A robust fusion method for multiview distributed video coding.

Salmistraro, Matteo; Ascenso, Joao; Brites, Catarina; Forchhammer, Søren

Published in:
Eurasip Journal on Advances in Signal Processing

Link to article, DOI:
10.1186/1687-6180-2014-174

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

A robust fusion method for multiview distributed video coding

Matteo Salmistraro[1], João Ascenso[2], Catarina Brites[2] and Søren Forchhammer[1*]

Abstract

Distributed video coding (DVC) is a coding paradigm which exploits the redundancy of the source (video) at the decoder side, as opposed to predictive coding, where the encoder leverages the redundancy. To exploit the correlation between views, multiview predictive video codecs require the encoder to have the various views available simultaneously. However, in multiview DVC (M-DVC), the decoder can still exploit the redundancy between views, avoiding the need for inter-camera communication. The key element of every DVC decoder is the side information (SI), which can be generated by leveraging intra-view or inter-view redundancy for multiview video data. In this paper, a novel learning-based fusion technique is proposed, which is able to robustly fuse an inter-view SI and an intra-view (temporal) SI. An inter-view SI generation method capable of identifying occluded areas is proposed and is coupled with a robust fusion system able to improve the quality of the fused SI along the decoding process through a learning process using already decoded data. We shall here take the approach to fuse the estimated distributions of the SIs as opposed to a conventional fusion algorithm based on the fusion of pixel values. The proposed solution is able to achieve gains up to 0.9 dB in Bjøntegaard difference when compared with the best-performing (in a RD sense) single SI DVC decoder, chosen as the best of an inter-view and a temporal SI-based decoder one.

Keywords: Distributed video coding; Multiview video coding; Side information fusion; Learning

1 Introduction

Distributed video coding (DVC) [1-3] is a coding paradigm based on the theoretical results of distributed source coding (DSC): the Slepian-Wolf [4] and the Wyner-Ziv (WZ) theorems [5]. These foundations establish a different way to compress information, namely, by independently coding the source data but jointly decoding it. Thus, in DVC, the source correlation is exploited at the decoder, as opposed to the widely adopted predictive coding solutions where the encoder is responsible for exploiting all the correlation. One of the key blocks of every DVC decoder is the side information (SI) generation module which estimates the WZ frame to be decoded. Typically, in monoview systems, the SI creation exploits the temporal redundancy by making assumptions of the apparent motion in a video stream, e.g. linear motion between reference frames is assumed [6]. Then, at the encoder, parity bits (or syndromes) are generated and transmitted to the decoder, and the use of channel decoders allows obtaining the decoded frames given the available SI. The channel decoder requires soft inputs for the source data to be decoded, which can be calculated from a correlation noise model. This correlation noise model statistically describes the relationship between the SI and the source and is obtained by computing an online residual, without using the original WZ frame.

An efficient DVC system must be able to minimize the amount of data sent from the encoder for a certain decoded quality level. Therefore, the SI has high importance for the rate-distortion (RD) performance of the DVC decoder; in fact, having a high-quality SI, characterized by few errors, allows the transmission of less error correcting data (requiring a lower bitrate) and enables improving the decoded WZ frame quality.

In monoview DVC codecs, every frame is independently coded without any reference to other decoded frames. This allows a low encoding complexity since the complex task of exploiting the temporal correlation (using motion estimation/compensation) is performed at the decoder.

* Correspondence: sofo@fotonik.dtu.dk
[1]DTU Fotonik, Technical University of Denmark, Ørsteds Plads, 2800 Kongens Lyngby, Denmark
Full list of author information is available at the end of the article

When different views of the same visual scene are coded in different camera nodes, e.g. in visual sensors networks, inter-view coding can further improve the coding performance, exploiting inter-camera redundancy. If a predictive multiview video codec is used, e.g. multiview video coding (MVC) [7], inter-camera communication is needed. MVC relies on the same coding tools used in H.264/AVC: decoded frames belonging to other views are inserted in the reference picture lists and used for disparity estimation/compensation. This approach requires inter-camera communication to enable one camera to use the frames of another camera for disparity compensation.

On the other hand, in DVC solutions for the multiview scenario, each camera can independently code the frames, relying on the decoder to exploit the correlation between the views [8,9]. Typically, the multiview DVC (M-DVC) decoder tries to exploit, at the same time, temporal intra-view and inter-view correlation, generating two SI frames: (1) temporal SI, by means of motion estimation and interpolation, e.g. employing overlapped block motion compensation (OBMC) [6] and (2) inter-view SI, generated by leveraging the inter-view redundancy [3]. To exploit the best part of each estimated SI frame, it is necessary to fuse the frames, choosing the best regions of each estimated SI frame to create a final SI frame that is used for decoding [8,9]; typically, the regions are chosen according to an estimation of their quality. SI fusion is a hard problem, and there are many fusion techniques available in the literature [8] with various degrees of efficiency. The goal of an efficient frame fusion technique is to deliver an RD performance better than the best-performing single SI decoder out of the one using the inter-view SI and the one using the temporal SI. In general, the larger the difference in RD performance between the SIs, the harder the fusion task is because fusing incorrectly a region of the frame may lead to consistent losses in RD performance.

Considering these challenges, the main contributions of this work are the following:

(1). A novel inter-view SI generation system called overlapped block disparity compensation (OBDC) is presented. This method is able to cope with high camera distance and detect occlusions due to a part of the scene outside the field of view of one camera. It is also able to adapt to unknown camera distances
(2). The fusion of the estimated distributions of the DCT coefficients of the SI
(3). A novel learning technique based on the refinement of the quality of the fused SI along the decoding process exploiting already decoded data

The three items are combined in a DVC set-up providing a novel learning-based M-DVC scheme. The fusion of distributions here is proposed as an alternative to the pixel-level fusion of the SI frames. The use of distributions to estimate the reliability of the regions of the SI allows exploiting high-performance noise modelling algorithms developed in literature. This learning algorithm allows correcting wrong initial estimations of the quality of the SIs, leading to superior RD performance for the next steps of the decoding process.

This paper is structured as follows: Section 2 deals with related works on inter-view SI creation and pixel- and block-based SI fusion techniques. An overview of the DVC coding process is given in Section 3. The novel fusion algorithm as well as the SI generation method is described in Section 4. In Section 5, the performance of the proposed tools is assessed and compared with state-of-the-art distributed coding solutions, as well as monoview predictive codecs.

2 Related work

2.1 Inter-view SI creation

Disparity compensation view prediction (DCVP) [10] is one of the simplest inter-view SI generation techniques, where the same algorithm used for temporal interpolation is applied between adjacent views to perform disparity estimation and compensation. However, the DCVP SI quality deteriorates when the distance between views is increased. The majority of the studies proposed in literature focus on really close cameras; for example, the distance between the cameras in [8] is 6.5 cm, and the problem of cameras moving with respect to each other is not addressed.

A different way to address the SI generation problem was proposed in [11], where multiview motion estimation (MVME) was presented. The key idea of MVME is to estimate a single SI frame by jointly exploiting the motion of neighbouring views and projecting the motion field in the current view. MVME generates the SI in two separate steps: (1) motion estimation is performed on the available lateral (left and right) views and (2) motion compensation using the reference (decoded) frames in the view to decode (the central view). A fusion step is performed in MVME to fuse various joint motion and disparity estimations, while in the previous work the fusion was performed between a purely inter-view SI and a purely temporal one. MVME demonstrates high performance in fast-motion sequences, but it is outperformed by motion compensation and interpolation techniques in slow-motion cases [11]. More recently [12], a modified version of the temporal motion estimation algorithm employed in DISCOVER [13] is proposed for inter-view SI generation. The key novelty is the penalization of small disparities, which characterizes background blocks.

2.2 SI fusion techniques

In recent years, SI fusion methods which use estimated distributions of the DCT coefficients were proposed for

monoview DVC [14,15] and applied to M-DVC [16,17]. In [14], optimal reconstruction for a multi-hypothesis decoder was proposed. In [16], the authors enhanced [14], proposing a cluster-based noise modelling system and fusion. In [15], the concept of parallel decoding was introduced: the distributions of the available SIs were fused using different weights, generating, in the aforementioned case, six different fused distributions. From each fused distribution, it is possible to calculate a set of conditional probabilities which are fed into six parallel LDPCA decoders. Thereafter, the decoders try to reconstruct the source bitplane considered in parallel for each new chunk of received parity bits. The process stops when the bitplane is successfully decoded by at least one LDPCA decoder. The method proposed in [15] can be seen as a brute-force rate-based optimization approach but it suffers from high computational complexity; to perform an efficient SI fusion, several channel decoders need to be used. In [17], the method proposed in [15] was applied to stereo M-DVC to fuse an inter-view and temporal SI frames. Nevertheless, the issue related to the complexity of [15] was not addressed, since [17] still relies on parallel LDPCA decoding.

In M-DVC, pixel- and block-based fusion techniques are widely adopted [8,9]. The results of [8] show that finding a fusion method able to perform robustly for a wide range of different video sequences is difficult, in particular, when the quality of the two SIs is very different and therefore the probability of making errors in the fusion process is high. A different approach for fusion in M-DVC is proposed in [9], where a past decoded WZ frame and its corresponding SI are used to train a support vector machine classifier, which is then used to perform the fusion task, classifying the reliability of each pixel in the SIs. In [12], the fusion is performed according to an occlusion map: temporal SI is used if pixels belonging to the left or right views are estimated to be occluded. In [12], adaptive validation is also introduced: for a small subset of the WZ frames, the parity bits are requested for correct inter-view and temporal SIs, introducing an overhead. If the two SIs require similar rates, the fused SI is chosen; otherwise, the single SI providing the lower rate is chosen.

However, the partially decoded information obtained during the decoding process can be used to enhance the RD performance of a DVC codec by improving the correlation noise [6,18] or the SI [19] or, as it is proposed in this work, the fusion process in a multiview decoder. In [20], the WZ frame is first decoded using either inter- or intra-view SI, according to the motion activity of the video. Then the completely reconstructed WZ frame is used as basis for the generation of a refined SI, either disparity or motion compensation is used on a block basis. Lastly, the refined SI is used in a new reconstruction step obtaining a higher quality reconstruction.

In [10], the encoder sends information to improve the fusion process: since the encoder has access to the original WZ frame and the key frames (KFs), a fusion mask can be generated based on the difference between the KFs and WZ frame (both known at the encoder). The mask is then compressed and sent to the decoder to drive the fusion process. However, when the encoder participates in the fusion process, its computational complexity is increased which may be impractical for some applications. In addition, the overhead can lead to a significant increase of the bitrate, which may severely limit the improvements obtained from having a higher quality fused SI frame. However, none of the works above used past decoded information to perform a better fusion process in a multiview decoder, as proposed in this work and described next.

2.3 Benchmarks for SI fusion

In [8,9], many SI fusion solutions were reviewed and presented. However, it is worth describing one method often used for comparison, MDCD-Lin [8] and two (ideal) SI fusion solutions often used as benchmark in the MV-DVC literature. In addition, these benchmarks are used to assess the proposed technique in Section 5.

Consider that the original WZ frame is denoted as X. The SIs employed for fusion, in all the benchmarks, are generated through OBMC and OBDC and denoted as Y_{OBMC} and Y_{OBDC} respectively. The corresponding estimated residuals are denoted as R_{OBMC} and R_{OBDC}. The following SI fusion benchmarks were considered.

Motion and disparity compensated difference linear fusion (MDCD-Lin) is a multiview fusion technique [8] used as benchmark in [9,12]. The techniques presented in [9] are shown to perform either as well as MDCD-Lin or as well as the best single SI decoder. Therefore, MDCD-Lin and two single SI decoders are usually employed as benchmarks. The MDCD-Lin fuses pixel values, using the estimated residuals as weights for generating the fused SI, for the pixel having position $\boldsymbol{x}$. The weight is calculated as follows:

$$w(\boldsymbol{x}) = \frac{|R_{OBMC}(x)|}{|R_{OBMC}(x)| + |R_{OBDC}(x)|} \quad (1)$$

The final SI is calculated as follows:

$$Y(\boldsymbol{x}) = w(\boldsymbol{x})Y_{OBDC}(\boldsymbol{x}) + (1 - w(x))Y_{OBMC}(\boldsymbol{x}) \quad (2)$$

The residual for the final SI is calculated using the same weighted average for the residuals.

Ideal fusion (IF) is also considered [8,9], which is sometimes referred to as oracle fusion. This is a quite common bound in M-DVC literature. It is often used as an upper bound to the performance a fusion technique can achieve. The fused SI is calculated as follows:

$$Y(\boldsymbol{x}) = \begin{cases} Y_{OBDC}(\boldsymbol{x}) & \text{if } |X(\boldsymbol{x}) - Y_{OBMC}(\boldsymbol{x})| > |X(\boldsymbol{x}) - Y_{OBDC}(\boldsymbol{x})| \\ Y_{OBMC}(\boldsymbol{x}) & \text{otherwise,} \end{cases} \quad (3)$$

and the same rule is applied to the residuals, in order to fuse them, obtaining the final residual. The technique requires that the original WZ frame, X, is known at the decoder, and therefore, the technique is not applicable in a practical scenario, but it may be used as a bound for the performance of the system. Even though IF is often used as upper bound (e.g. [9]), it is not an upper bound in a strict sense, since it performs a distortion-based optimization on the quality of the SI, and an improved PSNR of the SI need not always lead to superior RD performance.

Block-based (BB) ideal fusion (IF), (IF BB), is also introduced here. Given a block B, of 4×4 pixels, corresponding to a DCT block, the SAD (sum of absolute differences) of the block between the SI and the corresponding block in the original WZ frame is calculated and used as reliability measure to calculate the weight:

$$w_B = \frac{\Sigma_{\boldsymbol{r} \in B} |X(\boldsymbol{r}) - Y_{OBMC}(\boldsymbol{r})|}{\Sigma_{\boldsymbol{r} \in B} |X(\boldsymbol{r}) - Y_{OBMC}(\boldsymbol{r})| + \Sigma_{\boldsymbol{r} \in B} |X(\boldsymbol{r}) - Y_{OBDC}(\boldsymbol{r})|} \quad (4)$$

The weight w_B is then used to fuse each pixel $\boldsymbol{r}$ belonging to B as in (2) as well as it is used to generate the residual of the fused SI. Since IF BB requires the knowledge of the original WZ frame, X, this technique cannot be employed in a realistic scenario (as for IF), but it is a useful bound for what concerns the performance which can be reached using the learning approach presented in the next section.

3 Proposed M-DVC codec architecture

The M-DVC solution proposed in this paper adopts the widely used three-view scenario, although it may be generalized to other scenarios with more cameras. In this scenario, all the views are independently encoded without exploiting any inter-view correlation. However, the central view is decoded exploiting the inter-view correlation, while the left and right views are also independently decoded with respect to the other views and used to generate the SI for the current view. At the decoder, the M-DVC solution has access to the decoded frames from the lateral and central views, as shown in Figure 1. To generate the SI, OBMC only needs to access the decoded frames $I_{c,t-1}$ and $I_{c,t+1}$ since only the temporal correlation is exploited and OBDC requires also the decoded frames $I_{r,t}$ and $I_{l,t}$ since the disparity correlation is exploited, and X is the WZ frame of the central view, unknown at the decoder. The *central view* is WZ encoded; the *lateral views* (left and right views) are H.264/AVC Intra coded. The architecture of the proposed DVC codec is depicted in Figure 2 for the encoder and Figure 3 for the central view decoder (in Figure 3, the proposed tools are shaded). The overall encoding process for the multiview DVC encoder can be described as follows:

Central view encoder (Figure 2)

1. First, the *Video Splitting* module classifies the video frames into WZ frames and key frames according to the group-of-pictures (GOP) structure. In a GOP, the first frame is a KF, the others are WZ frames. The frames selected as KFs are encoded by a *H.264/AVC Intra encoder* and sent to the decoder.
2. For the WZ frames X, a DCT *transform* is applied, in this case an integer, 4×4 DCT. The DCT coefficients are uniformly quantized (according to the selected RD point) and divided into bitplanes by the *Quantization* module.
3. Each bitplane is fed as input to an *LDPCA encoder* [21], which generates syndromes which are stored in a *buffer* and sent upon request from the decoder.

Lateral view encoders (Figure 2)

In general, the only multiview codec requirement is that the lateral views (Figure 1) are encoded independently, i.e. without exploiting any past decoded frames of the same view or from the central view. In this setup, the lateral view frames ($\tilde{I}_l$, $\tilde{I}_r$) are coded with the H.264/AVC Intra Encoder but other solutions could be used, e.g. monoview DVC codec.

The overall decoding process for the multiview DVC decoder can be described as follows:

Lateral view decoders

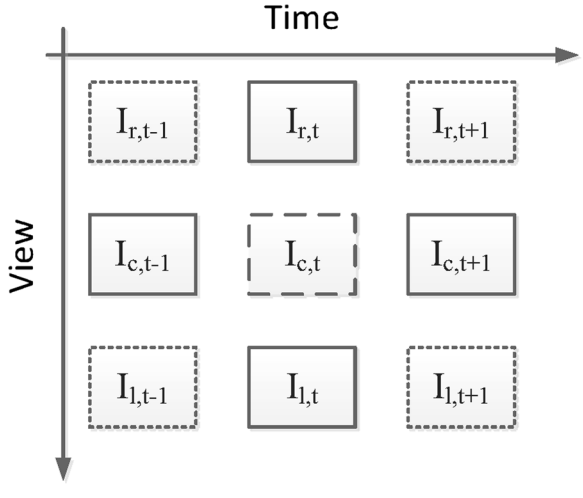

Figure 1 Stream structure. The frames in solid are KFs in which the decoder has access to.

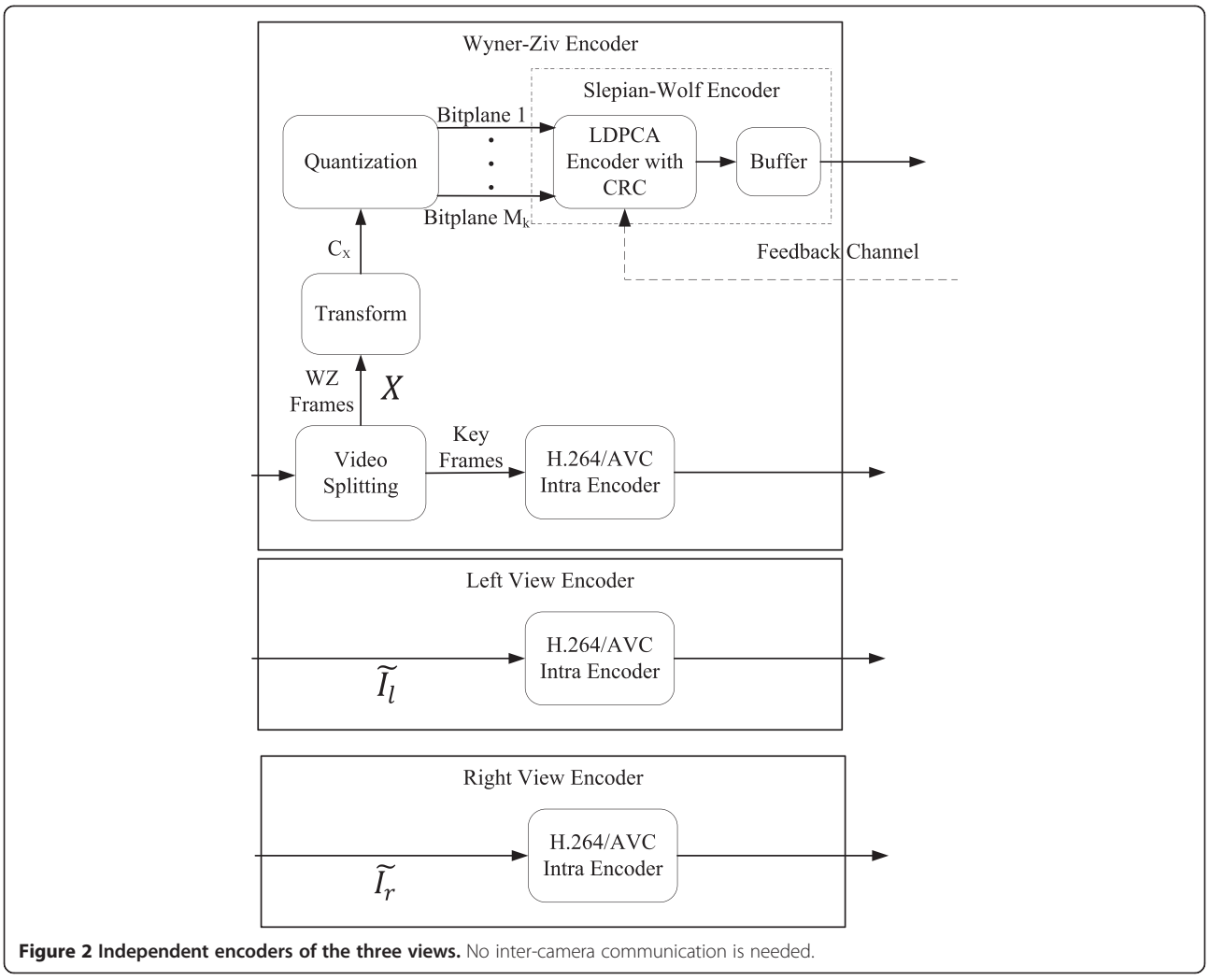

Figure 2 Independent encoders of the three views. No inter-camera communication is needed.

In this case, the lateral view frames are H.264/AVC Intra decoded but, as previously stated, other solutions could be used, e.g. monoview DVC codec. The left and right reconstructed frames are denoted as I_l and I_r, respectively.

Central view decoder (Figure 3)

1. The KFs are decoded first, using an H.264/AVC decoder, obtaining $I_{c,t-1}$ and $I_{c,t+1}$. In addition, the key frame quality should match the quality of the reconstructed WZ frame on average. Thus, to avoid quality fluctuations appropriate quantization step sizes for the WZ and KF DCT coefficients must be selected.
2. Then, $I_{c,t-1}$ and $I_{c,t+1}$ are used by the *OBMC SI generation* module to calculate the SI Y_{OBMC} and the (online) residual R_{OBMC}. Thereafter, Y_{OBMC} and R_{OBMC} are DCT transformed, and two sets of DCT coefficients C_{OBMC} and $C_{R,OBMC}$ are obtained. In this work, online residual estimation, as detailed in [6], is employed to estimate the relationship between the original WZ and SI frames without requiring access to the original WZ frame. The residual DCT coefficients $C_{R,OBMC}$ are used by the *Noise Modelling* module to calculate the parameter α_{OBMC} of the laplacian distribution of the correlation noise model [6].
3. The *OBDC SI generation* module calculates Y_{OBDC} and the corresponding residual R_{OBDC}. In OBDC, pre-aligned frames $I_{l,t}^{(a)}$ and $I_{r,t}^{(a)}$ are generated from the left-view $I_{l,t}$ and right-view $I_{r,t}$, respectively, removing lateral regions where no correspondence exists between frames. These regions cannot be interpolated using disparity compensation and thus, the co-located pixels in Y_{OBMC} are used. The SI frame and residual are both DCT transformed, generating C_{OBDC} and $C_{R,OBDC}$, respectively. Again, $C_{R,OBDC}$ is used by the Noise Modelling module to

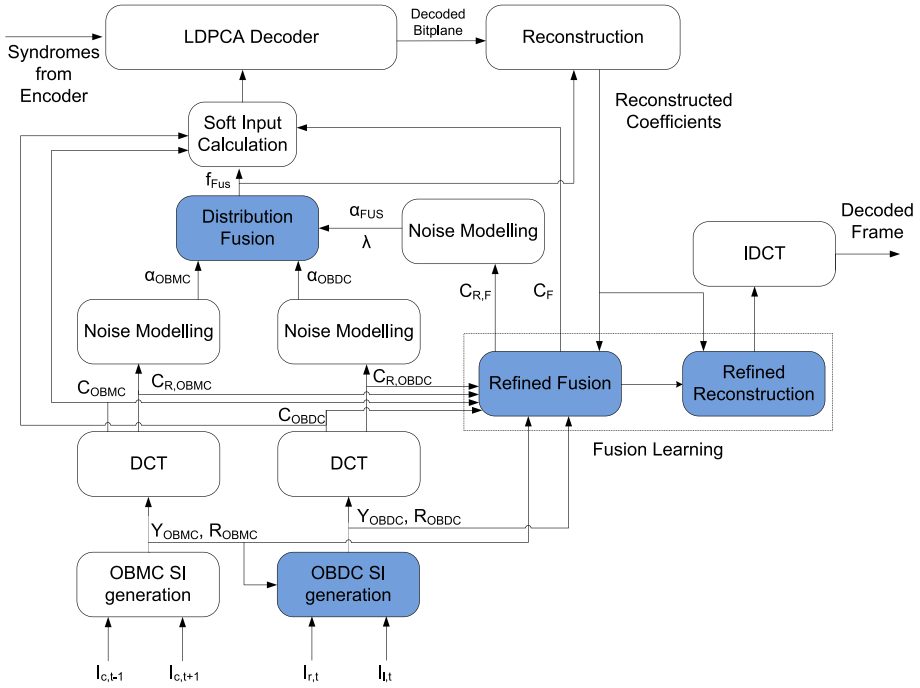

Figure 3 Architecture of the proposed central view decoder.

calculate the parameter α_{OBDC} of the laplacian distribution of the correlation noise model [6].

4. The *Refined Fusion* module generates the fused SI coefficients $C_F^{b_k}$ for DCT band b_k. The calculation of the corresponding residual coefficient $C_{R,F}^{b_k}$ after fusion is also performed. Both sets of coefficients (SI and residual) are calculated as weighted averages of the corresponding coefficients (or residuals) of OBMC and OBDC. The weights are calculated using the mean absolute differences (MAD) distortion metric between the partially decoded WZ frames and the SI frames; see Section 4.3 for more details.
5. The *Distribution Fusion* module calculates the joint distribution $f_{Fus}^{b_k}$ from the three correlation noise models: OBMC, OBDC and the fused SI. Then, the joint distribution is used by the *Soft Input Calculation* module to calculate the conditional probabilities for the LDPCA decoder. The joint distribution allows the systems to effectively fuse the three different SIs, taking into account the previously decoded information.
6. The *LDPCA decoder* requests syndromes from the encoder using a feedback channel: initially, a subset of syndromes is received by the decoder, which attempts to decode the source (bitplane). If the LDPCA decoding succeeds and an 8-bit CRC does not detect any error, the bitplane is assumed to be decoded, otherwise new syndromes are requested via the feedback channel, until successful decoding is achieved.
7. Once all the bitplanes of the band b_k are decoded, the DCT band is reconstructed by the *Reconstruction* module, using $f_{Fus}^{b_k}$, employing the optimal reconstruction technique outlined in [14].
8. At last, when all the bands are successfully decoded, the OBMC and OBDC are fused again. The newly fused SI is used in a last reconstruction step in the *Refined Reconstruction* module to further improve the quality of the decoded WZ frame.

4 Multiview decoding tools

In this section, the proposed techniques are described and analyzed. Thus, the novel contributions are inter-view OBDC SI generation, distribution fusion and the *Fusion Learning*, which can be divided into two distinct elements: the *Refined Fusion* used during the decoding process and the refined reconstruction used at the end of the decoding process (Figure 3).

4.1 Inter-view side-information generation

When using DCVP for inter-view SI generation, the same algorithm applied for motion interpolation is

applied between lateral views. This generates errors; for example, the appearance and disappearance of objects from the scene can create areas of wrong matches because an object in one view may have few or no matches in the other view. Thus, wrong disparity vectors can be estimated which in turn may lead to erroneous predictions. Typically, when content is acquired in a multiview system, there are regions which are present in one view but are occluded in another view, since the objects of the scene could be partially or totally occluded from the field-of-view of one camera when compared to another camera. This occurs quite often in the lateral areas of the frames. On the other hand, there are regions where there are clear correspondences between two views. In addition, when disparity between views is high, a higher search range is needed to have correct correspondences between views. This may lead to wrong matches in lowly textured areas. A way to mitigate these two aforementioned problems is to remove the lateral areas from the two frames by aligning them. Naturally, disparity estimation and compensation still needs to be performed, as each object has its own disparity due to the distance of the object to the cameras of the multiview system.

4.1.1 Overlapped block disparity compensation

As stated in the previous section, OBDC is conceptually similar to the idea of DCVP; but to allow for larger disparities, $I_{r,t}$ and $I_{l,t}$ shall be pre-aligned. This is done by finding the minimum average disparity and removing unmatched areas as described below. Consider that each frame of the multiview system has $n \times m$ spatial resolution. The average disparity d_{avg} between two views is calculated by the following:

$$d_{avg} = \underset{q\in[-r,r]}{\operatorname{argmin}} \sum_{i=0+q\chi(q)}^{m+q(\chi(-q))-1} \sum_{j=0}^{n-1} \frac{\left|I_{l,t}(i,j) - I_{r,t}(i-q,j)\right|}{(m-|q|)n} \tag{5}$$

where $\chi(q)$ is an indicator function, with $\chi(q) = 1$ if $q \geq 0$, and $\chi(q) = 0$ otherwise. r is the positive bound of the search range. If $d_{avg} > 0$, the pixels belonging to the area having i coordinates in the interval $[0, |d_{avg}| - 1]$ are removed from $I_{l,t}(i,j)$ frame, generating $I_{l,t}^{(a)}$, and for $I_{r,t}$ the pixels in the area $[m - 1 - |d_{avg}|, m - 1]$ are removed. In case $d_{avg} < 0$, the roles of the two frames are inverted as can be seen from the interval covered by the i variable in the first sum for a negative q.

The pixels contained in the lateral areas cannot be used for the disparity estimation and interpolation, since they have no match in the other area; therefore, these two areas are removed, generating the aligned frames $I_{l,t}^{(a)}$ and $I_{r,t}^{(a)}$, to which OBMC is applied, generating $Y_{OBDC}^{(a)}$. However, in $Y_{OBDC}^{(a)}$ there are now two areas, $|d_{avg}|/2$ pixels wide, which cannot be interpolated since their corresponding pixels are visible only in one KF view. The assumption for the structure of the areas in $Y_{OBDC}^{(a)}$ comes from the symmetrical structure of the placement of the cameras. Therefore, the unmatched pixels are substituted with the co-located pixels in Y_{OBMC}. A schematic of the algorithm is depicted in Figure 4. The same substitution is applied to the residual of OBDC, since it suffers from the same problem.

Using the pre-alignment phase, the length of the disparity vectors is reduced. This allows using a smaller search range, more reliable estimation (fewer wrong matches) and also lowering computational complexity. In addition, the calculation of the disparity field in the unmatched areas is not performed, allowing more robust motion estimation for the other blocks. In OBMC (which is the core of OBDC, see Figure 4) and in many similar motion estimation algorithms, smoothing is done on the motion field after its initial calculation. Erroneous disparity vectors may influence correct ones; therefore, with the alignment, the propagation of the error is avoided.

4.2 Fusion based on weighted distribution

The techniques previously proposed in literature make use of the residual or similar features to estimate the reliability of a given pixel (or block) for the two SI estimations. Once the SI reliability is estimated locally, it is possible to fuse each estimate, combining the SI estimates to achieve a higher reliability. Traditionally, many fusion methods for DVC use a binary mask which indicates how the two SI estimations should be fused to maximize the final SI frame quality. However, using this approach a hard decision is made which could be far from optimal and the generation of a new correlation noise model for the fused SI frame is difficult. Here, a different approach is proposed by fusing the correlation noise model distributions obtained for the two SI estimations independently, thus avoiding the need to calculate a residual for the fused SI. The better the residual and correlation noise model estimation is, the better the fusion process works. In addition, fusing the distributions according to the correlation model can be improved, as better correlation noise models are proposed in the literature. First, the correlation noise modelling presented in [6] is summarized here for completeness. Defining $C_R^{b_k}$ as the DCT transform of the estimated residual for band b_k, $D(u,v)$ measures the distance between individual

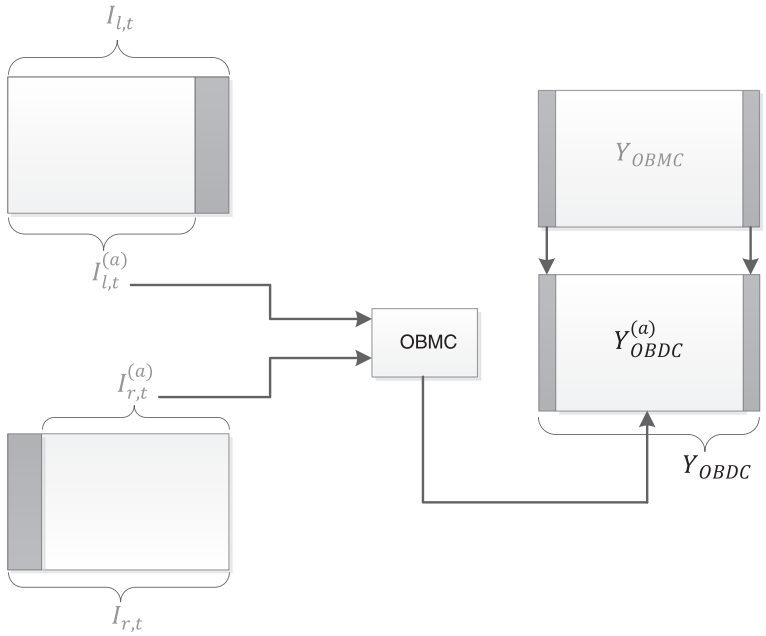

Figure 4 Illustration of the OBDC SI generation module.

coefficients and the average value of coefficients within band b_k:

$$D(u,v) = \left|C_R^{b_k}(u,v)\right| - E\left[\left|C_R^{b_k}\right|\right] \tag{6}$$

The parameter $\alpha^{b_k}(u,v)$ of the laplacian distribution used in the noise modelling is calculated as in [6]:

$$\alpha^{b_k}(u,v) = \frac{\alpha_c^{b_k}\beta E\left[\left|C_R^{b_k}\right|\right]}{\beta E\left[\left|C_R^{b_k}\right|\right] + (1-\beta)D(u,v)} \tag{7}$$

where $E[\cdot]$ denotes the expectation. The possible values of β are described in [6]. $\alpha_c^{b_k}$ is calculated as follows and it is based on the cluster c (inliers or outliers) the position (u,v) belongs to

$$\alpha_c^{b_k} = \frac{N_c}{\Sigma_{(u,v)\in c}\left|\left|C_R^{b_k}(u,v)\right| - E\left[\left|C_R^{b_k}\right|\left|(u,v)\in c\right.\right]\right|} \tag{8}$$

where N_c is the number of positions belonging to cluster c.

To determine which cluster the coefficient $C_R^{b_k}(u,v)$ belongs to, a mapping function is used based on the classification (inliers or outliers) on the already decoded coefficients [6]. This classification is based on the estimated variance of the coefficient and $D(u,v)$ [6]. Once the already decoded coefficients are classified, the classification of the coefficients of band b_k is estimated by the mapping function as in [6]. The algorithm employed is more complex [6], but here the main elements necessary to understand the rest of the work are provided.

Using the procedure outlined above for the generic laplacian parameter $\alpha^{b_k}(u,v)$, two sets of laplacian parameters can be defined: one set for the OBMC SI and one set for the OBDC SI, $\alpha_{OBMC}^{b_k}(u,v)$ and $\alpha_{OBDC}^{b_k}(u,v)$, respectively. The weight for fusing the distribution is calculated as proposed in [16]:

$$w^{b_k}(u,v) = \frac{\left(\alpha_{OBMC}^{b_k}(u,v)\right)^2}{\left(\alpha_{OBMC}^{b_k}(u,v)\right)^2 + \left(\alpha_{OBDC}^{b_k}(u,v)\right)^2} \tag{9}$$

Once the weights are calculated, the joint distribution for each position is defined as follows:

$$f^{b_k,(u,v)} = w^{b_k}(u,v)f_{X|Y_{OBMC}}^{b_k,(u,v)} + \left(1 - w^{b_k}(u,v)\right)f_{X|Y_{OBDC}}^{b_k,(u,v)} \tag{10}$$

where $f_{X|Y}^{b_k,(u,v)}$ is the estimated distribution for the coefficient (u,v) in band b_k given Y. The idea is that the weights give an indication of the reliability of the SIs and therefore they are used to fuse the distributions. This may be applied both in pixel-based and block-based approaches. This system is compatible with and

exploits the efficient block-based correlation noise estimations available in literature.

4.3 Fusion learning

The SI fusion process described in the previous section can be improved using a learning-based approach to leverage the knowledge of the already decoded bands. The idea is to use the already decoded bands to perform a more reliable SI fusion. Assuming that band b_k, with $k > 0$, is being decoded (b_0 indicates the DC coefficient) and that the decoding follows a zig-zag scan order, the previously decoded bands b_l, $l < k$ can be used to guide the fusion for each SI DCT coefficient. Consider a 4×4 DCT block in Y_{OBMC}, denoted as B_{OBMC} and its corresponding block in the partially reconstructed frame B_{Rec}. Let $C_{OBMC}^{b_k}(u,v)$ denote the coefficient in band b_k having position (u,v). First, the non-decoded coefficients are forced to be zero in B_{OBMC} and in the partially reconstructed block B_{Rec}. Then, both DCT blocks are inverse DCT transformed and the MAD between the two blocks is calculated, and it is denoted as the weight $w_F^{OBMC}(u,v)$ as shown in Figure 5. The MAD is an indicator of how close the previous SI DCT coefficients were to the ones belonging to the original WZ frame. It has to be noted that the WZ frame is not used in this process. The same procedure can be repeated for OBDC, using B_{OBDC} and B_{Rec}, generating the weight $w_F^{OBDC}(u,v)$. The higher the weight, the lower the reliability of the corresponding SI. Therefore, $w_F^{OBMC}(u,v)$ is used as weighting factor for OBDC, while $w_F^{OBDC}(u,v)$ is used as weighting factor for OBMC.

The set of weights is used to generate the fused SI coefficient:

$$C_F^{b_k}(u,v) = \frac{w_F^{OBMC}(u,v)C_{OBDC}^{b_k}(u,v) + w_F^{OBDC}(u,v)C_{OBMC}^{b_k}(u,v)}{w_F^{OBDC}(u,v) + w_F^{OBMC}(u,v)} \quad (11)$$

and the corresponding residual estimation for the fused coefficient of the SI:

$$C_{R,F}^{b_k}(u,v) = \frac{w_F^{OBMC}(u,v)C_{R,OBDC}^{b_k}(u,v) + w_F^{OBDC}(u,v)C_{R,OBMC}^{b_k}(u,v)}{w_F^{OBDC}(u,v) + w_F^{OBMC}(u,v)} \quad (12)$$

To use the correlation noise model of [6], the coefficients $C_F^{b_k}(u,v)$ need to be divided into the inlier cluster and outlier clusters. Therefore (11) is used to calculate $C_F^{b_l}(u,v)$, $0 \le l < k$. The coefficients $C_F^{b_l}(u,v)$ and the estimation function defined in [6] are used to segment the coefficients $C_F^{b_k}(u,v)$ in the two clusters. The three SIs for $k > 0$ are fused using the distribution fusion framework. The final joint distribution is defined as follows:

$$f_{Fus}^{b_k,(u,v)} = \lambda f^{b_k,(u,v)} + (1 - \lambda) f_{X|Y_{Fus}}^{b_k,(u,v)} \quad (13)$$

where

$$\lambda = \frac{1}{2^k} \quad (14)$$

and $f^{b_k,(u,v)}$ is defined in (10).

The adaptive computation of the λ parameter assures that a low weight is selected for the fused SI when the fused SI is not reliable, but it increases rapidly, in line with the expected increase in reliability of the fused SI. The conditional probability of each bit in the SI can be calculated, taking into account the previously decoded bitplanes and the correlation noise model described by $f_{Fus}^{b_k,(u,v)}$. The decoded bitplanes determine the intervals $[L,U)$ in which each coefficient belongs to. To reconstruct the coefficient in position (u,v), the optimal reconstruction proposed in [14] is used, which is the expectation of the coefficient given that the available SIs are the following:

$$C_{Rec}^{b_k}(u,v) = \frac{\int_L^U x f_{Fus}^{b_k,(u,v)}(x)dx}{\int_L^U f_{Fus}^{b_k,(u,v)}(x)dx} \quad (15)$$

This procedure is carried out for each band b_k, $0 \le k \le N_b$, where N_b is the maximum number of decoded bands, every time updating the weights $w_F^{OBMC}(u,v)$ and $w_F^{OBDC}(u,v)$. Once the band b_{N_b} is decoded, $C_F^{b_k,(u,v)}$ is calculated for each $N_b < k \le 16$, and they are used as coefficients in the reconstructed frame. For what concerns the reconstruction of the bands b_k, $0 \le k \le N_b$, they are reconstructed a second time to enhance the quality of the reconstructed frame. The segmentation into the inlier cluster and outlier cluster is calculated using the already reconstructed frame, i.e. the actual value of the decoded coefficient is used to determine the cluster it belongs to, as opposed to using the mapping function employed in the previous steps [6]. As residual, the difference between the previously decoded frame and the fused SI is used. In this case $\lambda = 0$ in the reconstruction since at this stage, the reliability of the fused SI is so high that it is not necessary to use the inter-view or temporal SIs.

5 Experimental results

In this section, the proposed coding tools of the previous section are evaluated using the DVC codec described in Section 3. Before presenting the experimental results obtained, the test conditions are first defined. Then, OBDC

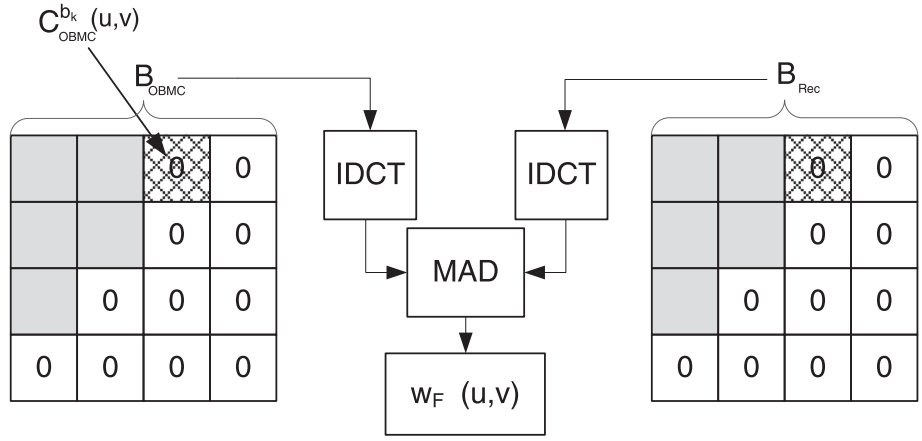

Figure 5 Calculation of the weights used for the refined fusion for OBMC.

is compared with DCVP, demonstrating the gains resulting from the pre-alignment phase. For fairness, DCVP employs OBMC for disparity estimation and compensation. Furthermore, the fusion algorithm performance is analysed comparing it with single SI decoders and alternative fusion techniques, using cameras at relatively close distance. Finally, the case of unknown disparity is analysed, examining the RD performance of the proposed decoder for 18 different camera configurations.

5.1 Test conditions

In the experiments, two sequences with still cameras and two sequences with moving cameras at constant inter-camera distance are analysed, in order to test the robustness of the system to global motion. The stream structure for the central view has GOP size 2. The full length of *Outdoor* and *Book Arrival* [22], 100 frames, is coded, and the first 10 s of *Kendo* and *Balloons* [22], i.e. 300 frames, is coded. For what concerns the spatial-temporal resolution, all the sequences are downsampled to CIF resolution:

- *Test sequences:* Outdoor, Book Arrival, Kendo and Balloons [22]. These sequences are characterized by different types of motion content, depth structures and camera arrangements, providing a meaningful and varied set of test conditions as outlined in Table 1; in the 'Interval of used views' column, '1' corresponds to the rightmost view (among the recommended views [23]). In the experiments, the central view is kept fixed while the distance between the central and the lateral cameras is increased, spanning the intervals detailed in Table 1. The distance between two consecutive cameras is 6.5 cm [24] for Outdoor and Book Arrival, while the distance between two consecutive cameras in Kendo and Balloons is 5 cm [22].
- *WZ frames coding:* The WZ frames are encoded at four RD points (Q_i, $i = 1, 4, 7, 8$) corresponding to four different 4×4 DCT quantization matrices [13]. The RD point Q_i corresponds to the lowest bitrate and quality and the RD point Q_8 to the highest bitrate and quality. The remaining test conditions associated with the DCT, quantization, noise modelling and reconstruction modules are the same as in [6]. For the LDPCA coding, a code length of 6,336 bits is used, and a CRC check of 8 bits is employed to check the correctness of the decoded result.
- *KFs coding*: The KFs in the central view are H.264/AVC Intra-coded (main profile) as it is commonly done in e.g. [6]. The quantization parameter (QP) of the KFs is selected in order to have a similar decoded quality between WZ frames and KF for the same RD point. In Table 2, the QPs used for each RD point are reported. As previously said, the lateral views are coded with the same parameters as the KFs of the central view.
- *Quality and bitrate*: Only the bitrate and PSNR of the luminance component is considered, as it is

Table 1 Characteristics of the test sequences

Sequence	Depth structure	Motion content	Moving cameras	Interval of used views	Central view	Frame rate (fps)
Outdoor	Medium	Complex	No	1-15	8	15
Book arrival	Complex	Medium	No	1-15	8	15
Kendo	Medium/Complex	Complex	Yes	1-5	3	30
Balloons	Medium/Complex	Medium	Yes	1-5	3	30

Table 2 Quantization parameters for the test sequences

Sequence	Q_1	Q_4	Q_7	Q_8
Outdoor	38	32	28	23
Book arrival	39	36	29	25
Kendo	39	36	29	22
Balloons	33	30	24	20

commonly done in literature. Both WZ frames and KFs are taken into account in rate and PSNR calculations. The rate and PSNR of the lateral views are not taken into account in order to better assess the performance of the proposed M-DVC solution.

5.2 OBDC-based SI performance assessment

In this section, the RD performance of the DVC solution using OBDC, with the sliding window approach, is assessed and compared with the one achieved when DCVP is used to generate the (inter-view) SI; the only difference between OBDC and DCVP is the pre-alignment phase. Table 3 shows the Bjøntegaard bitrate savings (BD-Rate) and Bjøntegaard PSNR gains (BD-PSNR) [25] between OBDC and DCVP when using as lateral views the ones closest to the central view (lowest disparity case), i.e. views 7 and 9 for Outdoor and Book Arrival and views 2 and 4 for Kendo and Balloons. Both SIs are evaluated using the same single SI decoder [6]. For DCVP, the parameters (e.g. search range, strength of the motion smoothing) are adapted to obtain the best average result in terms of RD performance and then the same parameters are used for OBDC. Such parameters are used in OBDC for all the sequences and for all the configurations (distance of the lateral cameras). As it can be observed from Table 3, OBDC allows improvements of the DVC codec RD performance when compared to DCVP, with PSNR gains up to 1.17 dB for the Book Arrival sequence, which is characterized by a complex depth structure. No appreciable gains are reported for Outdoor, the sequence displaying the simplest depth structure. Table 4 shows the BD-Rate savings and BD-PSNR gains between OBDC and DCVP when using as lateral views the ones furthest away from the central view (according to the view interval indicated in Table 1), i.e. views 1 and 15 for Outdoor and Book Arrival, and views 1 and 5 for Kendo and Balloons. In this case, the parameters for OBDC are the same as those used for generating the results in Table 3. On the other hand, the performance of DCVP is maximized through extensive simulations, finding, for each sequence, the parameters giving the best RD performance. It was not possible to find parameters which were able to perform well for all the sequences for DCVP, while, with the pre-alignment phase in OBDC, the disparity between views is normalized, leaving to the disparity estimation module the task to accommodate for minor differences.

Table 3 BD-Rate savings and BD-PSNR gains for OBDC with respect to DCVP, lower disparity

Sequence	BD-PSNR (dB)	BD-Rate (%)
Outdoor	0.00	0.00
Book arrival	1.17	−17.29
Kendo	0.18	−2.80
Balloons	0.27	−3.94

The lateral views are the ones closest to the central view.

Table 4 BD-Rate savings and BD-PSNR gains for OBDC with respect to DCVP, higher disparity

Sequence	BD-PSNR (dB)	BD-Rate (%)
Outdoor	0.63	−9.33
Book Arrival	1.52	−21.26
Kendo	0.90	−13.04
Balloons	0.90	−12.26

The lateral views are the ones furthest away from the central view.

5.3 M-DVC RD performance assessment

In this section, the RD performance of the proposed M-DVC coding solution is assessed and compared directly with the M-DVC scheme MDCD-Lin. The RD performance for distributed decoding based on only-motion SI and only-inter-view SI is also presented. Finally, the performance of predictive monoview codecs is provided for further comparison. The left, right and central views used in the experiments are reported in Table 5.

5.3.1 Coding benchmarks

The proposed M-DVC coding solution (described in Section 4) is compared with the following DVC-based codecs:

- OBMC: Single SI decoder, as presented in [6]. It is a single-view DVC solution, since it exploits the temporal correlation only.
- OBDC: Single SI decoder; OBDC is used as SI (outlined in Section 4.1). It exploits the inter-view

Table 5 Views used for assessing the proposed M-DVC coding solution RD performance

Sequence	Number of right view	Number of central view	Number of left view
Outdoor	6	8	10
Book arrival	6	8	10
Kendo	2	3	4
Balloons	2	3	4

Table 6 BD-Rate savings and BD-PSNR gains for the proposed M-DVC coding

Sequence	OBDC		OBMC		DISCOVER		MDCD-Lin	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
Outdoor	**0.90**	**−12.17**	1.12	−14.55	2.05	−25.36	*0.58*	*−7.70*
Book arrival	1.05	−15.64	**0.72**	**−10.96**	1.01	−15.47	0.74	−11.27
Kendo	**0.79**	**−11.73**	0.94	−13.92	1.53	−22.36	*0.58*	*−9.00*
Balloons	1.50	−20.23	**0.50**	**−7.14**	0.68	−9.81	0.59	−8.65
Average	1.06	−14.94	0.82	−11.64	1.32	−18.25	0.62	−9.16

The results are provided using boldface and italics. Boldface indicates the best-performing single SI-based DVC solution. Italics indicates the cases where MDCD-Lin is robust.

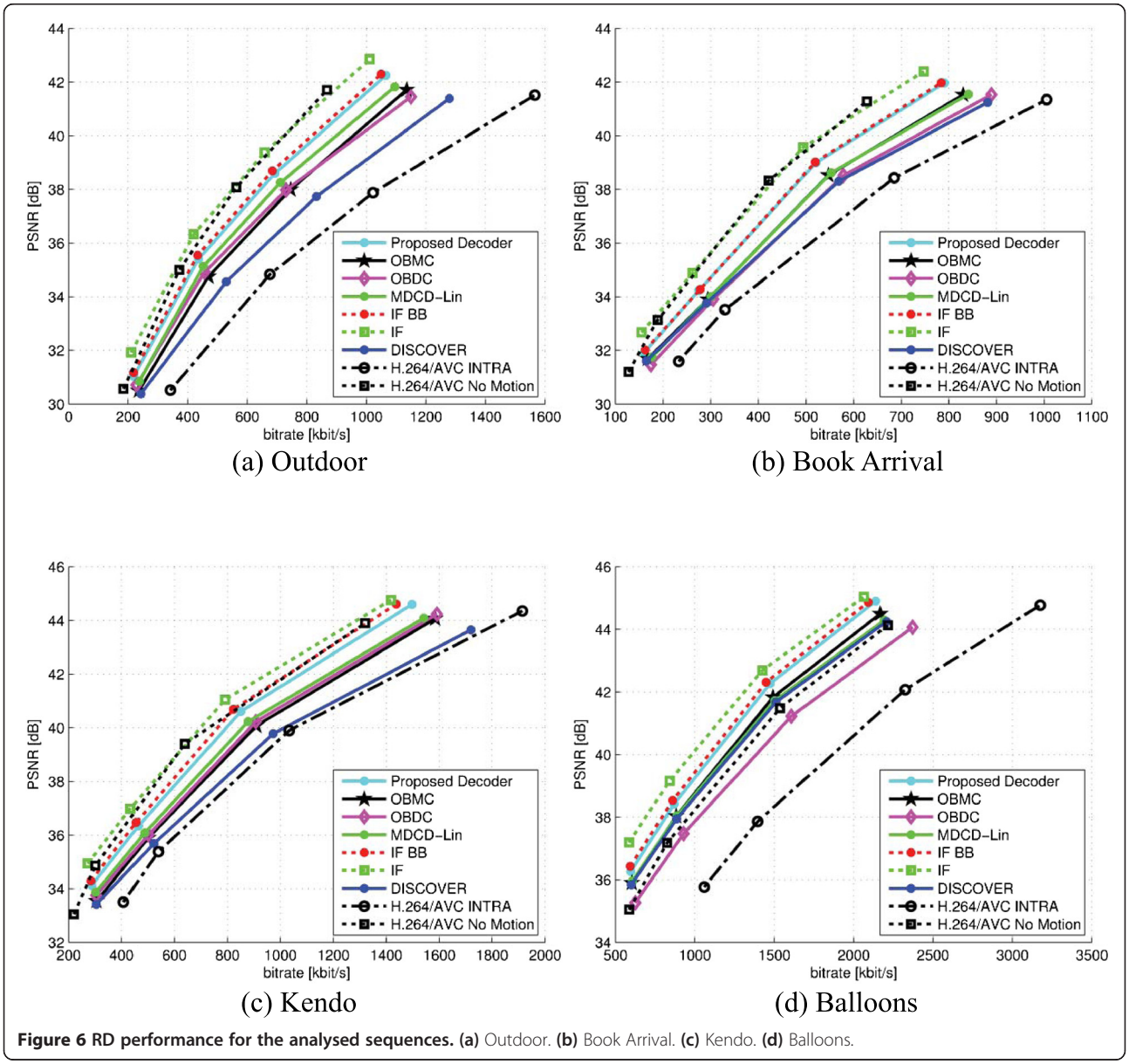

Figure 6 RD performance for the analysed sequences. (a) Outdoor. **(b)** Book Arrival. **(c)** Kendo. **(d)** Balloons.

Table 7 Outdoor video sequence: improvements for the proposed M-DVC solution for different Δ values

Δ	OBDC		OBMC		MDCD-Lin	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
1	**0.78**	**−10.53**	1.33	−16.94	*0.64*	*−8.43*
2	**0.90**	**−12.17**	1.12	−14.55	*0.58*	*−7.70*
3	0.97	−13.10	**0.96**	**−12.69**	*0.55*	*−7.38*
4	1.10	−14.87	**0.79**	**−10.58**	*0.56*	*−7.41*
5	1.42	−18.76	**0.60**	**−8.22**	*0.59*	*−8.02*
6	1.31	−17.66	**0.65**	**−8.82**	*0.50*	*−6.88*
7	1.39	−18.56	**0.56**	**−7.71**	*0.54*	*−7.30*

The results are provided using boldface and italics, following the conventions of the previous section.

correlation for the majority of the frame, while the temporal correlation is used for the rest.

- MDCD-Lin: Motion and disparity compensated linear fusion is the main benchmark. It is summarized in Section 2 and implemented following [8]. The weights (calculated from the on-line residuals) used to fuse the SIs are also used to fuse the corresponding residuals of the two SIs, to take into account that a wrong fusion has repercussions not only on the SI quality but also on the quality of the residual (which impacts the correlation model accuracy). The SI and the residual estimation are fed into the single SI decoder of [6]. While newer techniques were proposed [9], they were unable to provide consistent gains over MDCD-Lin. Therefore, MDCD-Lin is employed as benchmark.
- DISCOVER: this DVC-based codec [13] is still widely used as benchmark in literature. The system used as basis for the codec [6] has a structure which is similar to DISCOVER, but it uses an enhanced SI generation module (OBMC) and an advanced noise modelling algorithm. DISCOVER is reported only for completeness, but the focus will be the comparison with the other DVC coding solutions: the OBMC and OBDC-based baseline decoders, in order to make clear how the proposed tools improve the RD performance of the system.

For comparison, the performance of the proposed method is also compared with bounds given by ideal fusion techniques:

- IF BB: Summarized in Section 2. The SI and the residual estimation are fed into the single SI decoder detailed in [6]. The weights are used to fuse SIs and estimated residuals of the SIs.
- IF: Summarized in Section 2. The SI and the residual estimation are fed into the single SI decoder detailed in [6]. The weights are used to fuse SIs and estimated residuals of the SIs.

The proposed M-DVC decoder is finally compared with the following standard predictive coding schemes for reference:

- H.264/AVC Intra: It is the H.264/AVC codec (Main profile) with only the Intra modes enabled. It is also used for coding the KFs and lateral views. It is also a low-complexity encoding architecture;
- H.264/AVC No Motion: Exploits the temporal redundancy in an IB prediction structure setting the search range of the motion compensation to zero; therefore, the motion estimation part, which is the most computationally expensive encoding task, is not performed: the co-located blocks in the backward and/or forward reference frames are used for prediction.

Table 8 Book arrival video sequence: improvements for the proposed M-DVC solution for different Δ values

Δ	OBDC		OBMC		MDCD-Lin	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
1	**0.63**	**−9.45**	1.00	−14.85	*0.63*	*−9.58*
2	1.05	−15.64	**0.72**	**−10.96**	0.74	−11.27
3	1.47	−21.48	**0.52**	**−8.02**	1.00	−15.25
4	1.76	−25.39	**0.42**	**−6.55**	1.23	−18.62
5	1.95	−27.85	**0.26**	**−4.02**	1.40	−21.03
6	2.30	−32.31	**0.08**	**−1.25**	1.56	−23.35
7	3.24	−42.82	**−0.06**	**0.99**	2.25	−32.48

The results are provided using boldface and italics, following the conventions of the previous section.

Table 9 Kendo video sequence: improvements for the proposed M-DVC solution for different Δ values

Δ	OBDC		OBMC		MDCD-Lin	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
1	**0.79**	**−11.73**	0.94	−13.92	*0.58*	*−9.00*
2	1.01	−14.86	**0.62**	**−9.36**	*0.62*	*−9.37*

The results are provided using boldface and italics, following the conventions of the previous section.

5.3.2 RD performance

Table 6 reports the BD-Rate savings and BD-PSNR gains for the proposed M-DVC coding solution when compared to the baseline OBMC and OBDC-based DVC coding solutions, using the tools proposed in [6]. For each sequence, the best-performing single SI-based DVC solution is identified in boldface. The proposed M-DVC video coding solution is able to consistently outperform the best single SI-based DVC solution, with PSNR gains up to 0.9 dB. In the worst-case scenario, Balloons, the improvement is still significant, allowing a bitrate reduction up to around 7%. The results for the DISCOVER codec are also provided, and the average BD-Rate savings are around 18%. For what concerns the comparison with MDCD-Lin, the proposed method shows an average BD-PSNR gain of 0.62 dB. The improvement is robust, ranging from 0.58 to 0.74 dB. The gains of the proposed method over MDCD-Lin are in italics if MDCD-Lin is robust, i.e. if it is able to outperform both the single SI OBMC-based decoder and the single SI OBDC decoder.

Figure 6 reports the RD performance results obtained for the Outdoor, Book Arrival, Kendo and Balloons, for the nine coding solutions mentioned above. The proposed solution outperforms OBMC, OBDC, DISCOVER and MDCD-Lin, which are all four truly distributed decoders, i.e. they do not require the WZ frame. More specifically, the BD-PSNR gains of the proposed solution are up to 1.5 dB when compared with OBDC and up to 1.12 dB when compared with OBMC. The proposed decoder is able to outperform DISCOVER by up to 2 dB because DISCOVER uses less advanced SI generation systems and correlation noise model. MDCD-Lin is able to robustly fuse the SIs for Outdoor, Book Arrival and Kendo but not for Balloons. Furthermore, for the first three sequences, the improvements achieved with MDCD-Lin are lower when compared with the proposed solution, achieving BD-PSNR gains up to 0.33 dB for Outdoor. Therefore, the proposed solution, leveraging the fusion based on the distributions and the learning process, is able to outperform the other realistic distributed decoders. The use of weights derived from the distributions allows a more precise fusion because the correlation noise modelling is built on the premise that the residual may have errors. The learning process allows a refinement of the fused SI while decoding the frame, improving the SI quality accuracy by performing a more accurate SI fusion process. The ideal fusion-based coding solutions, IF and IF BB, require the original WZ frame. Therefore, they provide a bound but they cannot be used in practice. The BD-PSNR gains of IF BB over the proposed coding solution range from 0.02 dB for Book Arrival to 0.28 dB for Kendo. This shows that the proposed system is able to reach performance close to an ideal block-based fusion technique. However, pixel-level ideal fusion shows gains by up to 1.14 dB BD-PSNR, over the proposed coding solution for the Outdoor sequence. For what concerns the reference predictive coders, H.264/AVC Intra is outperformed by every distributed coding solution, regardless of the SI generation method. The proposed decoder is able to reach RD performance comparable with H.264/AVC No Motion for Kendo and Balloons. For Outdoor and Book Arrival, the only distributed decoder able to compete with H.264/AVC No Motion is the one with a pixel-level IF. However, notice that H.264/AVC No Motion requires much higher encoding complexity since it has to test several Intra and Inter modes using as reference the neighbouring or co-located blocks. It is difficult to provide a complete comparison with more recent works, such as [12], given that resolution and the distance between cameras are different, i.e. different test conditions are used. Nevertheless, for the same views used in [12], we produced results for MDCD-Lin. The technique proposed in [12], referred to as AV, is able to outperform MDCD-Lin by 0.61 dB as average of the BD-PSNR values for the four sequences. It has to be noted that MDCD-Lin is used to fuse MCTI and DCVP, while the results for AV in [12] are based on fusing better

Table 10 Balloons video sequence: improvements for the proposed M-DVC solution for different Δ values

Δ	OBDC		OBMC		MDCD-Lin	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
1	1.50	−20.23	**0.50**	**−7.14**	0.59	−8.65
2	1.90	−24.92	**0.34**	**−4.88**	0.75	−10.74

The results are provided using boldface and italics, following the conventions of the previous section.

performing SIs. The proposed method is able to achieve a similar improvement over MDCD-Lin (0.62 dB), but in this case the comparison is done using the same SIs for both fusion architectures. Direct comparison with [12] is difficult because different resolutions are used. Nevertheless, for the four analysed sequences, AV is able to perform well on Balloons (2.2 dB gain [12]), but the gains are minor (0.0 to 0.13 dB [12]) for the other three sequences we consider. The proposed method is, on the other hand, able to provide reasonably robust gains (0.58 to 0.74 dB, Table 6) on all four sequences. As a final note, it can be seen that the occlusion detection mechanism presented in [12] addressed occlusions in the areas where the different views overlap. The proposed method removes the areas that are occluded because they do not belong to the part of the views that overlap. It is reasonable to think that combining both approaches can lead to even higher gains.

5.4 Camera distance impact

This section assesses the impact of varying the distance between the lateral and the central views on the M-DVC codec RD performance. The test conditions are similar to the ones used in the previous subsection except for the choice of the lateral views. Tables 7, 8, 9, 10 show the BD-Rate savings and BD-PSNR gains for the proposed M-DVC solution with respect to the baseline OBMC and OBDC-based DVC coding solutions when varying the distance between the cameras for the Outdoor, Book Arrival, Kendo and Balloons sequences. The BD gains of the proposed method with respect to MDCD-Lin are also provided. (The results are provided using boldface and italics, following the conventions of the previous section.) The Δ value refers to the difference between the index of the central camera and the index of the right camera. It has to be noted that the same value of Δ may refer to different inter-camera spacing depending on the cameras arrangement. According to the results obtained, the proposed M-DVC solution is robust to changes in disparity: Outdoor, which is characterized by a simpler depth structure, shows a much more stable performance when compared with Book Arrival. Only in one case, out of the 18 examined cases, the proposed fusion solution is unable to perform better than the best single SI based DVC solution, but the performance loss is negligible, and the BD between the RD performance of the two single SI decoders (one using OBMC, the other using OBDC) is more than 3 dB, making the problem of increasing the performance by fusion extremely hard. For what concerns the performance comparison with MDCD-Lin, the gains of the proposed method, in BD-PSNR, range from 0.50 dB (Outdoor, $\Delta = 6$) to 2.25 (Book Arrival, $\Delta = 7$). The proposed method shows higher stability and robustness when compared with MDCD-Lin, which is unable to efficiently fuse SI having too different quality. It has to be noted that, as opposed to [12], MDCD-Lin fuses the same SIs used by the proposed method; therefore here the assessment is purely based on the performance of the fusion algorithm.

6 Conclusions

In this paper, a novel fusion approach is proposed, based on learning and fusion of the distributions, rather than fusion of the pixels of the SIs. This allows simplifying the problem of estimating the residual of the fused SI and allows the M-DVC solution to leverage well-known techniques for residual estimation and correlation noise model calculation developed for single SI DVC schemes. The proposed M-DVC coding solution proved to be robust to both increments and decrements of the distance between the cameras, which could be a desirable feature in systems where cameras can move with respect to each other or in systems where the distance between cameras is unknown. The proposed learning approach achieved a superior RD performance, on average, when compared with single SI decoders and it showed higher robustness than a residual-based SI fusion technique. The proposed fusion reached performance similar to the performance bounds obtained with a block-based ideal fusion, which relies on the knowledge of the original WZ frame. In case of cameras moving with respect to the scene, but keeping a fixed disparity, the M-DVC solution was able to achieve results that are close to H.264/AVC No Motion, and in the case of fixed cameras, the difference is relatively small, in particular, when compared with the RD performance loss of single SI DVC solutions.

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

Author details
[1]DTU Fotonik, Technical University of Denmark, Ørsteds Plads, 2800 Kongens Lyngby, Denmark. [2]Instituto Superior Técnico, Lisbon 1049-001, Portugal.

Received: 29 April 2014 Accepted: 12 November 2014
Published: 4 December 2014

References

1. B Girod, AM Aaron, S Rane, D Rebollo-Monedero, Distributed video coding. Proc. IEEE **93**(1), 71–83 (2005)
2. 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)
3. C Guillemot, F Pereira, L Torres, T Ebrahimi, R Leonardi, J Ostermann, Distributed monoview and multiview video coding. Signal Process. Mag. IEEE **24**(5), 67–76 (2007)
4. D Slepian, J Wolf, Noiseless coding of correlated information sources. IEEE Trans. Inf. Theory **19**(4), 471–480 (1973)
5. A Wyner, J Ziv, The rate-distortion function for source coding with side information at the decoder. IEEE Trans. Inf. Theory **22**(1), 1–10 (1976)
6. X Huang, S Forchhammer, Cross-band noise model refinement for transform domain Wyner-Ziv video coding. Signal Process. Image Commun. **27**(1), 16–30 (2012)
7. A Vetro, T Wiegand, GJ Sullivan, Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. Proc. IEEE **99**(4), 626–642 (2011)

8. T Maugey, W Miled, M Cagnazzo, B Pesquet-Popescu, Fusion schemes for multiview distributed video coding, in *Proceedings of European Signal Processing Conference* (Glasgow, 2009), pp. 559–563
9. F Dufaux, Support vector machine based fusion for multi-view distributed video coding, in *Proceedings of Digital Signal Processing (DSP)* (Corfu, 2011), pp. 1–7
10. M Ouaret, F Dufaux, T Ebrahimi, Multiview distributed video coding with encoder driven fusion, in *Proceedings of the 2007 European Signal Processing Conference (EUSIPCO-2007)* (Poznan, 2007)
11. X Artigas, F Tarrés, L Torres, Comparison of different side information generation methods for multiview distributed video coding, in *Proceedings of SIGMAP 2007* (Barcelona, 2007)
12. G Petrazzuoli, M Cagnazzo, B Pesquet-Popescu, Novel solutions for side information generation and fusion in multiview DVC. Eurasip J. Adv. Signal Process **1**, 154 (2013)
13. X Artigas, J Ascenso, M Dalai, S Klomp, D Kubasov, M Ouaret, The DISCOVER codec: architecture, techniques and evaluation, in *Proceedings of Picture Coding Symposium (PCS) 2007* (Lisbon, 2007)
14. D Kubasov, J Nayak, C Guillemot, Optimal reconstruction in Wyner-Ziv video coding with multiple side information, in *Proceedings of IEEE MMSP 2007* (Chania, Crete, 2007), pp. 183–186
15. X Huang, C Brites, J Ascenso, F Pereira, S Forchhammer, Distributed video coding with multiple side information, in *Proceedings of Picture Coding Symposium (PCS) 2009* (Chicago, Illinois, 2009), pp. 385–388
16. Y Li, H Liu, X Liu, S Ma, D Zhao, W Gao, Multi-hypothesis based multi-view distributed video coding, in *Proceedings of Picture Coding Symposium (PCS) 2009* (Chicago, Illinois, 2009), pp. 1–4
17. M Salmistraro, M Zamarin, S Forchhammer, Multi-hypothesis distributed stereo video coding, in *Proceedings of MMPS 2013* (Pula, Sardinia). 30 September to 2 October 2013
18. H Luong, LL Raket, X Huang, S Forchhammer, Side information and noise learning for distributed video coding using optical flow and clustering. IEEE Trans. Image Process. **21**(12), 4782–4796 (2012)
19. C Brites, J Ascenso, F Pereira, Learning based decoding approach for improved Wyner-Ziv video coding, in *Proceedings of PCS 2012* (Krakow, 2012), pp. 165–168
20. M Ouaret, F Dufaux, T Ebrahimi, Iterative multiview side information for enhanced reconstruction in distributed video coding. J Image. Video. Process **2009**, 3:1–3:17 (2009)
21. D Varodayan, A Aaron, B Girod, Rate-adaptive codes for distributed source coding. Eurasip. Signal. Process. J **86**(11), 3123–3130 (2006)
22. A Smolic, G Tech, H Brust, *Report on Generation of Stereo Video Database*. Technical Report d2. 1 July 2010
23. Nagoya University - Tanimoto Laboratory, Kendo specifications, in www.tanimoto.nuee.nagoya-u.ac.jp/~fukushima/mpegftv/yuv/Kendo/readme.txt. Accessed 18 November 2014
24. I Feldmann, M Mueller, F Zilly, R Tanger, K Mueller, A Smolic, P Kauff, T Wiegand, *HHI Test Material for 3D Video* (ISO, Archamps, 2008). May 2008
25. G Bjøntegaard, Calculation of average PSNR differences between RD curves, in *VCEG 13th Meeting* (Austin, Texas, 2001)

doi:10.1186/1687-6180-2014-174
Cite this article as: Salmistraro *et al.*: **A robust fusion method for multiview distributed video coding.** *EURASIP Journal on Advances in Signal Processing* 2014 **2014**:174.