penalization parameter $\gamma$ is chosen, IPLD is able to correctly select the block $B_2$.

This algorithm is well suited for the inter-view estimation because, in this case, the disparity is strongly related to the depth. However, it would not be as much efficient for temporal interpolation since objects with higher velocity do not necessarily occlude objects with smaller velocity, even though if two objects have the same speed, the one closer to the camera will have a larger apparent velocity. This intuition is confirmed by preliminary experiments.

The increase of computational complexity of IPLD with respect to DISCOVER is negligible: when Equation 1 is replaced by Equation 2, for each candidate trajectory, we only need the computation of $-\gamma\|\mathbf{v}(\mathbf{q})\|^2$ and an additional sum. This computational cost is dominated by two multiplications. Now, since in the splitting step only a small number of candidate trajectories is considered, we estimate the increase of computational complexity of our method in less than 20 multiplications per block. This is to be compared to the much heavier cost of motion estimation in the DISCOVER algorithm.

In Figure 4, we show an inter-view-interpolated image from the 'newspaper' sequence, using DISCOVER (Figure 4a) and IPLD (Figure 4b): our technique is much more successful in discriminating the background from the foreground.

## 4 Proposed fusion method

Since HOMI and IPLD allow the production of good temporal and inter-view interpolations of images, as a first novel contribution, we consider a multiview DVC system that just uses an existing fusion algorithm (such as those proposed in [44]) on these improved SI images; we expect that such a solution outperforms a reference system based on state-of-the-art interpolation methods. However, we can further improve the system performance with a novel fusion scheme based on occlusion detection.

Moreover, as it was observed in the past (e.g., in [44]), even though for the majority of test video sequences fusion improves upon temporal and inter-view interpolations, for some others, the opposite is true. In other words, there are sequences for which temporal or inter-view interpolation is much better than fusion; as a consequence, the rate-distortion (RD) performance comparison between fusion and interpolation is not very consistent even though, on the average, fusion has been reported for being the best solution.

In the second part of this section, we propose a novel Bayesian method for deciding among fusion, temporal interpolation or inter-view interpolation, based on the observation of the parity bit rate needed by the decoder. The resulting adaptive fusion method shows a much more consistent behavior and improved performance with respect to previous techniques.

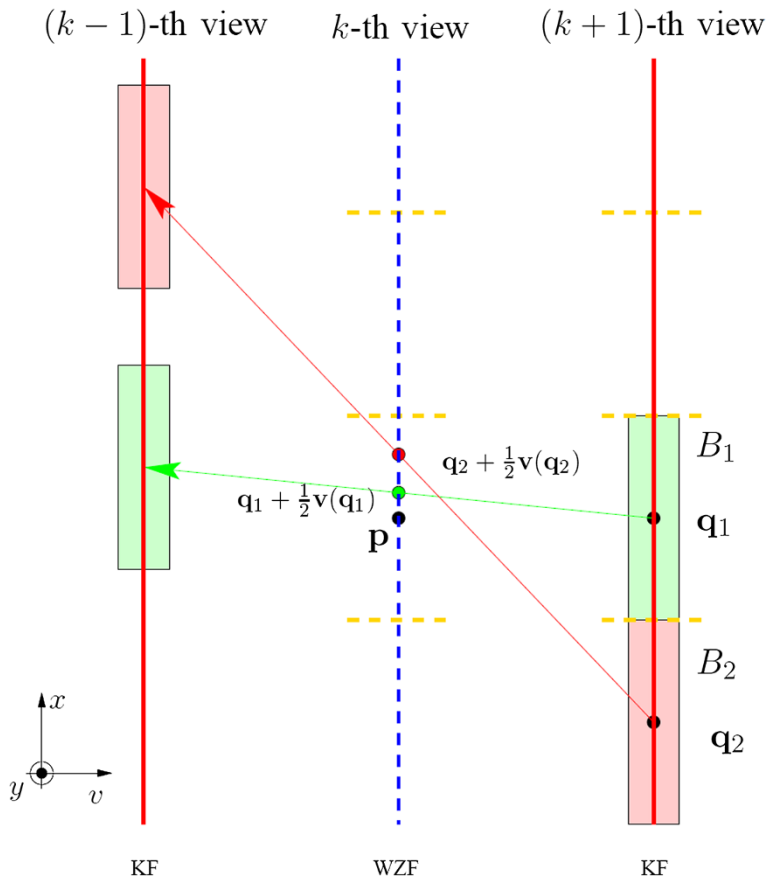


**Figure 3 A 1D representation for the IPLD algorithm.** A 1D representation of a possible situation of two disparity vectors after the forward disparity estimation of DISCOVER. With DISCOVER, we would choose the green vector in the splitting procedure because it is closer to **p** than the red vector, but arguably, the object in the red block occludes the object in the green block within the frame in the *k*th view.

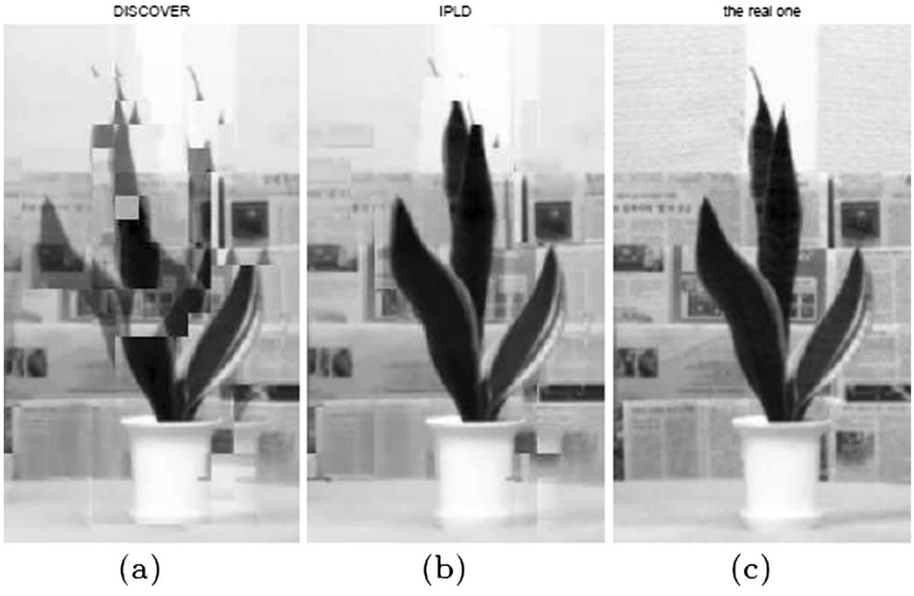


**Figure 4 A detail of the 'newspaper' sequence.** Inter-view interpolation generated **(a)** by DISCOVER (peak signal-to-noise ratio (PSNR) = 15.1 dB), and **(b)** by IPLD (PSNR = 28.3 dB). The original image is also shown **(c)**.

### 4.1 Occlusion detection-based fusion

For the fusion problem, we consider a scheme where a WZF $I_{t,k}$ has four neighboring KFs: $I_{t,k-1}$, $I_{t,k+1}$, $I_{t-1,k}$, and $I_{t+1,k}$. For example, we can consider an hybrid camera in a hybrid 1/2 scheme (Figure 2b) or any camera (except the first and the last) in a symmetric 1/2 scheme (Figure 2c). For the time being, any temporal interpolation and inter-view interpolation algorithms can be used to produce the partial estimates, $\widehat{I}_T$ and $\widehat{I}_V$. The proposed fusion between these images is performed as follows.

Let us suppose that we have the disparity fields $d_{k-1,k+1}(m,n)$ (from $I_{t,k-1}$ to $I_{t,k+1}$) and $d_{k+1,k-1}(m,n)$ (from $I_{t,k+1}$ to $I_{t,k-1}$). They can be obtained as side product of the inter-view interpolation process, or they can be estimated from scratch for the fusion process. In our implementation, we compute the disparity fields at the decoder from scratch using the method proposed by Miled and Pesquet [51] on the received KFs.

The coherence of the disparity fields is estimated by a left-right consistency crosscheck (LR-CC) [52-54]: if the pixel $\mathbf{p} = (m,n)$ in the image $k_1$ corresponds to the pixel $(m, n + d_{k_1,k_2}(m,n))$ in the image $k_2$, the disparity of the former should be the opposite of the disparity of the latter. Therefore, the absolute sum of these quantities is a measure of the disparity coherence in the direction from $k_1$ to $k_2$. We can write

$$R_{k_1,k_2}(m,n) = |d_{k_1,k_2}(m,n) + d_{k_2,k_1}(m, n + d_{k_1,k_2}(m,n))| \tag{3}$$

For a perfectly coherent disparity, we would obtain $R = 0$ everywhere. However, we take into account a tolerance term $\tau \geq 0$. In the scientific literature, a common value for $\tau$ is 1 pixel [55,56], at least in the context of stereo matching. Since our target is slightly different (image fusion in the context of multiview DVC), we do not use the value in the literature but rather optimize it experimentally. We define the occlusion map from $k_1$ to $k_2$ as follows:

$$O_{k_1,k_2}(m,n) = u(R_{k_1,k_2}(m,n) - \tau) \tag{4}$$

where $u(\cdot)$ is the left continuous Heaviside step function. Using Equation 4, we perform an occlusion detection: $O_{k_1,k_2}$ shows the points of $k_1$ occluded (i.e., not visible) in $k_2$.

Finally, we use the occlusion maps $O_{k-1,k+1}$ and $O_{k+1,k-1}$ to improve the fusion process. For each pixel $\mathbf{p}$, we decide whether to use only the temporal interpolation (since an occlusion has been detected) or to use a fused image, such as the linear fusion proposed in [44]. Now, the inter-view estimate $\widehat{I}_V(\mathbf{p})$ has been produced (using DISCOVER or IPLD) as the average of two disparity-compensated values:

$$\widehat{I}_V(m,n) = \frac{1}{2} I_{t,k-1}(m, n+e(m,n)) + \frac{1}{2} I_{t,k+1}(m, n+f(m,n)) \tag{5}$$

where $e$ and $f$ are respectively the estimates of $d_{k,k-1}$ and $d_{k,k+1}$, and they have been obtained by splitting.

Looking at Equation 5, we argue that $\widehat{I}_V(\mathbf{p})$ is affected by the possible occlusion of the pixel $(m, n + e(m,n))$ in the left image and of the pixel $(m, n + f(m,n))$ in the right image. Therefore, the proposed fusion formula is as follows:

$$\widetilde{I}(\mathbf{p}) = \begin{cases} \widehat{I}_T(\mathbf{p}) & \text{if } O_{k-1,k+1}(m, n+e(m,n)) = 1 \\ \widehat{I}_T(\mathbf{p}) & \text{if } O_{k+1,k-1}(m, n+f(m,n)) = 1 \\ I_{\mathrm{LIN}}(\mathbf{p}) & \text{otherwise} \end{cases}$$

In other words, we use only the temporal interpolation if one or both of the two pixels that contribute to $\widetilde{I}(\mathbf{p})$ is estimated as occluded. Otherwise, we use the linear fusion proposed in [44], indicated as $I_{\mathrm{LIN}}(\mathbf{p})$, with the only difference that we merge the estimates obtained by HOMI and IPLD instead of those obtained by DISCOVER.

We observe explicitly that in order to obtain our estimate of the WZF, we use four disparity fields ($d_{k-1,k+1}$, $d_{k+1,k-1}$, $e$, and $f$), and this feature characterizes our method with respect to existing techniques based on LR-CC.

### 4.2 Adaptive validation

As shown in the scientific literature and confirmed in our tests, even though fusion improves the quality of side information on the average, for some sequences, this is not true; on the contrary, the image generated by fusion is much worse than the one generated by inter-view interpolation only or temporal interpolation only. This happens for example for sequences where one interpolation is much better than the other (as shown for the 'outdoor' sequence in [49] and in our experimental section). For these sequences, the correlation along one axis (temporal or inter-view) is much stronger than the correlation along the other.

Therefore, we would need a method to decide which image among $\widehat{I}_T$, $\widehat{I}_V$, and $\widetilde{I}$ should be used as side information. This process can be seen as a final step of the fusion process, which validates $\widetilde{I}$ or goes back to $\widehat{I}_T$ or $\widehat{I}_V$: we call the proposed method adaptive validation (AV).

We recall that the best side information is the one that requires the smallest bit rate in order to be corrected by the channel decoder. Let us consider a given WZF $I_{t,k}$ and let us call $R_T$, $R_V$, and $R_F$, the bit rate (in bits per pixel) for correcting temporal, inter-view, and fused SIs, respectively. The central idea of the proposed method is very simple: we take the decision among the three strategies based on observing, for a small number of WZFs, the parity bit rate needed to correct the temporal and inter-view interpolations. The rationale behind this idea is that fusion is ineffective only when one of the two interpolations is much worse than the other. Therefore, it suffices

to observe the two rates $R_T$ and $R_V$ to estimate whether fusion is viable or not, while the value of $R_F$ is less relevant to this end.

Let us first discuss about the rate overhead associated to the evaluation of $R_T$ and $R_V$. This is performed by actually asking the turbo encoder to send the parity bits needed to correct the temporal and the inter-view side information. More precisely, we use the turbo encoder of DISCOVER that consists of two interleaved punctured 1/2-convolutional encoders. The DCT bands are quantized according to the quantization index, and each DCT band is independently coded. Let $P$ be the puncturing period (usually set to 48). Then, for each DCT band, the parity bit stream is structured into sub-blocks of size $\frac{MN}{16P}$ bits (one for each convolutional encoder) [57], where $M \times N$ is the spatial resolution of the frames and 16 is the number of pixels per block. At the first step, only a sub-block for each DCT band and for each bit plane is sent. Then, if needed, the decoder requests (via the feedback channel) a certain number of sub-blocks for the $b$th band for the $p$th bit plane. We call this number $N_j(p,b)$, where the index $j$ can be $T$ or $V$, in order to identify respectively the case of temporal or inter-view SI.

The total bit rate per pixel ($R_j$ with $j \in \{T, V\}$) would be

$$R_j = \frac{1}{MN} \sum_{b=1}^{16} \sum_{p=1}^{P(q_i,b)} \left( 2\frac{MN}{16P} + 2N_j(p,b)\frac{MN}{16P} \right) \qquad (6)$$

where $P(q_i, b)$ is the number of bits used for quantizing the $b$th band depending on the quantization index $q_i$. The number of parity bits sent to the decoder in order to correct the side information is given by two contributions: an amount depending only on the image resolution and on the quantization index that we will call $R_0$ and a variable amount depending on the SI quality expressed in terms of parity bit requests per band per bit plane.

Then, for two different SIs (temporal and inter-view), the value of $N_j(b,p)$ may be different for each bit plane and band. In the worst case, the sets of parity bits are disjoint (except for the common part $R_0$), so the rate needed to send both sets is $R_T + R_V - R_0$. In the most favorable case, one set of parity bits is a subset of the other; then, it suffices to send the parity bits needed for the worst case: the rate is $\max(R_T, R_V)$. In the general case, we need a bit rate between these two extreme cases.

Let us show the operation of the proposed system with an example taken from the encoding of the 'book arrival' sequence. We use a symmetric 1/2 scheme and consider image number 1 from view 1, encoded as a WZF. The key frames have been encoded with QP = 31. In Table 1, the values of $N_T(p,b)$ and $N_V(p,b)$ for the corresponding SI images $\widehat{I}_T$ and $\widehat{I}_V$, generated using images 0 and 2 of view 1 and image 1 of views 0 and 2, respectively. The numbers of requests in italic means that $N_V(p,b) > N_T(p,b)$, the number of requests in bold means that $N_V(p,b) < N_T(p,b)$, and for the others, $N_V(p,b) = N_T(p,b)$.

**Table 1 Number of requests per bit plane and per band**

| | $N_T(p,b)$ | | | | | $N_V(p,b)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Bitplane | | | | | Bitplane | | | | |
| **Band** | **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** |
| 1 | 0 | 1 | 3 | 4 | 6 | 0 | 1 | 3 | 4 | 6 |
| 2 | **3** | *0* | *0* | *3* | 0 | **2** | *1* | *1* | *4* | 0 |
| 3 | *2* | 0 | *0* | *3* | 0 | *3* | 0 | *1* | *4* | 0 |
| 4 | *1* | 0 | *1* | 0 | 0 | *2* | 0 | *2* | 0 | 0 |
| 5 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 6 | **2** | 0 | 2 | 0 | 0 | **1** | 0 | 2 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | *0* | 0 | 0 | 0 | 0 | *2* | 0 | 0 | 0 |

Number of requests (sequence 'book arrival', QP = 31; view 1, frame 1) split up per bit plane and per DCT band. Left [right]: number of bits sent for correcting the temporal [inter-view] SI DCT coefficients; in bold and in italic, coefficients that need more correction bits when estimated respectively from $\widehat{I}_T$ and from $\widehat{I}_V$.

It results that, for correcting the temporal SI, the encoder would send $R_T = 0.164$ bits per pixel (bpp). For correcting the inter-view SI, the encoder would send $R_V = 0.185$ bpp. Since a subset of the parity bits is in common for the two SIs, once, for example, the temporal parity bits have been sent, we need only $R_{\text{estimation}} = 0.216$ bpp for correcting the inter-view SI. In conclusion, with this rate, the decoder is able to know $R_T$ and $R_V$. The overhead for this operation is given by the difference between $R_{\text{estimation}}$ and the smallest among $R_T$, $R_V$, and $R_F$. In this case, using the occlusion detection (OD) fusion, we would have obtained $R_F = 0.135$ bpp, so the estimation overhead is $R_{\text{estimation}} - R_F = 0.081$ bpp.

The key point is to ask for this additional rate only for a subset of frames and to use the information about the parity rates $R_T$ and $R_V$ to decide, for the following WZFs, which side information to use. In other words, if we request the two sets of parity bits for all the WZFs, we incur in a large rate overhead and end up with a very inefficient system, which, in the best case, matches the performance of the worst side information. On the contrary, instead of using $R_T + R_V - R_0$ or $\max(R_T, R_V)$ bits, we would like to use a bit budget of $R_D$ bits, where $D$ is the optimal decision, i.e.,

$$D = \arg \min_{d \in \{T,V,F\}} R_d \qquad (7)$$

We estimate the optimal decision by observing the additional parity bit rate only for $n$ over $N$ WZFs, and for the following $N - n$, we use this decision. More precisely, for

$n$ frames out of $N$, we ask the channel coder the correction bits for both $\widehat{I}_T$ and $\widehat{I}_V$. We call $\delta_R = R_T - R_V$ the difference between the two rates. We expect the value of $\delta_R$ to be quite correlated to the optimal decision: if this parameter is large in absolute value, one estimate is much better than the other, so it should be used; if it is small, the two interpolations have close quality, so the fusion should be used. This threshold-based approach is actually equivalent to a maximum *a posteriori* (MAP) estimation of the optimal decision, as shown below.

Let us model the optimal decision $D$ as a discrete random variable with three possible values $\{T, V, F\}$. We want to estimate the best decision (according to Equation 7) given the parameter $\delta_R$ modeled as a continuous random variable. We use the MAP value obtained by applying the Bayes' law:

$$\begin{aligned}\widehat{D} &= \arg\max_d P(D = d|\delta_R) \\ &= \arg\max_d \frac{p(\delta_R|D = d)P(D = d)}{p(\delta_R)} \\ &= \arg\max_d p(\delta_R|D = d)P(D = d) \end{aligned} \quad (8)$$

where $P(\cdot)$ is the probability mass function (marginal or conditional) for $D$ and $p(\cdot)$ is the probability density function for $\delta_R$ (marginal or conditional). In order to solve this problem, we have to find the three functions $f_d(\delta_R)$ for $d \in \{T, V, F\}$, defined as

$$f_d(\delta_R) = p(\delta_R|d)P(d).$$

These functions are estimated off-line on a training set.

A couple of further issues has to be addressed, as illustrated in Section 5.3 in order to be able to use this method. First, as it was observed in some preliminary tests, the $f_d$ functions vary with the QP, so different thresholds must be determined as its function. Second, we have to decide how frequently to update the decision $\widehat{D}$ during the coding process. We compute $\delta_R$ for $n$ WZFs out of $N$ frames. For these $n$ frames, we pay an overhead rate cost since we ask twice the channel bits to the encoder. Therefore, we cannot update $\delta_R$ too often; otherwise, we lose all the advantage of having a better side information. On the other hand, a too small ratio $n/N$ could affect the reliability of the decision. This dilemma is solved experimentally, but we can anticipate that small values of the $n/N$ ratio (in the order of 1/100) work very well if scene changes do not occur.

## 5 Experimental results

In this section, we present the experimental results of the proposed SI generation and fusion methods. We consider a DISCOVER DVC system and modify the side information generation and fusion steps, using the methods proposed in the previous sections. The other tools remain the same: turbo codes are used for the WZFs; the correlation error is modeled as a Laplacian random variable. The statistical properties are evaluated on the residual error between the forward and the backward motion-compensated frames (or the left and the right disparity-compensated frames for inter-view correlation). The KFs are coded with the INTRA codec of H.264/AVC with four QPs: 31, 34, 37, and 40.

In order to evaluate the contributions of our techniques, we consider several configurations.

First, we consider the inter-view interpolation method IPLD: at this end, we use an asymmetric scheme (Figure 2a), such that there is no temporal interpolation or fusion whose performance could affect the evaluation. Only the performance on the WZ views is considered.

Then, we consider the global fusion scheme, made up of the temporal interpolation, the inter-view interpolation, and the fusion algorithm; in order to isolate the contributions of the occlusion detection fusion and of the adaptive validation, we consider separately the cases when the latter is turned off or on. For these experiments, we need a scheme where both temporal and inter-view interpolations are available: we can consider the even views in a hybrid 1/2 scheme (Figure 2b) or all but the side views in a symmetric 1/2 scheme (Figure 2c).

Therefore, we now evaluate separately the contributions of IPLD, occlusion detection, and AV; for each of the three methods, we also report the experiments performed in order to tune their parameters. Then, we validate their effectiveness on the end-to-end rate-distortion performance for the encoded view(s). The latter is measured using the Bjontegaard metrics [58] with respect to the reference DISCOVER system: both the peak signal-to-noise ratio (PSNR) variation ($\Delta_{\mathrm{PSNR}}$) and the percent rate variation ($\Delta_{\mathrm{R}}$) are reported. For the IPLD technique, we also report the SI quality (measured in terms of PSNR with respect to the original WZF).

These experiments have been performed on a test set composed of nine popular multiview sequences, listed in Table 2. They are characterized by different baseline distances, spatial resolutions, disparity ranges, and amount of motion in order to represent a sufficiently wide range of experimental conditions.

### 5.1 Inter-view interpolation results

For all the experiments in this subsection, we consider an asymmetric scheme (Figure 2a).

In the first set of experiments, we have to set the value of parameter $\gamma$ of Equation 2, which defines the penalization for small disparity (i.e., background) blocks. In order to set the value of $\gamma$, we test our algorithm on five rectified multiview video sequences, with 20 frames per sequence. The neighboring views are encoded with H.264/AVC INTRA. In our experiments, the value maximizing the PSNR of the

**Table 2 The multiview sequences test set**

| Sequence | Resolution |
|---|---|
| Balloons[a] | 1,024 × 768 |
| Book arrival[b] | 512 × 384 |
| Door flowers[b] | 1,024 × 768 |
| Leaving laptop[b] | 1,024 × 768 |
| Kendo[a] | 1,024 × 768 |
| Lovebird[c] | 1,024 × 768 |
| Newspaper[d] | 1,024 × 768 |
| Pantomime[a] | 1,280 × 960 |
| Outdoor[b] | 512 × 384 |

The multiview sequences test set. Sources: [a]Tanimoto Laboratory; [b]Heinrich Hertz Institute; [c]ETRI/MPEG Korea Forum; [d]Gwangju Institute of Science and Technology.

SI with respect to the WZF is $\gamma = 0.6$. By experiments, we have found that the optimal value of $\gamma$ is negligibly affected by the QP. Therefore, we keep this value.

In the second experiment, we compute the SI for all the test sequences with 100 frames per sequence at various QP for the reference KFs. The resulting PSNR (with respect to the original WZF) is compared with the one achieved by DISCOVER, and the difference is reported in Table 3. We observe that the proposed method improves almost always the DISCOVER quality, sometimes with very significant gains (up to 3.34 dB). This is because, as expected, our method succeeds at well reconstructing objects belonging to the foreground. We observe that for sequences such as 'lovebird', 'pantomime', 'outdoor', the proposed method does not improve with respect to DISCOVER since they have a small range of disparity values; in such cases, using disparity to discriminate foreground and background is less successful. On the contrary, when the disparity range is larger, like in 'newspaper' and 'balloons', the improvement is remarkable.

**Table 3 SI improvement (dB): IPLD versus DISCOVER**

| Sequence / QP | 31 | 34 | 37 | 40 |
|---|---|---|---|---|
| Balloons | 2.21 | 2.26 | 2.23 | 2.23 |
| Book arrival | 0.09 | 0.46 | 0.43 | 0.41 |
| Door flowers | 0.86 | 0.70 | 0.65 | 0.50 |
| Leaving laptop | 0.98 | 0.88 | 0.60 | 0.55 |
| Kendo | 1.10 | 0.92 | 0.80 | 0.75 |
| Lovebird | −0.52 | −0.33 | −0.28 | −0.17 |
| Newspaper | 3.34 | 3.13 | 3.05 | 2.95 |
| Pantomime | 0.00 | 0.00 | −0.11 | −0.11 |
| Outdoor | −0.10 | −0.02 | −0.02 | −0.05 |
| *Mean* | 0.88 | 0.89 | 0.81 | 0.78 |

In the last experiment, we compare the end-to-end rate-distortion performance of a complete DVC system using IPLD for inter-view interpolation with another using DISCOVER, i.e., the temporal interpolation described in Section 2 is applied along the view domain. The results, shown in Table 4, are related to the Wyner-Ziv camera only since the encoding of the KFs does not change in the two cases. We observe that the large improvement in SI quality shown in Table 3 is actually translated into a significant bit rate reduction for the end-to-end system, in particular for sequences such as 'newspaper' and 'balloons', where background and foreground are hard to be distinguished. For these sequences, we obtain a bit rate reduction up to more than 20% and a gain in PSNR of around 1 dB. As for the previous experiments, for the sequences with small disparity range, we observed some small losses. However, on average, the IPLD allows a bit rate reduction of more than 7.7% achieved with a very small complexity increase.

### 5.2 Results for fusion with occlusion detection

Fusion techniques can be used when both temporal and inter-view interpolations are available; therefore, we can consider the hybrid cameras in a hybrid 1/2 scheme or all the views (except for the first and the last) in a symmetric 1/2 scheme. The side information produced in such a scheme depends on the two interpolation techniques and on the fusion. In order to isolate the contribution of each element to the global performance, we consider the following cases:

- *DT*. SI generated with DISCOVER along the temporal axis;
- *HT*. SI generated with HOMI along the temporal axis;
- *DV*. SI generated with DISCOVER along the view axis;

**Table 4 RD performance of IPLD vs. DISCOVER**

| | $\Delta_R$ (%) | $\Delta_{PSNR}$ (dB) |
|---|---|---|
| Balloons | -22.71 | 1.05 |
| Book arrival | -5.36 | 0.53 |
| Door flowers | -5.40 | 0.24 |
| Leaving laptop | -7.62 | 0.32 |
| Kendo | -9.26 | 0.50 |
| Lovebird | 2.38 | -0.09 |
| Newspaper | -22.09 | 0.97 |
| Pantomime | 0.19 | 0.00 |
| Outdoor | 0.35 | -0.03 |
| *Mean* | −7.72 | 0.38 |

Bjontegaard metrics: IPLD versus DISCOVER (for the WZF camera in an asymmetric scheme).

- *IV*. SI generated with IPLD along the view axis;
- *FD*. SI obtained by linear fusion of DV and DT as in [44];
- *FHI*. SI obtained by linear fusion of HT and IV;
- *OD*. FHI with the occlusion detection;
- *AV*. OD with adaptive validation of fusion.

Comparing all these methods allows us to have a deeper insight on the contribution of all the proposed tools to the final performance. The FD method is considered the reference since its performance is better than much of the state of the art and only slightly worse than (but very close to) more recent fusion techniques, as shown in [49].

The first experiments are conducted to set the disparity tolerance $\tau$ in Equation 4. To this end, we evaluate the PSNR of the SI information generated with the OD method for 10 frames of seven test sequences for four different QPs. We repeat the test while varying the value of $\tau$, and the resulting SI PSNR is reported in Figure 5. We find that the best value is $\tau = 2$, and this is independent from the QP.

Now, in order to assess the impact of OD, we compute the RD performance of OD and FHI with respect to the reference FD [44]. The results are shown in Table 5. We observe that just using better temporal and inter-view interpolations than DISCOVER (FHI column) allows a non-negligible gain in the average (1.6% bit rate reduction). These results are fairly improved when the occlusion detection technique is used. On the average, we gain almost another 5%, achieving a remarkable 6.57% rate reduction. However, for some sequences, the OD approach actually worsens the performance due to false-positive errors on occlusion detection.

Some insights about the impact of different methods are shown in Figure 6. In the first row, we show a detail of the 'balloons' sequence, for which three views are available (left, center, and right). We notice that there are several occluded areas between the two external views, such as the musical notes on the background. In this case, the block-based IPLD (bottom-left figure) does not provide a good estimate of the central view: the blocks near the balloon contours are badly estimated. Using fusion (in particular FHI), we achieve a significant improvement, as shown in the bottom-center figure. The pixel-wise adaptivity of the fusion allows the use temporal interpolation where the inter-view is judged as not reliable; yet, there are some residual artifacts since not all the occlusions are correctly detected. When we apply the occlusion detection technique, we can then exclude the inter-view contribution from the fusion for the affected pixels, and this further improves the quality of the SI (bottom-right figure).

### 5.3 Adaptive validation results

Now we consider the AV technique. The first experiments are devoted to the estimation of the distribution of $\delta_R$, the rate difference between the corrections of $\widehat{I}_T$ and $\widehat{I}_V$, given the optimal decision $D$. This process is performed off-line. We consider a training set of 100 frames from seven multi-view sequences. For each frame, we compute the temporal interpolation $\widehat{I}_T$, the inter-view interpolation $\widehat{I}_V$, and the fusion with the OD algorithm, $\widetilde{I}$. Then, we evaluate the number of parity bits needed to correct

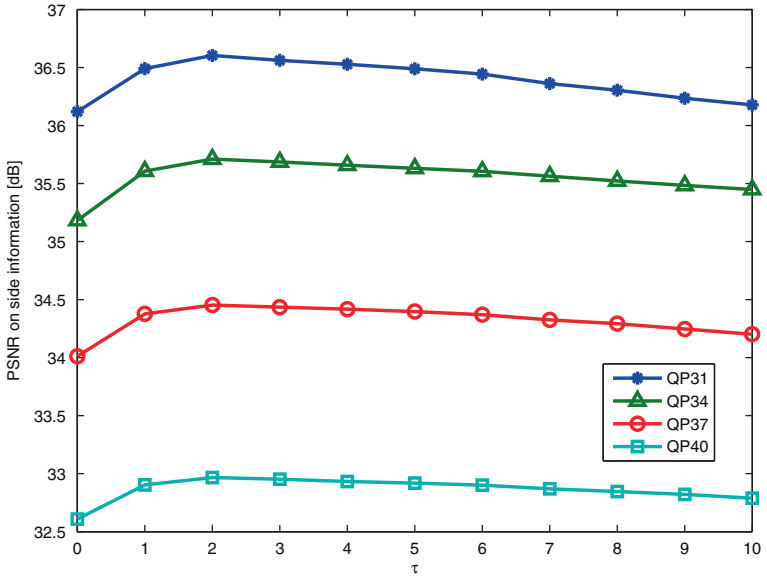


**Figure 5 PNSR of the fused SI.** PNSR of the fused SI versus the value of $\tau$ for different QPs.

**Table 5 RD performance of the fusion methods vs. state-of-the-art technique [44]**

| | FHI | | OD | |
|---|---|---|---|---|
| Sequence | $\Delta_R$ (%) | $\Delta_{PSNR}$ (dB) | $\Delta_R$ (%) | $\Delta_{PSNR}$ (dB) |
| Balloons | -9.42 | 0.51 | -29.69 | 1.66 |
| Book arrival | -0.96 | 0.06 | -2.73 | 0.16 |
| Door flowers | -0.22 | 0.01 | -6.12 | 0.29 |
| Leaving laptop | -2.18 | 0.09 | -3.66 | 0.17 |
| Kendo | -1.76 | 0.09 | -2.67 | 0.16 |
| Lovebird | 0.46 | -0.02 | -6.73 | 0.30 |
| Newspaper | -4.27 | 0.23 | -18.78 | 0.98 |
| Pantomime | 3.08 | -0.17 | 4.18 | -0.23 |
| Outdoor | 0.91 | -0.06 | 7.08 | -0.45 |
| *Mean* | −1.60 | 0.08 | −6.57 | 0.34 |

the three images. This information gives us both the optimal decision (the one associated to the minimum rate) and the $\delta_R$ parameter. Repeating this operation for all the frames and for the four QP values, we obtain a large set of samples from the distribution of $\delta_R|D$, along with the samples from the distribution of $D$. The relative frequencies of the latter constitute the estimate of the PMF of $D$. The samples of $\delta_R|D$ can be used to estimate the corresponding PDF, using the Parzen window method [59]. We use thus a non-parametric estimate of these PDFs: this is a reasonable approach since our *a priori* knowledge of the problem hardly suggests a mathematical model for these distributions.

We are now able to compute the $f_d$ functions: in Figure 7, we show them for QP = 37. It is obvious that selecting the function with the maximum value for a given $\delta_R$ amounts to compare this parameter with a couple of thresholds, given by the intersections of the $f_d$ curves. Our experiments allow us to find these thresholds for different QP values. These values, reported in Table 6, have been estimated off-line and will be used for all the sequences. We also observe that the data used for these estimates and the data used for the validation form two disjoint sets.

For the running example given in Section 4.2, we would have obtained $\delta_R = -0.021$ bpp, assuming $n = 1$, i.e., only one image is used to estimate the rates. Since the QP is equal to 31, Table 6 makes us conclude that the best decision is to use the SI generated by fusion. This decision will be kept for a set of $N$ frames, and therefore, the overhead of 15,925 bits is to be shared among all these images.

We also have to find the optimal update frequency for $\delta_R$, i.e., the optimal values for parameters $n$ and $N$ defined in Section 4.2. In order to do this, we simply run the adaptive fusion algorithm with different values for these parameters as shown in Table 7. We find that the global RD performance (measured by the Bjontegaard metrics with respect to our reference FD) improves when $N$ increases and when $n$ decreases. In particular we observe that they roughly depend on the ratio $n/N$ rather than separately on these two parameters. In conclusion, this

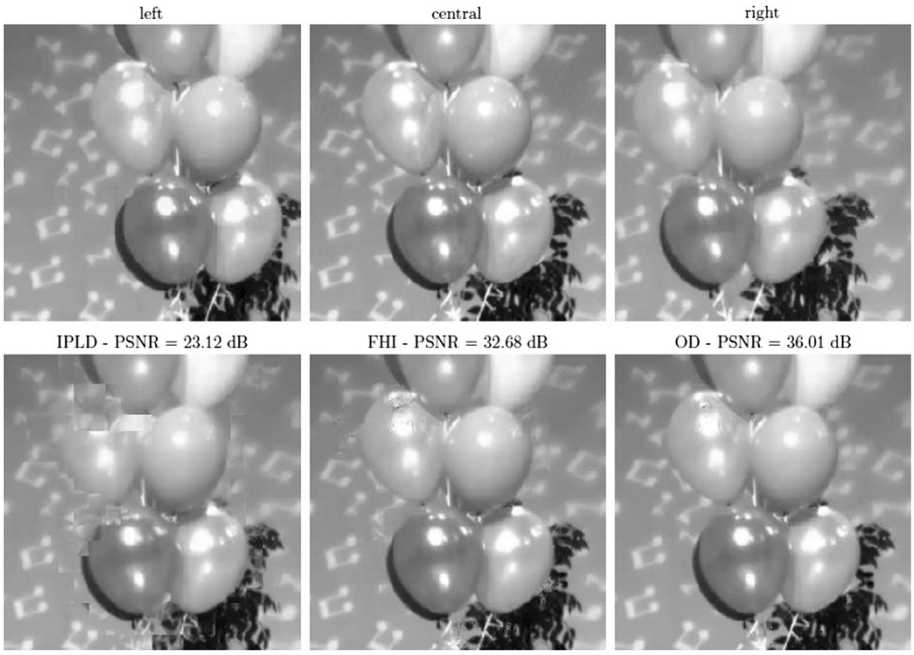


**Figure 6 Detail of the 'balloons' sequence.** In the first row, a detail of the left, the central, and the right view of sequence, are shown. The central one has to be estimated. In the second row, the SI obtained by IPLD, FHI, OD are shown.

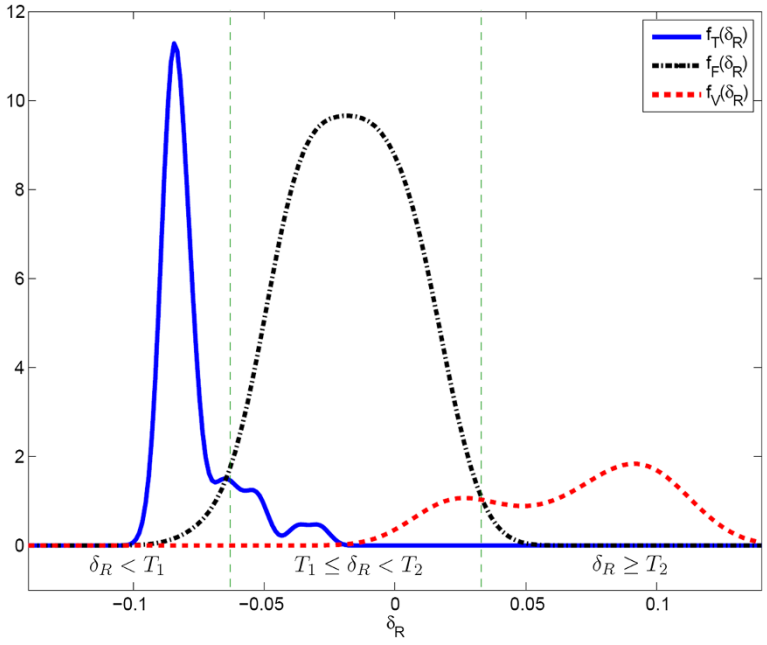


**Figure 7 Functions $f_d(\delta_R)$ for QP = 37.** If $\delta_R < T_1$, temporal interpolation is chosen; if $T_1 \leq \delta_R < T_2$, we decide for fusion; and if $\delta_R \geq T_2$, the inter-view interpolation is used.

experiment allows us to conclude that, even if we have a large $N$ and we use as few as $n = 1$ image to take the decision, we have very good performance, with a very small overhead since $n/N$ is small. This means that the optimal decision for fusion stays almost unchanged within a given sequence, and then the rate overhead due to a too frequent update is not worth. Therefore, in the following we use $n = 1$ and $N = 100$, and so we have a (small) rate overhead only for one over one hundred frames. Nevertheless, this allows us to take the best decision for the next WZFs.

For the running example, the rate overhead needed to perform the adaptive validation amounts to 15,925 per $N = 100$ images, i.e., to 0.00081 bits per pixel.

These results seem to point out that the update frequency could be as low as the scene-change frequency since, within a single sequence, the best decision seems not to change. As a consequence, another approach could be envisaged: we could update $\delta_R$ only when a scene change is detected.

These considerations are validated by introducing the concept of oracle, which is a decisor that *a priori* knows which the best side information is among temporal, inter-view, and fusion. Consequently, it uses directly the best SI without paying the cost of sending the redundant parity bits. The oracle is thus equivalent to the AV algorithm with $n = 1$ and $N = 1$ but with zero overhead. Of course, on one hand, the oracle is not realizable in practice; on the other hand, none of the proposed techniques can outdo it. Therefore, it is rather useful to understand whether the proposed estimator works well or not. In Table 8, for each sequence, we list the decision provided by the AV algorithm and compare them to the decisions taken frame-by-frame by the oracle. We observe that for seven sequences out of nine, the oracle chooses almost always the same SI, which is the one decided by the AV method. This is quite a good news: for these sequences, the proposed method is able to pick the best SI even though it observes $\delta_R$ only for one image over 100. For the other sequences, even if the AV does not choose the best SI, this does not affect the performance a lot since this happens

**Table 6 Optimal values of $T_1$ and $T_2$ estimated off-line**

| QP | 31 | 34 | 37 | 40 |
|---|---|---|---|---|
| $T_1$ | -0.13 | -0.08 | -0.06 | -0.03 |
| $T_2$ | 0.06 | 0.05 | 0.04 | 0.03 |

**Table 7 RD Performance of AV w.r.t. FD [44] for different values of *n* and *N***

| N | $n = 1$ | | $n = 3$ | |
|---|---|---|---|---|
| | $\Delta_R$ (%) | $\Delta_{PSNR}$ (dB) | $\Delta_R$ (%) | $\Delta_{PSNR}$ (dB) |
| 10 | −4.17 | 0.25 | 6.92 | −0.33 |
| 20 | −6.93 | 0.40 | −0.90 | 0.10 |
| 50 | −8.49 | 0.49 | −5.11 | 0.33 |
| 100 | −9.11 | 0.42 | −7.70 | 0.45 |

Average rate-distortion performance for AV with respect to FD [44] for different values of *n* and *N*.

**Table 8 Decision for each sequence**

| | | Oracle | | |
|---|---|---|---|---|
| | **Decision** | **Time (%)** | **View (%)** | **Fusion (%)** |
| Balloons | Time | 90 | 0 | 10 |
| Book arrival | Fusion | 2 | 0 | 98 |
| Door flowers | Fusion | 24 | 0 | 76 |
| Leaving laptop | Fusion | 4 | 0 | 96 |
| Kendo | Fusion | 48 | 28 | 24 |
| Lovebird | Time | 100 | 0 | 0 |
| Newspaper | Time | 98 | 0 | 2 |
| Outdoor | View | 0 | 76 | 24 |
| Pantomime | Fusion | 30 | 48 | 22 |

Decision for each sequence and percentage of time-, view-, and fusion-interpolated frames for optimal decision.

**Table 10 RD performance of adaptive validation w.r.t. vs. [44] and H.264/AVC INTRA**

| **Sequence** | **AV w.r.t. FD [44]** | | **AV w.r.t. H.264/AVC INTRA** | |
|---|---|---|---|---|
| | $\Delta_R$ **(%)** | $\Delta_{PSNR}$ **(dB)** | $\Delta_R$ **(%)** | $\Delta_{PSNR}$ **(dB)** |
| Balloons | −37.03 | 2.20 | −27.15 | 1.72 |
| Book arrival | −2.08 | 0.13 | −28.53 | 1.78 |
| Door flowers | −5.01 | 0.24 | −22.39 | 1.04 |
| Leaving laptop | 0.21 | −0.06 | −12.65 | 0.54 |
| Kendo | −1.59 | 0.10 | 6.51 | −0.32 |
| Lovebird | −16.34 | 0.80 | −44.44 | 2.17 |
| Newspaper | −24.89 | 1.34 | −31.22 | 1.71 |
| Pantomime | 4.65 | −0.26 | −2.30 | 0.09 |
| Outdoor | 0.08 | 0.00 | −39.27 | 2.42 |
| *Mean* | −9.11 | 0.42 | −22.38 | 1.23 |

only when all the estimates have comparable qualities. To support this interpretation, we report in Table 9 the RD performance of HT, IV, OD, and AV with respect to the oracle. We observe that AV may not be the best method on one given sequence because of the rate overhead for $\delta_R$ estimation. For example, for the 'lovebird' sequence, the oracle would choose the HT method all the time; so does the AV method, but it needs a small bit rate overhead to take the decision.

However, the AV method is the only one that can adapt to the different characteristics of the sequences, and for this reason, it has the best average performance, and it never has catastrophic increases with respect to the oracle as it happens for all the other methods: the temporal interpolation can cost up to 21% more than the oracle, the inter-view interpolation can cost up to 70%, and even the OD fusion can have a loss of about 11%. The AV fusion needs only 6.8% additional bit rate with respect to the oracle in the worst case and about 2.5% in average. The consistency of the RD performance is one of the most valuable properties of the proposed technique with respect to the state of the art.

Finally, in Table 10, we give the RD results for AV versus FD [44] and H.264/INTRA. In H.264/AVC INTRA, all the frames are INTRA coded (high profile, CABAC), i.e., without exploiting temporal or inter-view correlations. We observe that our method can achieve an average bit rate reduction of 9.11% w.r.t. [44] and of 22.38% w.r.t. H.264/AVC INTRA.

## 6 Conclusions

In this paper, we propose several methods to improve the quality of the estimate of the Wyner-Ziv frames in multiview DVC. We use the HOMI algorithm for temporal interpolation, taking advantage of its possibility of detecting non-linear trajectories. For the inter-view

**Table 9 RD performance for HT, IV, OD and AV vs. the oracle**

| | **HT** | | **IV** | | **OD** | | **AV** | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta_R$**(%)** | $\Delta_{PSNR}$ **(dB)** | $\Delta_R$**(%)** | $\Delta_{PSNR}$ **(dB)** | $\Delta_R$**(%)** | $\Delta_{PSNR}$ **(dB)** | $\Delta_R$**(%)** | $\Delta_{PSNR}$ **(dB)** |
| Balloons | 1.30 | -0.08 | 65.42 | -4.12 | 10.96 | -0.63 | 1.81 | -0.11 |
| Book arrival | 9.36 | -0.56 | 12.70 | -0.76 | 0.02 | 0.00 | 0.72 | -0.04 |
| Door flowers | 3.30 | -0.14 | 28.89 | -1.32 | 0.36 | -0.02 | 1.24 | -0.06 |
| Leaving laptop | 10.98 | -0.51 | 17.44 | -0.80 | 1.01 | -0.05 | 1.47 | -0.07 |
| Kendo | 5.83 | -0.33 | 24.59 | -1.40 | 5.13 | -0.25 | 5.73 | -0.28 |
| Lovebird | 0.00 | 0.00 | 48.28 | -2.30 | 10.56 | -0.52 | 0.64 | -0.03 |
| Newspaper | 0.18 | -0.01 | 70.64 | -0.35 | 8.21 | -0.43 | 1.65 | -0.05 |
| Pantomime | 5.15 | -0.31 | 12.24 | -0.72 | 6.28 | -0.37 | 6.83 | -0.41 |
| Outdoor | 21.21 | -1.38 | 1.73 | -0.11 | 8.96 | -0.57 | 2.34 | -0.15 |
| *Mean* | 6.36 | −0.37 | 31.32 | −1.32 | 5.72 | −0.31 | 2.49 | −0.13 |

Average RD performance in Bjontegaard metric for HT, IV, OD, and AV vs. the oracle.

interpolation, we introduce a method that allows a better discrimination between background and foreground. Finally, as far as the fusion method is concerned, we introduce two innovations: an occlusion detection-based fusion and an adaptive fusion, capable of restoring the inter-view or temporal interpolations instead of the fusion when needed. The originality of the latter method resides in the fact that we use a MAP optimal decision based on the observation of the bit rate requested by the channel decoder. The resulting performance is significantly better than state-of-the-art SI generation and fusion techniques: we obtain up to 7.7% of bit rate reduction for inter-view interpolation. The occlusion detection-based fusion allows a rate reduction of 6.6%, while using adaptive validation, we achieve up to more than 20% rate reduction and obtain an average of more than 9%. As future work, we want to explore multiview video-plus-depth scenarios: depth information can improve the quality of the SI. Moreover, camera parameters and DIBR algorithm (using the depth maps) can help us in the generation on the SI along the view domain without communication between the cameras.

## Endnotes

[a] For simplicity, in the following, we always consider the case where every second frame is a KF, i.e., the GOP size is 2. All the algorithms can be easily extended to larger GOP sizes.

[b] The 'disparity vector' has of course a null vertical component. This vectorial notation allows a simple problem formulation.

**Competing interests**
The authors declare that they have no competing interests.

Received: 28 June 2013 Accepted: 18 September 2013
Published: 2 October 2013

**References**

1. D Slepian, JK Wolf, Noiseless coding of correlated information sources. IEEE Trans. Inform. Theory. **19**, 471–480 (1973)
2. A Wyner, J Ziv, The rate-distortion function for source coding with side information at the receiver. IEEE Trans. Inform. Theory. **22**, 1–11 (1976)
3. A Aaron, R Zhang, B Girod, Wyner-Ziv coding of motion video, in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers,* Pacific Grove, 3–6 Nov 2002 (IEEE, Piscataway, 2002), pp. 240–244
4. B Girod, A Aaron, S Rane, D D Rebollo-Monedero, Distributed video coding. Proc. IEEE. **93**, 71–83 (2005)
5. R Puri, A Majumdar, K Ramchandran, PRISM: a video coding paradigm with motion estimation at the decoder. IEEE Trans. Image Process. **16**(10), 2436–2448 (2007)
6. G Cheung, A Ortega, N Cheung, Interactive streaming of stored multiview video using redundant framestructures. IEEE Trans. Image Process. **20**(3), 744 –761 (2011)
7. G Petrazzuoli, M Cagnazzo, F Dufaux, B Pesquet-Popescu, Using distributed source coding and depth image based rendering to improve interactive multiview video access, in *18th IEEE International Conference on Image Processing (ICIP),* Brussels, 11–14, Sept 2011 (IEEE, Piscataway, 2011), pp. 605–608
8. A Gelman, P Dragotti, V Velisavljevic, Interactive multiview image coding, in *18th IEEE International Conference on Image Processing (ICIP),* Brussels, 11–14, Sept 2011 (IEEE, Piscataway, 2011), pp. 601–604
9. T Maugey, J Gauthier, M Cagnazzo, B Pesquet-Popescu, Evaluation of side information effectiveness in distributed video coding, in *IEEE Transactions on Circuits and Systems for Video Technology* (IEEE, Piscataway, 2013 in press)
10. X Artigas, J Ascenso, M Dalai, S Klomp, D Kubasov, M Ouaret, The DISCOVER codec: architecture, techniques and evaluation, in *Picture Coding Symposium (PCS'07),* (Lisbon, November 2007)
11. J Ascenso, C Brites, F Dufaux, A Fernando, T Ebrahimi, F Pereira, S Tubaro, The VISNET II DVC codec: architecture, tools and performance, in *European Signal Processing Conference (EUSIPCO 2010)*, (Aalborg, 23–27 August 2010)
12. X Huang, S Forchhammer, Improved side information generation for distributed video coding. IEEE 10th Workshop on Multimedia Signal Processing, Cairns, 8–10 Oct 2008 (IEEE, Piscataway, 2008), pp. 223–228
13. T Andre, B Pesquet-Popescu, M Gastaud, M Antonini, M Barlaud, Motion estimation using chrominance for wavelet-based video coding, in *Proceedings of IEEE Picture Coding Symposium*, (San Francisco, December 2004)
14. B Macchiavello, F Brandi, E Peixoto, R de Queiroz, D Mukherjee, Side-information generation for temporally and spatially scalable Wyner-Ziv codecs. J. Image Video Process. **2009**, 14 (2009)
15. J Ascenso, F Pereira, Advanced side information creation techniques and framework for Wyner-Ziv video coding. J. Vis. Commun. Image Representation **19**(8), 600–613 (2008)
16. D Kubasov, C Guillemot, Mesh-based motion-compensated interpolation for side information extraction in distributed video coding. IEEE International Conference on Image Processing, Atlanta, 8–11 Oct 2006 (IEEE, Piscataway, 2006), pp. 261–264
17. S Mys, J Slowack, J Skorupa, P Lambert, RV de Walle, Introducing skip mode in distributed video coding. Signal Process Image Commun. **24**(3), 200–213 (2009)
18. JE Fowler, An implementation of PRISM using QccPack. Technical report, Mississippi State University (2005)
19. M Cagnazzo, T Maugey, B Pesquet-Popescu, A differential motion estimation method for image interpolation in distributed video coding. IEEE Intern. Conf. Acoust., Speech and Sign. Proc. **1**, 1861–1864 (2009)
20. M Cagnazzo, W Miled, T Maugey, B Pesquet-Popescu, Image interpolation with edge-preserving differential motion refinement. 16th IEEE International Conference on Image Processing (ICIP), Cairo, 7–10 Nov 2009 (IEEE, Piscataway, 2009), pp. 361–364
21. S Ye, M Ouaret, F Dufaux, T Ebrahimi, Improved side information generation with iterative decoding and frame interpolation for distributed video coding. 15th IEEE International Conference on Image Processing (ICIP 2008), San Diego, 12–15 Oct 2008 (IEEE, Piscataway, 2008), pp. 2228–2231
22. S Ye, M Ouaret, F Dufaux, T Ebrahimi, Improved side information generation for distributed video coding by exploiting spatial and temporal correlations. J. Image Video Process. **2009**, 4 (2009)
23. A Abou-Elailah, F Dufaux, J Farah, M Cagnazzo, B Pesquet-Popescu, Fusion of global and local motion estimation for distributed video coding. Circuits Syst. Video Technol, IEEE Trans. **22**, 1–12 (2012)
24. A Abou-Elailah, F Dufaux, J Farah, M Cagnazzo, Fusion of global and local side information using Support Vector Machine in transform-domain DVC. Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, 27–31 Aug 2012 (IEEE, Piscataway, 2012), pp. 574–578
25. R Martins, C Brites, J Ascenso, F Pereira, Refining side information for improved transform domain Wyner-Ziv video coding, Circuits Syst. Video Technol. IEEE Trans. **19**(9), 1327–1341 (2009)
26. A Abou-Elailah, F Dufaux, M Cagnazzo, B Pesquet-Popescu, J Farah, Successive refinement of side information using adaptive search area for long duration GOPs in distributed video coding. 19th International Conference on Telecommunications (ICT), Jounieh , 23–25 April 2012 (IEEE, Piscataway, 2012), pp. 1–6
27. M Tagliasacchi, S Tubaro, Hash-based motion modeling in Wyner-Ziv video coding. Proceed. of IEEE. Intern. Conf. Acoust., Speech and Sign. Proc. **1**, 509–512 (2007)

28. S Park, YY Lee, CS Kim, SU Lee, Efficient side information generation using assistant pixels for distributed video coding. Picture Coding Symposium (PCS), Krakow, 7–9 May 2012 (IEEE, Piscataway, 2012), pp. 161–164
29. F Verbist, N Deligiannis, M Jacobs, J Barbarien, A Munteanu, P Schelkens, J Cornelis, Maximum likelihood estimation for the generation of side information in distributed video coding. 18th International Conference on Systems, Signals and Image Processing (IWSSIP), Sarajevo, 16–18 June 2011 (IEEE, Piscataway, 2011), pp. 1–4
30. G Petrazzuoli, M Cagnazzo, B Pesquet-Popescu, High order motion interpolation for side information improvement in DVC. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, 14–19 March 2010 (IEEE, Piscataway, 2010), pp. 2342–2345
31. G Petrazzuoli, M Cagnazzo, B Pesquet-Popescu, Fast and efficient side information generation in distributed video coding by using dense motion rep, in *18th European Signal Processing Conference (EUSIPCO)*, Aalborg, 23–27 August 2010
32. G Petrazzuoli, T Maugey, M Cagnazzo, B Pesquet-Popescu, Side information refinement for long duration GOPs in DVC. IEEE International Workshop on Multimedia Signal Processing (MMSP), Saint Malo, 4–6 Oct 2010 (IEEE, Piscataway, 2010), pp. 309–314
33. C Brites, J Ascenso, F Pereira, Side information creation for efficient Wyner–Ziv video coding: classifying and reviewing. Signal Process, Image Commun. **28**(7), 689–726 (2013)
34. T Maugey, B Pesquet-Popescu, Side information estimation new symmetric schemes for multi-view distributed video coding. J. Vis, Commun. Image Representation. **19**(8), 589–599 (2008)
35. H Shum, S Kang, A review of image-based rendering techniques. Proceed. Intern. Symp. Visual Comm and Proc. (2000). doi: 10.1117/12.386541
36. 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 Process. Mag. **24**(5), 67–76 (2007)
37. X Guo, Y Lu, F Wu, W Gao, Distributed multi-view video coding. Proceed. of Intern Symp. Visual Comm. and Image Proc. **6077**, 290–297 (2006)
38. X Artigas, E Angeli, L Torres, Side information generation for multiview distributed video coding using a fusion approach. Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG ), Rejkjavik, 7–9 June 2006 (IEEE, Piscataway, 2006), pp. 250–253
39. M Ouaret, F Dufaux, T Ebrahimi, Fusion-based multiview distributed video coding. Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, Santa Barbara, 23–27 Oct 2006 (ACM, New York, 2006), pp. 139–144
40. J Areia, J Ascenso, C Brites, F Pereira, Wyner-Ziv stereo video coding using a side information fusion approach. IEEE 9th Workshop on Multimedia Signal Processing (MMSP), Crete, 1–3 Oct 2007 (IEEE, Piscataway, 2007), pp. 453–456
41. M Ouaret, F Dufaux, T Ebrahimi, Multiview distributed video coding with encoder driven fusion. in European Signal Processing Conference (EUSIPCO 2007), Poznan, 03–07 Sept 2007 (InfoScience, Cary, 2007), pp. 3–7
42. P Ferre, D Agrafiotis, D Bull, Fusion methods for side information generation in multi-view distributed video coding systems. Proceed. of IEEE Intern, Conf. Image Proc. **6**, 409–412 (2007)
43. M Tagliasacchi, G Prandi, S Tubaro, Symmetric distributed coding of stereo video sequences. Proceed. of IEEE Intern. Conf. Image Proc. **2**, 29–32 (2007)
44. T Maugey, W Miled, M Cagnazzo, B Pesquet-Popescu, Fusion schemes for multiview distributed video coding, in *European Signal Processing Conference (EUSIPCO)*, Glasgow, August 2009
45. Y Li, H Liu, X Liu, S Ma, D Zhao, W Gao, Multi-hypothesis based multi-view distributed video coding, in *Picture Coding Symposium (PCS)*, Chicago, 6–8 May 2009 (IEEE Piscataway, 2009), pp. 1–4
46. D Kubasov, J Nayak, C Guillemot, Optimal reconstruction in Wyner-Ziv video coding with multiple side information. IEEE 9th Workshop on Multimedia Signal Processing (MMSP), Crete, 1–3 Oct. 2007 (IEEE, Piscataway, 2007), pp. 183–186
47. M Salmistraro, M Zamarin, LL Rakêt, S Forchhammer, D Fotonik, Ø Plads, Distributed multi-hypothesis coding of depth maps using texture motion information and optical flow, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)* (Vancouver Convention Exhibition Centre, Vancouver, 26–31 May 2013)
48. X Huang, L Raket, M Luong, HV Nielsen, F Lauze, S Forchhammer, Multi-hypothesis transform domain Wyner-Ziv video coding including optical flow. IEEE 13th International Workshop on Multimedia Signal Processing (MMSP), Hangzhou, 17–19 Oct 2011 (IEEE, Piscataway, 2011), pp. 1–6
49. F Dufaux, Support vector machine based fusion for multi-view distributed video coding. 17th International Conference on Digital Signal Processing (DSP), Corfu, 6–8 July 2011 (IEEE, Piscataway, 2011), pp. 1–7
50. I Daribo, B Pesquet-Popescu, Depth-aided image inpainting for novel view synthesis. IEEE International Workshop on Multimedia Signal Processing (MMSP), Saint Malo, 4–6 Oct 2010 (IEEE, Piscataway, 2010), pp. 167–170
51. W Miled, J Pesquet, Disparity map estimation using a total variation bound. The 3rd Canadian Conference on Computer and Robot Vision, Quebec, 7–9 June 2006 (IEEE, Piscataway, 2006), p. 48
52. P Fua, A parallel stereo algorithm that produces dense depth maps and preserves image features. Mach. Vis. Appl. **6**, 35–49 (1993). doi:10.1007/BF01212430
53. G Egnal, R Wildes, Detecting binocular half-occlusions: empirical comparisons of five approaches. IEEE Trans. Pattern Anal Mach. Intell. **24**(8), 1127–1133 (2002)
54. D Scharstein. R Szeliski, High-accuracy stereo depth maps using structured light. in IEEE Conference on Computer Vision and Pattern Recognition. **1**, 195–202 (2003)
55. H Hirschmüller, Improvements in real-time correlation-based stereo vision. Proceedings of the IEEEWorkshop on Stereo and Multi-Baseline Vision (SMBV 2001), Kauai, 9–10 Dec 2001 (IEEE, Piscataway, 2001), pp. 141–148
56. KJ Yoon, IS Kweon, Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal, Mach. Intell. **28**(4), 650–656 (2006)
57. DN Rowitch, LB Milstein, On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes. Commun. IEEE Trans. **48**(6), 948–959 (2000)
58. G Bjontegaard, Calculation of average PSNR differences between RD-curves, in *VCEG Meeting*, (Austin, 2–4 April 2001)
59. E Parzen, On estimation of a probability density function and mode. Ann. Math. Stat. **33**(3), 1065–1076 (1962)

doi:10.1186/1687-6180-2013-154
**Cite this article as:** Petrazzuoli *et al.*: Novel solutions for side information generation and fusion in multiview DVC. *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:154.