Depth-based multiview distributed video coding

Giovanni Petrazzuoli, *Member, IEEE,* Thomas Maugey, *Member, IEEE,* Marco Cagnazzo, *Senior Member, IEEE,* and Béatrice Pesquet-Popescu, *Fellow, IEEE*

***Abstract*—Multiview distributed video coding (DVC) has gained much attention in the last few years because of its potential in avoiding communication between cameras without decreasing the coding performance. However, the current results are not matching the expectations mainly due to the fact that some theoretical assumptions are not satisfied in the current implementations. For example, in distributed source coding the encoder must know the correlation between the sources, which cannot be achieved in the traditional DVC systems without having a communication between the cameras. In this work, we propose a novel multiview distributed video coding scheme, in which the depth maps are used to estimate the way two views are correlated with no exchanges between the cameras. Only their relative positions are known. We design the complete scheme and further propose a rate allocation algorithm to efficiently share the bit budget between the different components of our scheme. Then, a rate allocation algorithm for depth maps is proposed in order to maximize the quality of synthesizing virtual views. We show through detailed experiments that our scheme significantly outperforms the state-of-the-art DVC system.**

I. Introduction

Multiview video has recently gathered increased attention, thanks to the availability of new acquisition and rendering systems. This paves the way to numerous applications, such as 3D and free viewpoint TV [1]. In this context, the problem of efficient compression is more urgent than ever, in sight of the huge amount of data storage and transmission required by multiview video. On the one hand, the compression methods have evolved and are now able to exploit the correlation between viewpoints with increasing effectiveness (*e.g.,* view synthesis techniques [2]). On the other hand, the capture systems become complex and ambitious by covering larger scenes, such as sport/music events, museum, cities, etc. As a result, the acquisition system more than ever cannot provide communication between cameras, which makes the usage of traditional compression techniques impossible since they estimate the similarity between views at the encoder side relying on the knowledge of the content of every viewpoint.

In the last decade, an alternative paradigm has been developed in order to alleviate the inter-camera communication problems. The distributed source coding theory shows that two correlated sources X and Y can be transmitted with the same efficiency when they are jointly or independently encoded, as soon as the decoding is done jointly [3], [4]. One of the conditions for such a result is that the encoder knows the correlation between X and Y. Distributed codecs are based on the idea of using channel coding tools for source compression. Practical schemes for video have been developed mainly relying on the so-called Stanford approach [5], implemented within the DISCOVER project [6]. In this scheme the images are either key frames (KFs) or Wyner-Ziv frames (WZFs). KFs are INTRA coded and are used at the decoder to generate an estimation of the WZFs. This estimation, called side information (SI), is corrected by parity bits sent by the channel encoder. The underlying assumptions reflect those of the theory: i) the error between the side information and the Wyner-Ziv frame (or between Y and X) is stationary, and ii) the correlation between them is known, or, equivalently, the probability distribution function of this error is known at the encoder. However, none of these assumptions is completely verified in practice, which keeps the performance of distributed video coding schemes suboptimal compared to traditional compression standards [7]. While i) might be solved by improving side information generation techniques [7], the practical distributed coding schemes generally circumvent ii) by using a feedback channel or by relying on a light communication between cameras; both of these solutions may be very difficult to implement in practical DVC scenarios.

In the same period, the scene capture has also undergone a major change with the popularization of depth sensors systems. Indeed, with time-of-flight or structured light techniques [8], depth maps can now be efficiently and cheaply acquired. Then, the multi-view plus depth (MVD) format is becoming more and more popular (see 3D-HEVC [9]). Depth images offer a great potential for avoiding the two aforementioned limitations of multi-view DVC [10]. If both the depth and color images are available for a given viewpoint, one can estimate any other viewpoints using depth-based image rendering (DIBR) [11]. In these synthesized viewpoint images we can find unoccluded and occluded regions. The former are estimated using the texture from the reference viewpoint. If the depth data are perfect, they can be recovered completely under the Lambertian assumption[1]. On the contrary, the occlusion areas are not estimated at all, since they are parts of the scene that are not visible in the reference viewpoint. In these regions, the mean square error is equal to the variance of the image, since no estimation is provided. In summary, the knowledge of the depth map for a given viewpoint allows to first reconstruct part of other viewpoints and, second, localizes the errors in these estimations. The depth maps enable to estimate the level

G. Petrazzuoli, M. Cagnazzo and B. Pesquet-Popescu are with the Department of Image and Signal Processing, Institut Mines-Telecom, TELECOM-ParisTech 46 rue Barrault, F-75634 Paris, FRANCE, and with LTCI, CNRS. E-mail: {petrazzu,cagnazzo,pesquet}@telecom-paristech.fr. T. Maugey is with the Signal Processing Laboratory (LTS4), Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. E-mail: maugey@epfl.ch.

[1]The brightness of a point is the same if seen similarly from observers at different viewpoints.

of similarity of the current view with any other viewpoints, without requiring any communication between them.

In this paper, we develop a novel distributed coding architecture which relies on the depth information. As explained above, the idea is to use the depth maps to build the side information at the decoder and to estimate the level of similarity between two views at the encoder side. The occlusion areas to be sent are obtained by a double DIBR, and these areas are coded by a shape adaptive algorithm. The main advantage of our scheme is that we better fit to Wyner and Ziv's assumption, namely a knowledge of the correlation at the encoder side, without communication between cameras. The only assumption is that all cameras know the position of the other ones. This can be considered as an *a priori* knowledge. Another improvement of our scheme is that it is not linked to a particular error metric. More precisely, traditional Stanford-based distributed schemes only aims at minimizing the MSE, due to the fact that channel coders as turbocodes or LDPC are used for compression. On the contrary, our scheme avoids the channel coder (as in [12]) and so it can be independent from the correlation model among the sources. This makes possible to construct the scheme under the perspective of alternative error metrics without changing the architecture. In our experimental section, we however provide some tests in terms of Rate-PSNR and Rate-SSIM performance. The obtained results with these two quality metrics show that our scheme significantly outperforms the traditional DVC system and intra coding. Moreover, it sometimes even competes with schemes that do not respect distributed coding assumptions such as the layered-depth format [13]. Finally, our novel coding architecture allows to find a very effective solution to the rate allocation problem between key and non-key frames in order to maximize the PSNR of texture images. We are able to provide an allocation algorithm whose performance is very close to an ideal full-search allocation. This is a further improvement with respect to classical DVC systems, where the optimization of the rate allocation is made difficult by the rigid constraints of the channel coder (i.e. parity bits sent on demand in large chunks). Since depth maps allow to perform a free viewpoint navigation, we have also proposed a new technique for bit rate allocation for depth maps in order to maximize the PSNR on the virtual views.

The rest of the paper is structured as follows: in Section II, we position our work in the context of multiview distributed video coding. In Section III the new distributed architecture for MVD is described and a rate distortion allocation algorithm is proposed that maximizes the PSNR on original and on virtual views (Section IV). In Section V we show experimental results, while we draw conclusions and outline future work in Section VI.

II. Related work

Distributed video coding has been applied to multiview video mainly for avoiding inter-camera communication [14] or equivalently, a centralized encoding process. Distributed video coding schemes take mainly two forms in the literature: the PRISM [15] and Standford [5] approaches. The latter framework has proven to be the most competitive, leading to many more extensions such as the one developed by the European project DISCOVER [16]. Recently, other DVC techniques have been proposed in order to improve it, such as VISNET II [17]. One of the key aspects of Stanford-like DVC is the estimation of the WZF at the decoder, called side information. It can be generated, according to the frame repartition, by temporal interpolation or inter-view interpolation or by a fusion of them [18]–[22]. Several solutions [23], [24] have been explored in the literature to improve the quality of the SI and consequently the RD performance, despite improving the PSNR on SI does not imply directly maximizing RD performance [25]. As the quality of the inter-view generated SI is often poor, the RD performance of these schemes is not better than the INTRA mode of H.264/AVC [7], [26], [27]. Inter-view estimation methods suffer from a lack of knowledge about the scene geometry. This can be handled by using depth maps. Till now multiview video plus depth has not been deeply explored in the context of DVC. The information of the depth map can be used, for example, in order to improve the quality of the estimation of the WZF. DIBR algorithms can be used along with the camera parameters, in order to generate other views. For example, Artigas *et al.* [6] propose a method for texture SI generation based on depth maps. Given the KF image, the associated depth map and the camera parameters, it is possible to create a virtual viewpoint. The synthesized image suffers, nevertheless, from some drawbacks: occluded areas cannot be rendered, errors in depth maps generate annoying artifacts, and view-dependent image features such as reflections cannot be correctly interpolated. These problems are mitigated by an image fusion algorithm [18]–[21]. Recently, Salmistraro *et al.* [10], [28] have proposed different solutions in order to exploit depth information for SI generation for both texture and depth signals. In particular they propose optical flow techniques exploiting information given by the depth for the motion estimation. The generation of the texture SI jointly from texture and depth has not been further explored since it does not meet the essential hypotheses of DVC schemes based on DISCOVER [16]. First, the error between the generated SI and the real WZF is not Laplacian [29]. Second, this error is strongly non stationary: there are several regions not affected by errors and other ones affected by noise of high variance. Without these two hypotheses the channel bit allocation per band per bit plane is sub-optimal and this erroneous allocation strongly affects the Rate-Distortion performance of the whole system. Several works [30]–[32] have also proposed to remove the feedback channel, that leads to some problems in practical implementations. The drawbacks of these applications is that camera communication is needed and a loss in RD performance is observed (from 8% to 16% of bit rate increase). In this paper, we propose an alternative to traditional DVC systems, where we exploit depth maps as crucial information on the scene geometry (*e.g*, from which we derive a correlation model). This permits to completely get rid of inter-camera communication and also to suppress the feedback channel, without any communication between the sources at the encoder side. Therefore, the rate allocated for non-key cameras can be varied more finely, allowing improved

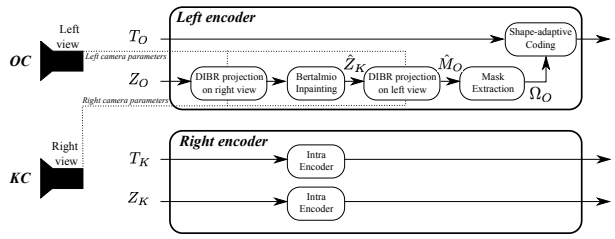


Fig. 1. Structure of the encoder: the right camera is classically encoded. For the left camera the occlusion mask is extracted by a double DIBR run on the depth map. Only the regions selected by this mask are encoded and sent to the decoder.

results.

III. Proposed architecture

We consider two range cameras that provide a texture image and the corresponding depth map each, and we discuss later in this section the case of a generic number of cameras. For one view (say, the right-hand one), texture and depth are encoded in INTRA mode (as in Fig. 1)[2]. Using these two images, at the decoder side, DIBR is applied and the left-hand view can be reconstructed except for the occluded regions, *i.e.* the areas that are visible in the left view but not in the right one. In order to fill in the occluded regions, the left-hand camera should know where the occluded regions are in the right view; however inter-camera communication is impossible in the distributed coding paradigm. Therefore, we propose to estimate the occluded regions without inter-camera communication, by only exploiting the geometric information, such as the depth map and the camera parameters. The left-hand camera will only send an efficient representation of the occluded areas.

Let us now establish some notation for describing our system. The camera that is INTRA encoded is called key camera (KC), a naming convention very common in the DVC literature. We call the other camera, for which only occluded regions are sent, an occlusion camera (OC). The texture and the depth captured by the KC are referred to as T_K and Z_K respectively (since our algorithm is independent from the time, we omit the dependence on a temporal variable). Likewise, texture and depth from OC are referred to as T_O and Z_O. As already mentioned, KC coding is straightforward, therefore let us consider OC coding. It consists of two parts: estimation of the occluded areas and their encoding by a shape-adaptive algorithm. The occlusion mask M_O for the OC is defined as

$$M_O(m,n) = \begin{cases} 1 & \text{if } (m,n) \text{ is visible in the OC but} \\ & \text{not in the KC} \\ 0 & \text{if } (m,n) \text{ is visible to both cameras} \end{cases}$$

This mask is not available at the OC, so it has to be estimated. To this end, we apply DIBR followed by a Bertalmio inpainting [33] to Z_O. Bertalmio inpainting is well suited for the depth maps, since it consists in an anisotropic diffusion [34]. Using the resulting estimated depth map $\widehat{Z}_K$, we perform a back projection into the OC coordinate system. We obtain thus the estimate of the occlusion map as seen by the KC, and we refer to it as $\widehat{M}_O$. Fig. 2 shows an example of estimated occlusion map for the "dancer" MVD sequence. In principle, all the texture pixels corresponding to the points with $\widehat{M}_O = 1$ should be encoded and sent, in order to allow the occlusion filling. We observe that many small regions appear in the estimated map (we define as a region a set of connected pixels where $\widehat{M}_O = 1$). These small regions are relatively costly to encode, while the corresponding pixels can be effectively filled in by inpainting at the decoder, since in many cases they result from noise or depth errors, rather than from actual occlusions. For this reason, we remove from $\widehat{M}_O$ the connected regions smaller than a certain number p of pixels. The threshold p cannot be too large, otherwise we risk to loose important information that cannot be recovered by the decoder. The optimal value of this parameter is empirically determined (Section V). The second processing step consists in slightly enlarging the mask, in order to take into account the fact that there may be other actually occluded pixels falling outside the estimated mask [35]. We show in Fig. 3 an example of this phenomenon: we mark in red occluded pixels that are not in M_O. These missing pixels will be hardly reconstructed by inpainting. The mask enlargement is performed by a dilation using a disk-shaped structuring element with a radius of ρ pixels. Increasing the disk radius would eventually assure that all the occluded pixels are included in the occlusion mask, but would also increase the coding cost of the occluded regions. As for the minimum region area p, the best value for the disk radius ρ is determined by experiments, as shown in Section V. The occlusion mask resulting from the dilation is referred to as Ω_O. We compare in Fig. 4 the estimated occlusion map $\widehat{M}_O$ with the its refined version: the latter appears more suitable for selecting the regions to be encoded.

The regions selected by the mask are encoded using the shape-adaptive (SA) algorithm proposed in [36], [37]. A SA wavelet transform [38] is carried out on each region. This transform preserves the spatial correlation and the self-similarity across subbands, which is crucial for the following zero-tree coding algorithm. The resulting coefficients are encoded by a SA version of SPIHT [39] that differs from the original in two major aspects: first, only nodes belonging to the support of the SA transform are considered while scanning a spatial orientation tree. Second, the baseband coefficients are no longer grouped into 2×2 square and a single root is considered instead. Finally, an optimal rate allocation among the different regions is performed.

The decoder structure is depicted in Fig. 5. Using the decoded KC texture and depth, the actual occlusions can be computed and part (hopefully all) of them can be filled by the regions encoded with the SA algorithm. If some areas in the texture image are still unfilled, they are recovered using Criminisi inpainiting [40]. Differently from depth maps, that are essentially textureless images, Criminisi inpainting is more suitable for textured image. Finally, the synthesized depth map is also inpainted, obtaining a decoded depth map $\widetilde{Z}_O$, to be

[2]We could also encode this view using temporal prediction, *i.e.* motion estimation. However, we consider in the following an All-INTRA encoding in order to keep the complexity low.

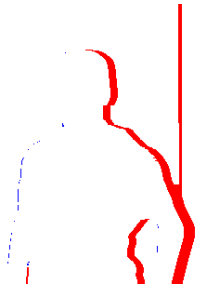


Fig. 2. Examples of occlusion regions in a detail from the dancer sequence (view 9, frame 1): the occlusions whose area is smaller than 50 pixels are in blue; occlusions larger than the threshold are in red.

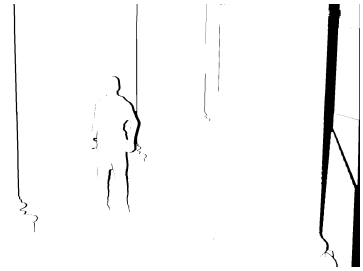
Fig. 3. A detail of an OC without dilation of occluded areas (view 9, frame 1). The non-filled occlusion areas are in red.

used for synthesizing other viewpoints.

This scheme can easily be extended to an arbitrary number of cameras, since the encoding of the KC and of the OC is totally independent of the number of cameras, perfectly abiding to the distributed coding paradigm. Of course, if we have many OCs and some of them are too far away from a KC, their estimated occlusion area may cover a large part of the scene, thus increasing the coding rate. However we remark that with respect to the classical DVC architecture, we have more flexibility in positioning key cameras and non-key cameras. For example, in the common case of three-camera configuration, we can use a central KC with two lateral OCs; in the case of a classical DVC system instead, inter-view estimation works well enough only when a Wyner-Ziv frame is interpolated from two adjacent key frames (extrapolation gives worse performance [41]).

IV. Bit rate allocation

Usually in DVC the rate allocation between KFs and WZFs is empirically obtained [16]: a quantization index for encoding the transform coefficients of WZFs is chosen according to the quantization parameter used for the KFs, with the goal of having the same distortion both for KFs and WZFs. Those coefficients are fed into a channel coder to produce the parity bits. The parity bits are sent on demand and in relatively large sets, called chunks, until the bit error rate BER at the decoder side drops below a given threshold. In summary, the rate allocation between KFs and parity bits is suboptimal and requires a feedback channel to be implemented. The DVC



(a)

(b)

Fig. 4. The unprocessed (a) and the processed (b) occlusion map (view 9, frame 1) from the *dancer* sequence

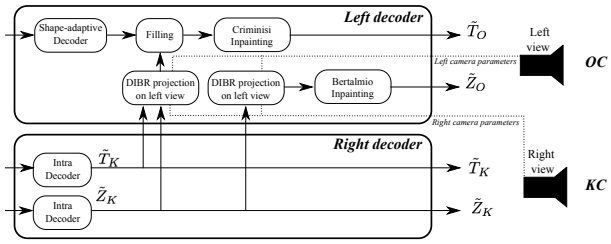


Fig. 5. Structure of the decoder: the right texture and depth can reconstruct the un-occluded regions of the left view. At the decoder side, the occluded regions are filled by the objects sent by the encoder.

architecture proposed in the previous section allows a better rate allocation procedure. We show here that the performance of an "oracle" system, that knows the distortion associated to any rate allocation choice, can be approached by a heuristic algorithm based on the characteristics of both our system and of the MVD signal. In a first part (Sec. IV-A) we will explore the effect of rate allocation on the original views. The impact on the synthesized views is investigated in Sec. IV-B.

A. Rate allocation for original views

Let us start by the problem formulation. The total available bit rate R must be allocated among the KC and the OC. Let R_K and R_Z be the rates associated to T_K and Z_K, i.e. the key texture and depth respectively. Moreover, let R_{SA} be the rate associated to the SA coding of occluded regions in OC. We call D_K and D_O the distortions for the KC and the OC textures, respectively. As metric for the distortion we use the MSE, but we could use other additive metric. Our goal is to minimize the total distortion subject to the constraint on the total rate:

$$\text{Minimize } D = D_K + D_O \quad \text{subject to} \quad R_{SA} + R_K + R_Z = R \tag{1}$$

We observe that the distortion D_K depends only on R_K, and the $D_K(R_K)$ function is in principle available at the KC encoder. The term D_O is more complex to evaluate: it is made up of the distortion on the unoccluded (synthesized) areas, D_U, plus the distortion on the occluded areas, encoded with the shape adaptive algorithm, D_{SA}. Both terms must be weighted by the relative sizes of the corresponding areas. In turn, the distortion D_U depends on the key texture and key depth rates: $D_U = D_U(R_K, R_Z)$. The relationship between R_K, R_Z and D_U is difficult to be mathematically modeled. However it is possible to numerically compute it, when one has simultaneously access to all the data T_K, Z_K, T_O, Z_O. Finally, $D_{SA}(R_{SA})$ is the rate-distortion curve of the shape adaptive encoder, and it is available at the OC encoder. In summary the total distortion may be written as:

$$\begin{aligned} D(R_K, R_Z, R_{SA}) = D_K(R_K) + \beta D_{SA}(R_{SA}) \\ +(1-\beta) D_U(R_K, R_Z) \end{aligned} \quad (2)$$

where β is the ratio between the number of pixels of the occlusions and total the number of pixels of T_O. The values of β for our test sequences range from 1.60 % to 8.40 %, with an average of 3.68 %.

The constrained minimization could in principle be solved by a full-search minimization (i.e. trying all the possible rate allocations). However this is only possible in a simulation environment, not only for its complexity, but also because the distortion computation is only possible when all decoded and original images are available. An analytic minimization is difficult to perform since it is hard to find a reliable model of $D_U(R_K, R_Z)$. Therefore, we design an effective heuristic rate allocation algorithm based on the characteristics of our system and of the MVD data. We show that our algorithm gives RD results that are very close to the full-search, at least regarding the allocation between the KC and the OC data.

More precisely, we start on a very common assumption about R_K and R_Z, i.e. the rates needed to encode a texture and its depth map. The relationship between them has been explored in many previous works [42]–[45], and in the reference software of the upcoming 3D-HEVC coding standard it is implemented as a simple empirical rule that associates the depth map quantization parameter to the one used for the texture [46], as shown in Tab. I. We use the same rule here and, as a consequence, rates R_Z and R_K are functions of the single quantization parameter QP. Now, we have to find the relationship between the optimal QP and R_{SA}. Since we have reduced the dimensionality of our problem from a three-dimensional search space to a bi-dimensional one, we can more easily compute the distortion in a simulation environment. We represent the total distortion as a function of two parameters:

$$\begin{aligned} D(QP, R_{SA}) = D_K(R_K(QP)) + \beta D_{SA}(R_{SA}) \\ +(1-\beta) D_U(R_K(QP), R_Z(QP)) \end{aligned} \quad (3)$$

We show in Fig. 6 the distortion vs. the total bit rate per pixel $r = \frac{R}{2MN}$ for several QPs and for a given sequence ("poznan street"), where $M \times N$ is the spatial resolution. The same general behavior was observed on all the other test sequences. For a given total rate $R = R^*$ we should be able to select the QP corresponding to the lowest curve at this rate. This could be difficult in general, but we observe that for a given QP, each curve is very steep at the beginning and then becomes practically horizontal. This means that the SA coding of occlusion areas is efficient with a few bits, but once the occluded regions are "well coded", it is practically useless to increase their rate, and it is rather worth to use a higher QP, i.e. to "jump" on the next curve. Since a very small rate R_{SA} is sufficient to minimize the distortion for a given QP, it looks like we should always use the smallest QP possible, i.e. the smallest such that $R_K(QP)+R_Z(QP) \leq R$, and we have to give only the residual rate to the shape-adaptive coding of occlusion.

This approach implies the assumption that a QP corresponding to a given curve in Fig. 6 is optimal as soon as the total rate is larger than the minimum rate associated to the curve. However, if we look more closely to the figure, we understand that this is not entirely true. For example, the curve associated to QP=30 has a minimum rate of r_0=0.100 bpp, but it is optimal only when a total rate at least as large as r_1=0.109 is available. Therefore, we design an empirical rule to perform rate allocation: the KC must select the lowest QP such that the corresponding rate R_K+R_Z is smaller than $R-f(R)$. In turn, $f(R)$ is a part of the total rate reserved for the encoding of the occluded areas. For simplicity, we have considered $f(R) = \alpha R$ with $\alpha \in [0, 1]$; nevertheless, as shown in the experimental part, this allows RD performance very close to the upper bound given by the full-search approach. The parameter α is experimentally determined, as shown later.

In conclusion, the heuristic rate allocation algorithm consists of the following steps. The KC camera encodes its data at the lowest possible QP such that $R_K + R_Z \leq (1-\alpha)R$, *i.e.*

$$QP = \arg\min_q R_K(q)+R_Z(q) \quad \text{s.t.} R_K(q)+R_Z(q) \leq (1-\alpha)R \quad (4)$$

Then, the OC encodes its data using the residual available rate. A problem arises here: the OC should use a rate $R_{SA} = \alpha R + (R - R_K - R_Z)$, but it does not know the rate already allocated to the key camera $R_K + R_Z$. We refer to the case where the OC uses exactly this rate as *ideal allocation* (IA).

We consider three practical solutions to this problem. The simplest one is to use only $R_{SA} = \alpha R$. This is equivalent to optimizing the rate allocation for a rate constraint equal to $R_K + R_Z + \alpha R$. Then, for the resulting final rate the RD performance is the same as IA, but we are not able to perfectly select this final rate. A second solution consists in having the OC camera estimating the QP used by the KC, simply by performing a dummy Intra coding of T_O and Z_O: if the two views are similar enough the estimation would be often very close to the right value. We refer to this solution as local estimation (LE). A third possible solution is to implement a very light inter-camera communication such that the KC can tell the OC how much rate it has consumed. This would exactly implement the IA but would however deviate from the distributed paradigm. In summary, the first solution is simple, gives ideal RD performance but does not allow a perfect rate control. LE allows rate control and, as shown in the experimental section, has RD performance very close to the IA,

TABLE I
RELATIONSHIP BETWEEN THE QUANTIZATION PARAMETER FOR TEXTURE QP_T AND THE ONE FOR DEPTH QP_Z

QP_T	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25
QP_Z	51	50	50	50	50	49	48	47	47	46	45	45	44	44	43	43	42	42	41	41	40	39	38	37	36	35	34

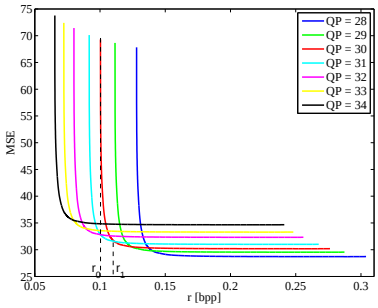


Fig. 6. The total distortion D vs. the bit rate per pixel $r = \frac{R}{2MN}$. For each different QP, we have a different color. This example is for a frame of the *poznan street* sequence

but it is more complex since the OC encoder must perform a dummy Intra coding. Finally the third solution implements the IA but demands a light communication between cameras. The choice among these solution is discussed in the experimental section.

This rate allocation algorithm can be easily extended to three cameras, where the central camera is a KC and the two lateral ones are OC. We will refer to this architecture in the next section as OKO. We observe that, for symmetry, the quality of synthesized regions at left and at right will be nearly the same. Then, we expect that the bit rate for coding the occluded areas for the two OCs would be nearly the same. Configuration tests with more than three cameras are less common in the context of 3D-TV [46] and will be studied as future works.

B. Quality of virtual viewpoints

In the scheme of Fig. 1, the depth of the OC Z_O is not explicitly encoded, and as shown in Fig. 5, it is obtained at the decoder by DIBR and inpainting. However, Z_O has a huge impact on the quality of the synthesized virtual viewpoints at the decoder side. In the previous discussion, the distortion of these viewpoints has not been considered, because otherwise the problem would not have been tractable. However here we want to investigate about the following issues: is it worth sending some more bits to improve the representation of Z_O? how much does it improve the virtual viewpoints quality? how can we allocate the total bit-rate considering the OC depth?

For simplicity, we assume that we have two cameras only and that the bit-rate allocation between R_K, R_D, R_{SA} has been performed according to the previous discussion. We call $R_0 = R_K + R_D + R_{SA}$ the rate allocated to T_K, Z_K and T_O. We can improve the representation of Z_O by encoding its occluded areas with the SA algorithm exactly as done for T_O. We have only to solve the rate allocation problem $R = R_0 + R_{Z_O}$ where R_{Z_O} is the rate for the SA encoding of the occluded areas in Z_O. The target will be to maximize the average PSNR of the original views and of the virtual views generated at the positions $\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$ of the baseline. The virtual views are obtained by interpolation algorithm of the view-synthesis reference software (VSRS), as for the 3D-HEVC Tests. The references for the virtual views are obtained by applying VSRS on the original texture and depth data (without compression), as in [47].

In order to explore the allocation problem between R_0 and R_{Z_O}, we perform a full search (FS) allocation for several combinations of values for the two parameters. The results for the *dancer* sequence are shown in Fig. 7. Each red curve corresponds to a fixed value of R_0 and we vary R_{Z_O}. As a consequence, the horizontal axis is the total bit-rate. The vertical axis is the average PSNR on virtual views. We also show the virtual viewpoint quality if the depth maps are not encoded but only inpainted as shown in Fig. 5. The optimal allocation corresponds to the upper envelope of the red curves.

Designing an algorithm for optimal rate allocation would be even more difficult than the previous case, given the complex, non-linear relationships between depth rate and virtual viewpoint quality. As a consequence, we propose a reasonable, heuristic allocation algorithm that has fairly good performance. Inspired by [46], we suppose that the ratio between occluded depth rate and occluded texture rate is the same as the ratio between key depth and key texture. However, since this rate may be very small if R_0 is close to $R_K + R_Z$, we add a small term assuring a suitable coding rate for the occluded depth. This term is a fraction of the rate R_0. In conclusion we assume:

$$R_{Z_O} = \frac{R_Z}{R_K} R_{SA} + \epsilon R_0 \tag{5}$$

The fraction ϵ of the total rate dedicated to the occluded depth is optimized by experiments, as shown in Section V-C. We show in Fig. 7 the results of the heuristic allocation: we observe that they are quite close to the optimal one, and better than those obtained by just inpainting the estimated depth map. Similar results were obtained on all the test sequences (see Section V-C).

V. EXPERIMENTAL RESULTS

In order to validate our proposed architecture and rate allocation method, we have performed tests on several MVD sequences (see Tab. II), using 3 views per sequence and 60 frames per view. The first 3 sequences (*mobile*, *dancer*, *GTFly*) are computer-generated and the depth data are perfect. For the others depth maps have been computed by a dense disparity estimation algorithm and so they are affected by errors. We use them since they are quite common in the literature, but we underline that the intended use case is the one where the depth maps are provided by range cameras [8], [48]–[51].

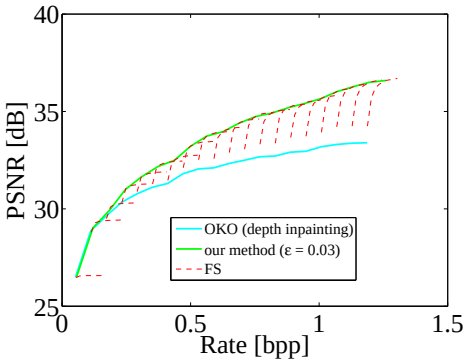


Fig. 7. PSNR of virtual frames for the *dancer* sequence. Red, dashed lines: full search (FS). Green line: the proposed heuristic allocation. Cyan: no depth map encoding.

TABLE II
THE MULTIVIEW PLUS DEPTH SEQUENCES TEST SET. SOURCES : A. PHILIPS; B. NOKIA; C. TANIMOTO LABORATORY; D. GWANGJU INSTITUTE OF SCIENCE AND TECHNOLOGY; E. POZNAN UNIVERSITY OF TECHNOLOGY [52]

Sequence	Resolution	views
mobile[a]	720×540	4, 5, 6
dancer[b]	1920×1088	1, 5, 9
GTFly[b]	1920×1088	1, 5, 9
balloons[c]	1024×768	0, 1, 2
kendo[c]	1024×768	1, 3, 5
newspaper[d]	1024×768	2, 4, 6
poznan street[e]	1920×1088	3, 4, 5

This section is organized as follows. In Sec. V-A we show how to tune the our parameters to obtain the best results. Then, in Sec. V-B we show the RD results on the original views, comparing the proposed architecture to many existing solutions. Finally, in Sec. V-C we provide the results for the virtual views.

A. Parameter tuning

We have to select the optimal values of the minimal region size p and of the structuring element radius ρ. They should be determined by considering their impact on the global RD performance, meaning that they should be computed jointly with the other system parameters. This would be very complex, therefore we perform a greedy optimization: each of these two parameters is varied while the other is kept fixed. We have found that with $p = 50$ and $\rho = 5$ the best performance can be obtained, and this result is independent from all the other system parameters with very good approximation. We will use these values for p and ρ in the rest of the paper.

Next, we have to set the parameter α introduced in Eq. 4 for the bit-rate allocation. Let P_{or} be the ideal PSNR that should be obtained if an oracle, full search rate allocation is performed, and let P_α be the obtained PSNR using the empirical rule in Eq. (4) for a given α. In Fig. 8 we compare P_{or} and P_α for different values of α for a given sequence. Finally, we have computed the average PSNR loss Δ_α as the mean value of the absolute difference between $P_{\text{or}}(\cdot)$ and $P_\alpha(\cdot)$:

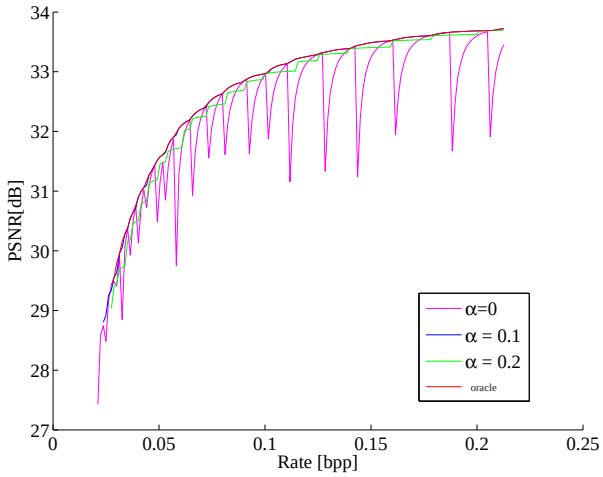


Fig. 8. PSNR [dB] vs. R for different values of α for the sequence *poznan street*.

$$\Delta_\alpha(P_{\text{or}}(\cdot), P_\alpha(\cdot)) = \frac{1}{R_{max} - R_{min}} \int_{R_{min}}^{R_{max}} (P_{\text{or}}(r) - P_\alpha(r))dr$$

where $[R_{min}, R_{max}]$ is the range of experiment rates. The best value of α is the one that minimizes Δ_α. We have computed the value of Δ_α as an average on 5 frames for 7 sequences, and the results are in Tab. III. The optimal value of α is 0.1, which corresponds to an average PSNR loss with respect to the full-search oracle case of just 0.04 dB. This is a remarkable result, since it means that our heuristic rule performs practically as well as the full search allocation. We could achieve even better results if we adapt the value of α to the sequence but, on one hand, this would give a very small gain even in the best case and on the other hand, we do not have for the moment any hint about how to select this parameter as a function of the sequence. In conclusion, using a fixed value of α gives excellent performances without having to adapt it to each sequence: in the following we only refer to the case $\alpha = 0.1$.

Finally, we compare the different methods for deciding the rate of the occlusion camera. The first and the third proposed solutions explained in Section IV have the same RD performance as IA, the only difference being that the first is perfectly distributed and the third allows rate control. The LE solution is distributed and allows rate control, but may introduce a small RD loss with respect to the IA. We compute this PSNR loss likewise the previous case of α: it is the mean value of the difference between $P_{\text{IA}}(\cdot)$ and $P_{\text{LE}}(\cdot)$. According to our experiments, the PSNR loss ranges between $3 \cdot 10^{-3}$ and $8 \cdot 10^{-5}$ dB and therefore is negligible. In conclusion, the LE method performs practically as well as the IA, but introduces some increase in complexity. One should choose among the three solutions according to the relative importance of the constraints on rate, complexity or distributedness of the system.

TABLE III
THE PNSR LOSS Δ_α [DB] WITH THE EMPIRICAL RULE 4 FOR DIFFERENT VALUES OF α.

Sequence	$\alpha = 0.000$	$\alpha = 0.050$	$\alpha = 0.075$	$\alpha = 0.100$	$\alpha = 0.125$	$\alpha = 0.200$
mobile	0.53	0.01	0.03	0.10	0.18	0.50
dancer	0.28	0.01	0.01	0.02	0.05	0.18
GTFly	0.22	0.01	0.01	0.03	0.07	0.20
balloons	0.61	0.09	0.04	0.01	0.01	0.04
kendo	1.13	0.17	0.07	0.01	0.01	0.05
newspaper	1.20	0.40	0.21	0.14	0.07	0.01
poznan street	0.36	0.03	0.01	0.01	0.01	0.10
average	0.61	0.10	0.05	**0.04**	0.06	0.15

B. Rate-distortion performance on original views

In this section, we evaluate the RD performance of the proposed architecture on the original views. The different methods that we consider for comparison can be classified into three families:

1) **Distributed methods (exploiting only inter-view correlation)**
   - **our proposed method (OKO)** - the central camera (texture + depth) is H.264/AVC INTRA coded and the two lateral ones are OC (see Fig. 9(b))
   - All INTRA - all the three cameras are Key cameras. This comparison is made because this configuration can be considered as a distributed one, since there is no communication between the cameras. For this case, cameras provide only texture images.
   - DISCOVER - the central camera is WZF coded (see Fig. 9(a)) in a classical DVC architecture: the SI for the central camera is obtained by the inter-view interpolation from external cameras, which in this setting are both K cameras [16]. A turbo encoder is used for Wyner-Ziv coding. This estimation is corrected by the turbo decoder through the parity bits sent by the WZ encoder. As for All INTRA, cameras provide only texture images. Other distributed codecs (such as VISNET II) have better performance than DISCOVER, so it cannot be considered as a state-of-the-art method; however it is a relevant benchmark in sights of its popularity and availability. In all the experiments we used our own DISCOVER implementation, that allows to manage large frame sizes and to change freely the SI generation technique. Since we use turbo codes as channel code, the modification is very small (we do not need the LDPCA matrices).
   - DISCOVERd-WKW - Here we test a different KF/WZF arrangement for DISCOVER. The central view (K) is INTRA coded (with its depth map) and the two WZFs are estimated by DIBR applied on the central camera. These two estimations are corrected by parity bits, as usual. In this scheme, the occluded zones on SI are inpainted, before turbo decoding.
2) **Simulcast methods (both distributed and not)**
   - DISCOVER simulcast - Each view is encoded separately by DISCOVER, by exploiting only temporal correlation. The GOP size is equal to 2.
   - H.264 simulcast - Each view is encoded independently by H.264/AVC, by exploiting only temporal correlation. The GOP size is supposed equal to 2 (IBIB).
   - HEVC simulcast - Each view is encoded independently by HEVC, by exploiting only temporal correlation. The GOP size is supposed equal to 2 (IBIB).
3) **Other methods**
   - LDVa - A state-of-the-art Layered Depth Video (LDV) (non-distributed) architecture [13] is implemented from the MVD data and the different layers are encoded by the aggregation method described in [53]. This architecture is very similar to our codec, but does not comply with the DVC paradigm: indeed, the inter-camera correlation is exploited at the encoder side. Another difference is the coding technique of the occluded areas: they are encoded by [53] and not by our shape adaptive coding techniques.
   - LDVo - A LDV (non-distributed) architecture is implemented and our shape adaptive coding is applied for the different regions. This architecture is the same as the one for LDVa: the difference is the coding technique for the occluded regions. In this case shape adaptive algorithm is applied for encoding the occluded areas.
   - MV-HEVC - the new model under test for Multi-View coding HEVC based in the version HTM 6.2, which is again not a DVC scheme, and is considered only for performance comparison purposes [9], being a future standard in multiview and MVD encoding. Inter-view residual prediction is used for exploiting correlation among the views. This is a not distributed architecture, because communication among the cameras is allowed. Moreover, the INTRA Frames are coded in HEVC INTRA mode.
   - DISCOVER fusion - WZFs and KFs are arranged in a quincunx scheme. Each WZF is estimated both by inter-view and temporal interpolation. Then, these estimates are fused by the algorithm proposed by [21]. The two side views only temporal interpolation is performed because inter-view interpolation is not possible.

We have considered two quality metrics, the PSNR and the SSIM [54], since the former is sometimes inconsistent with human perception. In particular, since we use DIBR for

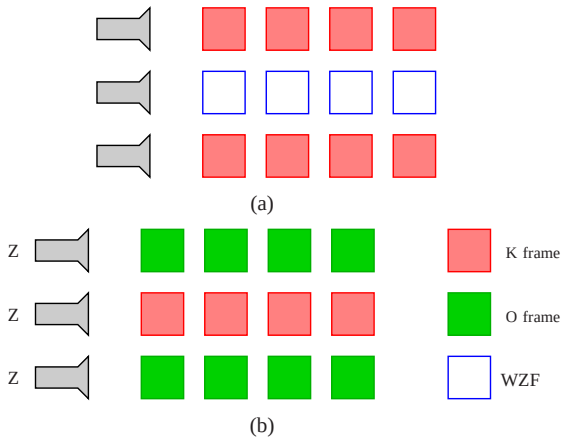


Fig. 9. Two DVC camera configurations: DISCOVER (a) and OKO (b)

synthesizing new views, a small misplacement of certain pixels may reduce significantly the PSNR of the synthesized frames, while their perceived quality is possibly not affected as much. In Fig. 10, 11 and 12, we show the Rate-PSNR and Rate-SSIM curves for the seven test sequences. The first figure refers to three computer-generated (CG) sequences (*mobile*, *dancer*, *GTFly*), while the other ones are natural sequences. The distributed method (exploiting only inter-view correlation) performances are indicated with a solid line, the simulcast methods have a dashed line, and the other methods a dash-dot line. Some results for Rate-PSNR improvement w.r.t. DISCOVER in terms of Bjontegaard metric [55] are also listed in Tab. IV[3].

Comparison with other distributed methods (exploiting only inter-view correlation).

From Fig. 10 we remark that for the CG sequences, the proposed method is the best within the distributed techniques exploited only inter-view correlation, except for relatively high-rates, where All-INTRA has slightly better performance. However, when we consider the SSIM, the proposed technique clearly outperforms all the competitors from the same family. As expected, the PSNR may penalize our technique since a small positioning error may result in a high MSE without necessarily a perceived quality loss. We also remark that the DISCOVER codec has always the worse performance both in PSNR and in SSIM. This is due to the lack of flexibility of the classical DVC methods, sending large chunks of parity bits to correct badly estimated side information images. In Fig. 11 and 12 we observe somewhat similar results for natural sequences: the proposed technique is better than DISCOVER in PSNR and SSIM, and almost always better than All-INTRA in SSIM, while for the PSNR it depends on the rate and on the sequence. We conclude that, as we may expected, our technique depends on the quality of the depth maps. When the depth maps are estimated (and then affected by errors), we observe a saturation effect in the PSNR (and partly in the SSIM). Our explanation is that errors in the depth may cause that some occluded areas are not correctly recognized as such: as a consequence, increasing the bit-rate has little effect on the quality of those areas. This performance saturation is an indirect consequence of giving up the feedback channel in our scheme: when the geometric information about the 3D scene is poor, we are not able to effectively correct the actual occluded areas. As a consequence, the proposed system is an effective alternative to classical schemes when the depth maps are not too poor. However, the steady improvement in both depth acquisition devices and in disparity estimation algorithms will make it reasonably easier to have high-quality depth maps for practical distributed systems. These results are summarized in Tab. IV, where we observe that our technique achieves an average bit rate reduction of 48.44% (and up to 66.94%) and an average PSNR improvement of 1.38 dB (and up to 3.50 dB) w.r.t. DISCOVER. In Tab. V, we have compared the gain in PSNR and in SSIM of our method w.r.t. All INTRA, when the bit rate is equal to 0.1 bpp. We remark that, although for some sequences we do not have a PSNR improvement, such as in the *newspaper* sequence (where we have a loss of 0.70 dB), in SSIM we have an improvement of 1.3%. Moreover, this gain in SSIM is nearly the same as that of the *GTFly* sequence where the corresponding PSNR improvement is of 0.85 dB. There does not exist a clear correspondence between the gain in PSNR and in SSIM.

[3]In order to compute the Bjontegaard metric, we have sampled the RD curves on four points

TABLE V
Δ_{PSNR} AND Δ_{SSIM} W.R.T. ALL INTRA AT A FIXED RATE OF 0.1 BIT PER PIXEL.

	Δ_{PSNR} [dB]	Δ_{SSIM}
mobile	4.08	0.091
dancer	0.93	0.062
balloons	-0.76	0.011
kendo	-2.63	-0.002
GTFly	0.85	0.012
newspaper	-0.70	0.013
poznan street	0.34	0.032

In order to improve the quality of the synthesized regions in our algorithm, we have also tried to send parity bits from a channel coder just as in classical DVC [5], [16]. However as shown in the previous results, using classical DVC on the MVD data is often even less efficient than INTRA coding: therefore, as expected, adding parity bits to our data does not improve the RD performance.

Finally, we spend a few words on a variant of DISCOVER, called DISCOVERd-WKW, where the depth and texture data for the central camera are INTRA coded and used for obtaining an estimation of the left and right views via DIBR extrapolation. Then, the occluded zones are filled by Bertalmio inpainting; the SI is finally corrected by the parity bits of the turbo encoder. We observe that the DIBR extrapolation of the DISCOVERd-WKW outperforms the extrapolation of DISCOVER if depth data are perfect (CG sequences). An opposite behavior is observed for natural sequences. However, both are always worse than the proposed technique.

Comparison with simulcast methods (both distributed and non-distributed)

Let us consider now the simulcast method performance in Fig. 10 to 12 (dashed lines). As expected, simulcast HEVC has the best performance (in PSNR and SSIM), but it is also by far the most complex technique. Simulcast H.264/AVC is a bit

TABLE IV
BJONTEGAARD METRIC FOR COMPARISON OF DIFFERENT TECHNIQUES W.R.T. DISCOVER.

	All INTRA		OKO		LDVa		LDVo	
	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]
mobile	1.31	-0.06	-66.94	3.50	-64.65	3.45	-79.45	3.92
dancer	-29.03	0.96	-57.50	1.58	-55.60	1.56	-81.49	1.99
balloons	-32.97	2.21	-22.48	-0.09	-17.10	0.07	-34.78	0.47
kendo	-30.98	2.10	-29.83	0.40	-27.63	0.41	-26.80	0.98
GTFly	-25.34	1.07	-65.47	1.96	-67.10	2.08	-81.46	2.44
newspaper	-68.56	4.06	-51.50	1.40	-40.58	1.28	-65.85	1.92
poznan street	-20.63	0.94	-45.39	0.93	-45.10	0.89	-48.83	1.21
Mean	-29.45	1.61	-48.44	1.38	-45.39	1.39	-59.80	1.84

worse and, for the computer-generated sequences, has a worst SSIM than the proposed method. This is a good result for our distributed technique. Our interpretation is that when inter-view correlation is correctly exploited (e.g. because the depth maps are not affected by errors), it may give better results than exploiting temporal correlation. A similar result is observed concerning simulcast DISCOVER: our technique has better SSIM (but worse PSNR) on CG images, while on two natural sequences out of four DISCOVER is better. However we remark that the most relevant comparison is the one between our method and the first family (distributed methods exploiting only inter-view correlation)), since in our algorithm, we do not take into account the temporal correlation. The goal of this paper is to exploit geometrical information to get compression. Then, as future work, we can integrate a technique that exploits temporal correlation. In this case, we can expect to improve DISCOVER simulcast in all the configurations.

Comparison with non distributed methods

For completeness, we have performed also a comparison with non distributed methods that allow a communication among the cameras. We can remark that our method performs also better than LDVa, partly thanks to the more effective coding of the occluded areas. We have compared our algorithm also with a variant of LDV where occluded areas are encoded with our SA algorithm (LDVo), and with the reference software of the upcoming MV-HEVC standard, the HEVC-based HTM 6.2 software [9]. Since in all the other schemes we do not exploit temporal correlation, we only use inter-view and inter-component prediction tools in the HTM. This method has the best performance for CG sequences, as we can expect, while HEVC Simulcast is better for three out of four natural sequences. Finally, we have also performed comparisons w.r.t. multiview DISCOVER codec with fusion of SI proposed in [21]: in a quincunx scheme, each WZF can be estimated both by temporal and inter-view interpolation. Then, these two estimates are fused by the algorithm proposed by [21]. We can observe that this method always outperforms DISCOVER simulcast, since the temporal estimation is enhanced by the inter-view one. We can remark that, even if our technique (OKO) only exploits inter-view correlation, sometimes we are able to outperform DISCOVER fusion method, namely for all the CG sequences.

Complexity issues: The encoder of our architecture performs a double DIBR, a Bertalmio inpainting, a mask processing and shape-adaptive encoding. DIBR has a complexity close to motion compensation (i.e. much less than motion estimation). Given the depth map, we obtain the disparity map. Then, from intrinsic and extrinsic camera parameters, each pixel of the reference frame is mapped in another pixel of the synthesized frame. A fixed amount of operations per pixel is needed. Mask extraction is intrinsic in DIBR then no additional computations have to be performed. The morphological operations are not computationally intensive. Fast algorithms exist with a complexity of a few (in the range 1.5 to 3) comparisons per pixel [56]. As for Bertalmio inpainting, since there is not template matching (such as in Criminisi), the complexity is very low. Finally, shape adaptive (SA) coding consists in a shape adaptive transform (whose complexity is basically the same as for an ordinary transform) and in SPIHT-like bit plane coding (that is extremely simple). As a consequence, SA coding has a per-pixel complexity comparable to INTRA, but since only a small part of the image is encoded with this technique, its impact on the total complexity is reduced as such. In summary, all the elements of the OC encoder have a complexity that is $O(N)$, where N is the number of occluded pixels in the image. Moreover, there is no time-consuming matching operation, in opposition to the case of motion estimation. Moreover, also at the decoder side the complexity of our algorithm is much lower than the DISCOVER one. The complexity of DIBR projection is negligible w.r.t. DISCOVER interpolation, that needs a full search algorithm of block matching. In our architecture, we also suppress the iterative channel decoding, which is responsible for the high complexity of the current DVC decoders (more complex even than the motion estimation). Moreover, the feedback channel is also eliminated, making the whole architecture more attractive for implementation in practical systems.

C. Rate-Distortion performance on virtual views

In this section, we discuss the bit rate allocation method for depth map coding of OC in order to maximize the PSNR on virtual views. At first, we have to select the best value of ϵ in Eq. (5). We have performed a full search for the value of R_{Z_0} maximizing the PSNR on the synthesized views, by varying the value of ϵ. We have averaged the results on 10 frames for each of 7 sequences for the virtual views and we have found that the bast value of ϵ is equal to 0.03. As shown in Fig. 7 for *dancer* sequence, this choice may achieve a performance very close to the optimum. Similar results have been obtained for other sequences.

After having tuned the rate allocation, we turn our attention to RD performance. We consider a configuration with three

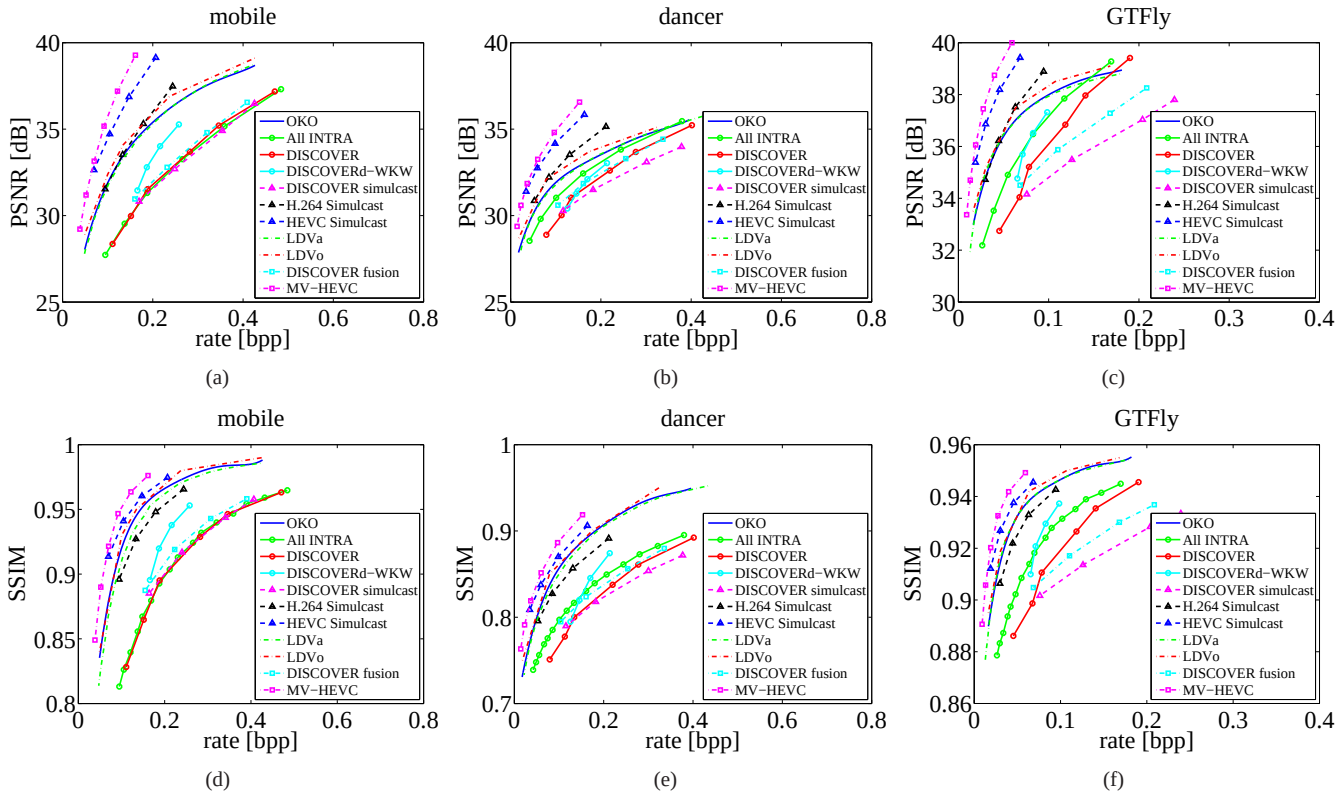


Fig. 10. Rate-PSNR and Rate-SSIM performance for *mobile*, *dancer* and *GTFly* sequence

range cameras. Related to the baseline, we refer to the left camera position as -1, to the central one's as 0 and the right one's as 1. Then, we consider the virtual viewpoints in $[\pm\frac{1}{4}, \pm\frac{1}{2}, \pm\frac{3}{4}]$ by the VSRS software version 6.2.

We have compared the following methods (also listed in Tab. VI):

- All INTRA(d) - All Frames are INTRA coded (both texture and depth independently).
- OKOi - We use our method (OKO) for coding the three original views and depth maps for O cameras are filled by Bertalmio inpainting.
- OKOalloc - We use our method (OKO) and we send depth maps occluded areas at the bit rate given by Eq. (5).
- DISCOVERd-WKW - it is the same as in the previous section (but the depth maps are sent for all cameras): SI for WZFs is obtained by extrapolation on the central KF.
- DISCOVER-V - Since the results of the previous scheme are not satisfactory, we have introduced a new DISCOVER-based codec, where all the depth maps of the three views are INTRA coded. The left and the right texture views are INTRA coded (as in DISCOVER). Texture of central camera is Wyner-Ziv coded. At the decoder side, the estimation of this central WZF is obtained by performing VSRS interpolation algorithm on the two INTRA coded views. As usual, this SI is corrected by parity bits.
- 3D-HEVC - the new model under test for Multi-View coding HEVC based in the version HTM 6.2, which is not a distributed scheme, and is considered only for performance benchmarking purposes. Inter-view residual prediction is used for exploiting correlation among the views. Depth INTRA mode (wedgelets) and view synthesis optimization are used [9].

For comparison, we have evaluated the average PSNR for the virtual views vs. the overall bit rate. We evaluated the Bjontegaard metric w.r.t. All INTRA (d). The results are in Tab. VII. The complete Rate-PSNR and Rate-SSIM curves for each sequence are in Fig. 13, 14 and 15. These results are very important since they confirm the superiority of the proposed approach with respect to comparable techniques. OKOalloc is largely better than comparable distributed methods (for PSNR and SSIM) and than the All-Intra approach (for SSIM and almost always for PSNR). Indeed, with respect to All INTRA, we are able to obtain an average bit rate reduction up to 38.77%. OKOalloc is on the average better than All-Intra, even though for natural sequences it may have a smaller PSNR. As in the case of the original views, also for virtual views the performance are affected by the quality of the depth maps, visible as a saturation of the PSNR for natural sequences in Fig. 14(a-c) and 15(a). However we remark that this phenomenon affects much less the SSIM, since the latter is less sensitive to small errors in objects' position. As for the other methods, we remark also that DISCOVER-V performs better than DISCOVERd-WKW. Indeed, even thought in DISCOVER-V we have two Key cameras instead of one (as in DISCOVERd-WKW), interpolation performs much better than extrapolation. Moreover, correcting depth maps by parity bits increases significantly the total bit rate, because

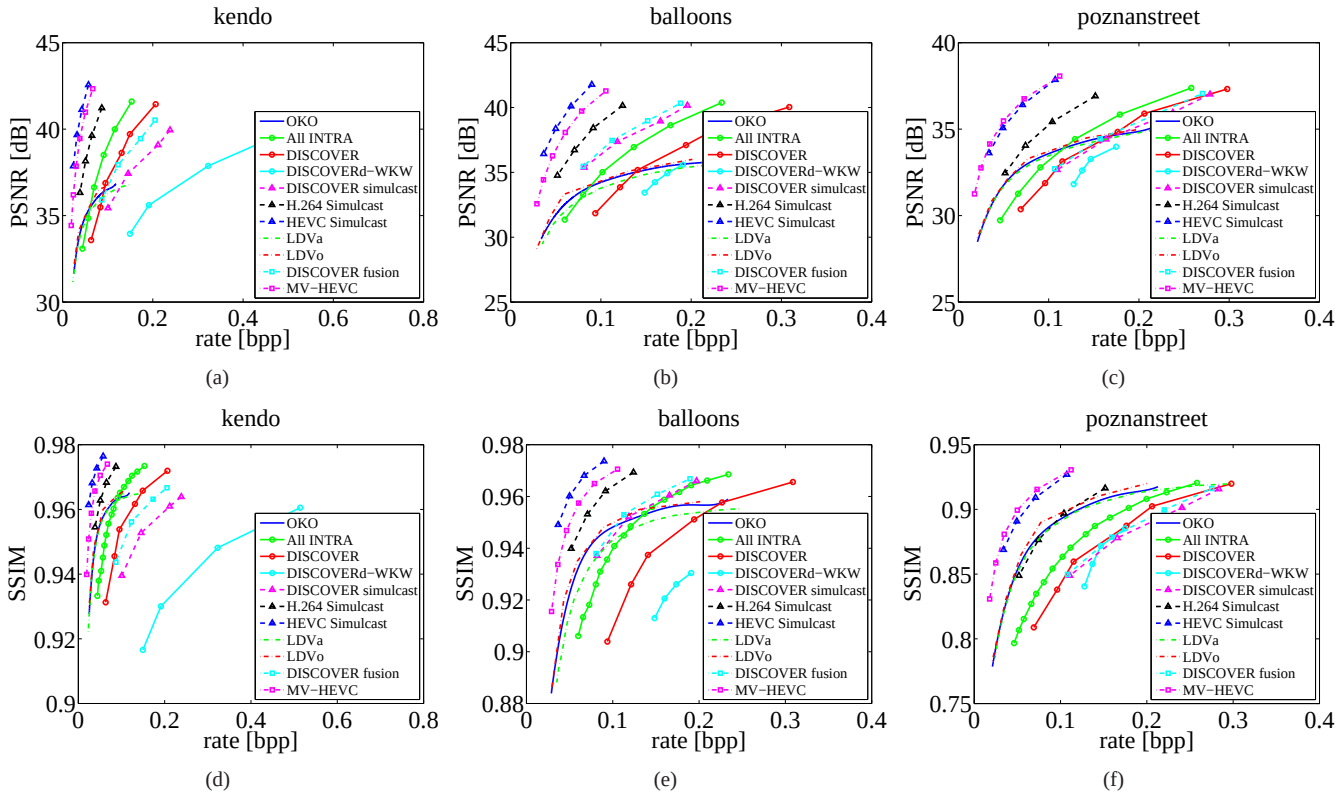


Fig. 11. Rate-PSNR and Rate-SSIM performance for *kendo*, *balloons* and *poznan street* sequence

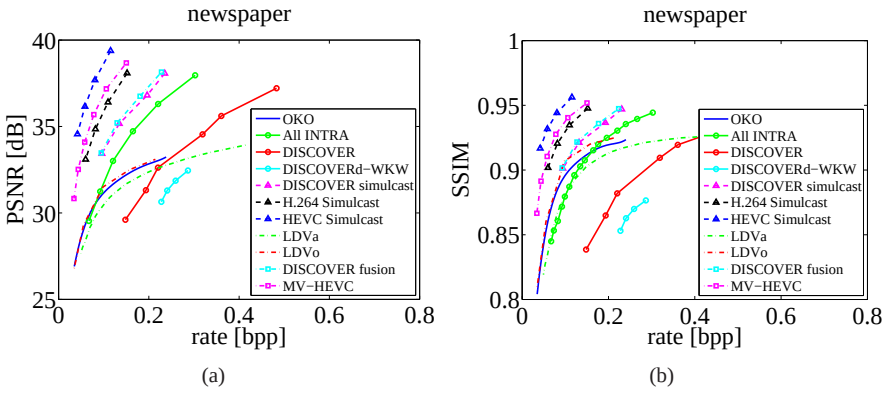


Fig. 12. Rate-PSNR and Rate-SSIM performance for *newspaper* sequence

depth data are not suitable for turbo encoding and decoding. Indeed, DVC turbo decoding would maximize the PSNR on the depth data, but we are interested in maximizing the PSNR on the virtual views. The proposed techniques are always better than other DISCOVER-based systems in terms of SSIM and very often in terms of PSNR. As for real views, we observe a saturation effect for large bit rate, due to artifacts and error on depth maps (in particular, for the sequences where the depth map is estimated) . Finally, as expected, 3D-HEVC has the best performance, but for the CG sequences our techniques have SSIM scores closer to 3D-HEVC than to distributed competitors.

VI. Conclusion and Future Works

In this paper, we have proposed a distributed architecture for the multiview video plus depth format. Our system consists in encoding one view, called Key camera, in the INTRA mode and sending only occluded areas of the other one, called O camera, which are obtained via a double DIBR. Differently from classical DVC architectures, the proposed system is not based on channel coding. Thus, it is not necessary to model the statistical properties of the error at the encoder side, as well as the feedback channel (typical of most DVC systems) which can be removed. Then, a rate allocation method between KC and OC has been proposed: we have found that the choice of the QP of the KC influences significantly the performance of the whole system, because the quality of OC depends strictly

TABLE VI
THE DIFFERENT TECHNIQUES USED IN OUR TESTS FOR VIRTUAL VIEWS. SA: SHAPE-ADAPTIVE; I: H.264 INTRA CODING; WZ: WYNER-ZIV; I-HEVC: HEVC INTRA CODING; P: PREDICTIVE INTER-VIEW CODING.

technique	Texture coding			Depth coding			Remarks
view	left	central	right	left	central	right	
OKO i	SA	I	SA	·	I	·	Depth maps for O cameras are inpainted
OKO alloc	SA	I	SA	SA	I	SA	
All INTRA (d)	I	I	I	I	I	I	
DISCOVERd-WKW	WZ	I	WZ	WZ	I	WZ	SI for WZF is obtained by extrapolation
DISCOVER-V	I	WZ	I	I	I	I	SI for WZF is obtained by interpolation
3D-HEVC	P	I-HEVC	P	P	I-HEVC	P	

TABLE VII
Δ_R AND Δ_{PSNR} W.R.T. ALL-INTRAD FOR SYNTHESIZED VIEWS (PSNR IS COMPUTED EXCLUSIVELY ON VIRTUAL VIEWS)

	OKOalloc		OKO-I		DISCOVERd-WKW		DISCOVER-V	
	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]
mobile	-69.36	4.26	-41.73	2.37	29.87	0.05	-18.87	1.24
dancer	-23.65	0.06	6.87	-0.34	80.22	-0.34	-11.65	0.34
balloons	-27.62	0.01	4.29	-0.07	114.11	-1.37	2.38	-0.21
kendo	-34.36	-0.02	-35.84	1.69	167.15	-2.84	9.34	-0.74
GTFly	-47.04	0.01	-34.92	2.02	81.23	0.21	0.11	0.02
newspaper	-13.95	-0.53	-4.84	-1.22	139.11	-2.32	9.95	-0.57
poznan street	-55.42	1.59	-59.39	1.70	81.56	-0.07	-6.11	0.30
mean	-38.77	1.44	-23.65	0.89	99.03	-0.96	-2.12	0.05

on the bit rate that we have used for encoding the Key camera. With our method we are able to obtain a bit rate reduction w.r.t. DISCOVER up to 67%. We have also tested our algorithm in terms of SSIM. This measure seems to be more suitable w.r.t. PSNR for evaluating methods using DIBR. Unfortunately, for actual cameras (where depth maps are not directly acquired), for high bit rate there is a saturation in terms of PSNR, because of the limited quality of synthesized areas, due to errors and artefacts on depth maps. This aspect is mitigated for synthetic computer-generated sequences. We have also explored a rate allocation method for occluded areas of OC depth maps in order to maximize the quality of syntactic views. With our method, we significantly outperform state-of-the-art algorithm. As future work, we plan to extend our rate allocation method for a generic number of cameras and we can exploit also the temporal correlation of the occluded regions within the OC sequence.

REFERENCES

[1] M. Tanimoto, M. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 67 –76, Jan. 2011.

[2] S. Yea and A. Vetro, "View synthesis prediction for multiview video coding," *Signal Processing: Image Communication*, vol. 24, pp. 89–100, 2009.

[3] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, pp. 471–480, Jul. 1973.

[4] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–11, Jan. 1976.

[5] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Asilomar Conference on Signals and Systems*, Pacific Grove, California, Nov. 2002.

[6] X. Artigas, E. Angeli, and L. Torres, "Side information generation for multiview distributed video coding using a fusion approach," in *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, june 2006, pp. 250 –253.

[7] C. Brites, J. Ascenso, and F. Pereira, "Side information creation for efficient Wyner-Ziv video coding: Classifying and reviewing," *Signal Processing: Image Communication*, vol. 28, pp. 689–726, 2013.

[8] G. Alenya and C. Torras, "Lock-in time-of-flight (ToF) cameras: A survey," *IEEE Sensors Journal*, vol. 11, pp. 1917–1926, 2011.

[9] [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HT 6.2

[10] M. Salmistraro, M. Zamarin, L. L. Rakêt, and S. Forchhammer, "Distributed multi-hypothesis coding of depth maps using texture motion information and optical flow," in *IEEE Intern. Conf. on Acoustics, Speech and Signal Proc.*, Vancouver, Canada, Sep. 2013.

[11] C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR)," in *Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, Benalmádena, Spain, Sep. 2003, pp. 482–487.

[12] W. Daio, G. Cheung, N.-M. Cheung, A. Ortega, and O. C. Auo, "Rate-distortion optimized merge frame using piecewise constant functions," in *Proc. of IEEE Int. Conf. Image Proc.* Melbourne, Australia: IEEE, 2013.

[13] K. Muller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, Cairns, Australia, oct. 2008, pp. 34 –39.

[14] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann, "Distributed monoview and multiview video coding: Basics, problems and recent advances," *IEEE Signal Proc. Mag.*, pp. 67–76, Sep. 2007.

[15] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: a video coding paradigm with motion estimation at the decoder," *IEEE Trans. Image Proc.*, vol. 16, no. 10, pp. 2436–2448, Oct. 2007.

[16] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The Discover codec: Architecture, techniques and evaluation," in *Proc. of Pict. Cod. Symp.*, Lisbon, Portugal, Nov. 2007.

[17] J. Ascenso, C. Brites, F. Dufaux, A. Fernando, T. Ebrahimi, F. Pereira, and S. Tubaro, "The VISNET ii DVC codec: architecture, tools and performance," in *Proc. of Europ. Sign. Proc. Conf.*, 2010.

[18] J. Areia, J. Ascenso, C. Brites, and F. Pereira, "Wyner-Ziv stereo video coding using a side information fusion approach," in *Proc. of IEEE Workshop on Multimedia Sign. Proc.*, Chania, Crete, Greece, oct 2007, pp. 453 –456.

[19] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-based multiview distributed video coding," in *ACM Intern. Workshop on Video surveillance and sensor networks*, Santa Barbara, CA, USA, 2006, pp. 139–144.

[20] P. Ferre, D. Agrafiotis, and D. Bull, "Fusion methods for side information generation in multi-view distributed video coding systems," in *Proc. of IEEE Int. Conf. Image Proc.*, vol. 6, San Antonio, Texas, 2007, pp. VI –409 –VI –412.

[21] T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Fusion schemes for multiview distributed video coding," in *Proc. of Europ. Sign. Proc. Conf.*, Glasgow, Scotland, 2009.

[22] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "Novel solutions for side information generation and fusion in multiview DVC," *EURASIP J. Applied Signal Proc.*, vol. 2013, no. 154, p. 17, Oct. 2013.

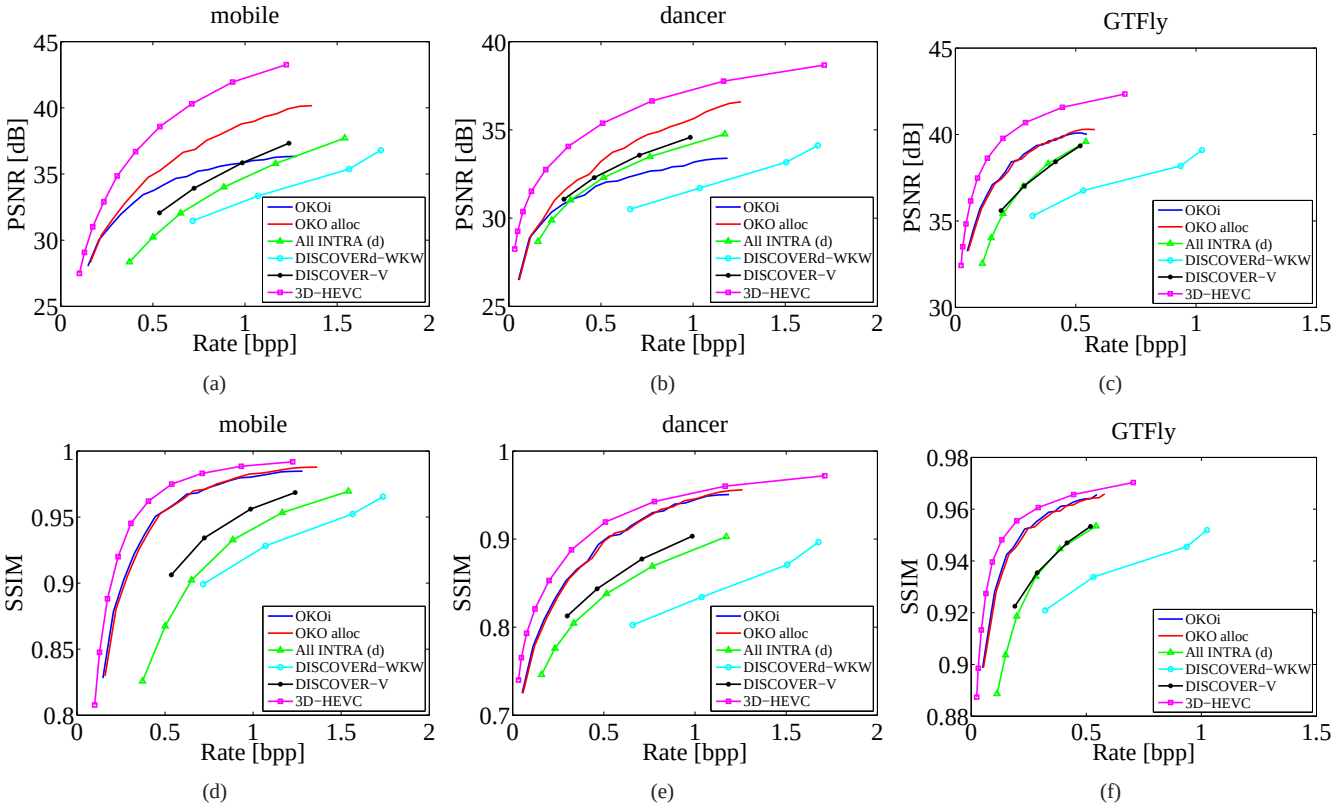


Fig. 13. Rate-PSNR and Rate-SSIM performance for *mobile*, *dancer*, *GTFly* sequence for virtual views

[23] H. Shum and S. Kang, "A review of image-based rendering techniques," *Proc. of SPIE Int. Symp. Visual Comm. and Image Proc.*, vol. 213, 2000.

[24] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," in *Proc. of SPIE Int. Symp. Visual Comm. and Image Proc.*, vol. 6077, 2006, pp. 290–297.

[25] T. Maugey, J. Gauthier, M. Cagnazzo, and B. Pesquet-Popescu, "Evaluation of side information effectiveness in distributed video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2116 – 2126, Dec. 2013.

[26] T. Maugey and B. Pesquet-Popescu, "Side information estimation and new symmetric schemes for multi-view distributed video coding," *Journal of Visual Communication and Image Representation*, vol. 19, no. 8, pp. 589–599, 2008.

[27] O. Mourad, F. Dufaux, E. Touradj *et al.*, "Iterative multiview side information for enhanced reconstruction in distributed video coding," *EURASIP Journal on Image and Video Processing*, 2009.

[28] M. Salmistraro, L. L. Rakêt, M. Zamarin, A. Ukhanova, and S. Forchhammer, "Texture side information generation for distributed coding of video-plus-depth," in *Proc. of IEEE Int. Conf. Image Proc.*, Melbourne, Australia, Sep. 2013.

[29] T. Maugey, J. Gauthier, B. Pesquet-Popescu, and C. Guillemot, "Using an exponential power model for Wyner-Ziv video coding," in *IEEE Intern. Conf. on Acoustics Speech and Signal Process.*, Dallas, Texas, Mar. 2010, pp. 2338–2341.

[30] D. J. Louw and H. Kaneko, "Suppressing feedback in a distributed video coding system by employing real field codes," *EURASIP J. Applied Signal Proc.*, vol. 2013, no. 1, pp. 1–19, 2013.

[31] F. Verbist, N. Deligiannis, S. M. Satti, P. Schelkens, and A. Munteanu, "Encoder-driven rate control and mode decision for distributed video coding," *EURASIP J. Applied Signal Proc.*, vol. 2013, no. 1, p. 156, 2013.

[32] J. L. Martinez, G. Fernandez-Escribano, H. Kalva, W. R. J. Weerakkody, W. A. C. Fernando, and A. Garrido, "Feedback free DVC architecture using machine learning," in *Proc. of IEEE Int. Conf. Image Proc.*, 2008, pp. 1140–1143.

[33] M. Bertalmio, A. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 355–362.

[34] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *Proc. of IEEE Workshop on Multimedia Sign. Proc.*, Saint Malo, France, Oct. 2010, pp. 167–170.

[35] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Distance dependent depth filtering in 3d warping for 3dtv," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, Chania, Crete, Greece, oct. 2007, pp. 312 –315.

[36] M. Cagnazzo, G. Poggi, L. Verdoliva, and A. Zinicola, "Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT," in *Proc. of IEEE Int. Conf. Image Proc.*, Singapore, Oct. 2004, pp. 2459–2462.

[37] M. Cagnazzo, G. Poggi, and L. Verdoliva, "Region-based transform coding of multispectral images," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2916–2926, Dec. 2007.

[38] S. Li and W. Li, "Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 5, pp. 725 –743, aug 2000.

[39] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, 1996.

[40] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *Image Pro., IEEE Trans.*, vol. 13, no. 9, pp. 1200 –1212, Sep. 2004.

[41] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-d video," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 453–465, 2011.

[42] G. Cheung, V. Velisavljevic, and A. Ortega, "On dependent bit allocation for multiview image coding with depth-image-based rendering," *IEEE Trans. Image Proc.*, vol. 20, no. 11, pp. 3179–3194, 2011.

[43] H. Yuan, Y. Chang, J. Huo, F. Yang, and Z. Lu, "Model-based joint bit allocation between texture videos and depth maps for 3-d video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 485–497, 2011.

[44] E. Bosc, V. Jantet, M. Pressigout, L. Morin, and C. Guillemot, "Bit-rate allocation for multi-view video plus depth," in *3DTV Conference:*

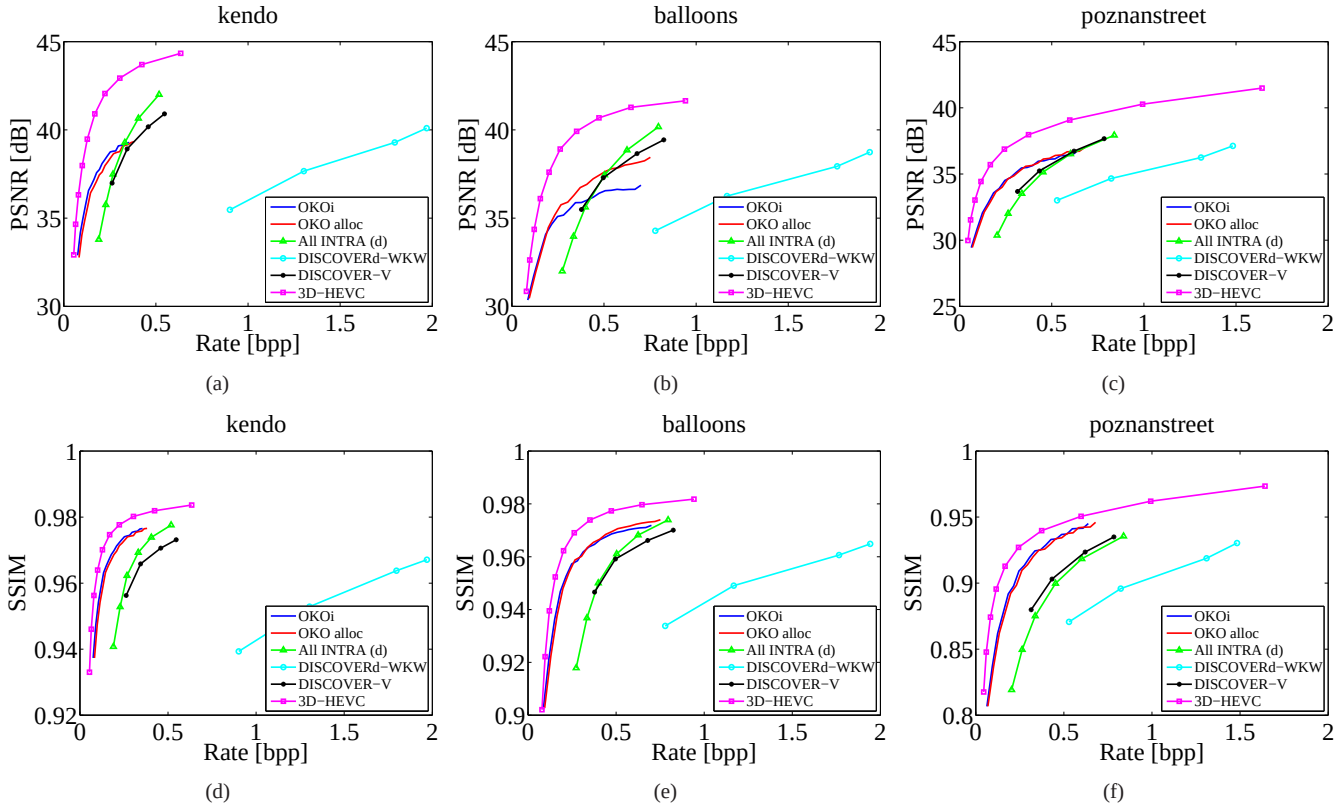


Fig. 14. Rate-PSNR and Rate-SSIM performance for *kendo*, *balloons*, *poznan street* sequence for virtual views

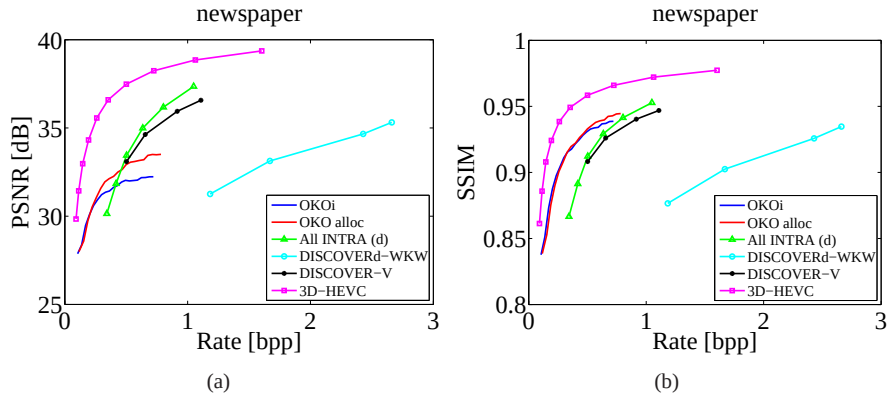


Fig. 15. Rate-PSNR and Rate-SSIM performance for *newspaper* sequence for virtual views

The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011.

[45] F. Shao, G. Jiang, M. Yu, K. Chen, and Y.-S. Ho, "Asymmetric coding of multi-view video plus depth based 3-D video for view rendering," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 157–167, 2012.

[46] D. Rusanovsky, K. Muller, and A. Vetro, "Common test conditions of 3DV core experiments," January 2013, ITU-T SG 16 WP 3 & ISO/IEC JTC1/SC29/WG11 JCT3V-C1100.

[47] H. Schwarz, C. Bartnik, S. Bosse, *et al.*, "3d video coding using advanced prediction, depth modeling, and encoder control methods," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012.

[48] T. Koninckx and L. Van Gool, "Real-time range acquisition by adaptive structured light," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 432 –445, march 2006.

[49] S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras: Principles, Methods and Applications*. Springer Briefs in Computer Science, 2012.

[50] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight sensors in computer graphics," in *Proc. Eurographics (State-of-the-Art Report)*, 2009, pp. 119–134.

[51] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 755–762.

[52] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan multiview video test sequences and camera parameters," *ISO/IEC JTC1/SC29/WG11 MPEG*, p. M17050, 2009.

[53] S.-U. Yoon and Y.-S. Ho, "Multiple color and depth video coding using a hierarchical representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1450–1460, 2007.

[54] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Proc.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[55] G. Bjontegaard, "Calculation of average PSNR differences between RD-

curves," in *VCEG Meeting*, Austin, USA, Apr. 2001.
[56] M. Van Droogenbroeck and M. Buckley, "Morphological erosions and openings: fast algorithms based on anchors," *Journal of Mathematical Imaging and Vision*, vol. 22, no. 2-3, pp. 121–142, 2005.

Giovanni Petrazzuoli Giovanni Petrazzuoli (S'09, M'13) obtained the Laurea Specialistica (equivalent to the M.S.) in Telecommunication Engineering from Federico II University (Naples, Italy), in November 2009. He obtained the PhD degree from Télécom ParisTech in January 2013, with a thesis on distributed video coding for multi-view and multi-view plus depth video. Since February 2013, he is postdoctoral researcher within the Image and Signal Processing Department at Télécom ParisTech (Paris, France). His research interests also cover interactive streaming, shape adaptive coding and super-resolution.

Thomas Maugey Thomas Maugey (S'09, M'11) graduated from Ecole Supérieure d'Electricité, Supélec, Gif-sur-Yvette, France in 2007. He received the M.Sc. degree in fundamental and applied mathematics from Supélec and Université Paul Verlaine, Metz, France, in 2007. He received his Ph.D. degree in Image and Signal Processing at Télécom ParisTech, Paris, France in 2010. Since October 2010, he is a postdoctoral researcher at the Signal Processing Laboratory (LTS4) of Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. His research interests include monoview and multiview distributed video coding, 3D video communication, data representation, video compression, network coding and view synthesis.

Marco Cagnazzo (M'05-SM'11) obtained the Laurea (equivalent to the M.S.) degree in Telecommunication Engineering from Federico II University, Napoli, Italy, in 2002, and the Ph.D. degree in Information and Communication Technology from Federico II University and the University of Nice-Sophia Antipolis, Nice, France in 2005.

He was a post-doc fellow at I3S Laboratory (Sophia Antipolis, France) from 2006 to 2008. Since February 2008 he has been Associate Professor at Institut Mines-Télécom, Télécom ParisTech (Paris), within the Multimedia team. He is author of more than 90 contributions in peer-reviewed journals, conferences proceedings, books and book chapters. His current research interests are three-dimensional video communication and coding, distribued video coding, robust video delivery, network coding.

Dr. Cagnazzo is an Area Editor for *Elsevier Signal Processing: Image Communication* and *Elsevier Signal Processing*. Moreover he is a reviewer for major international scientific reviews (IEEE TRANS. MULTIMEDIA, IEEE TRANS. IMAGE PROCESSING, IEEE TRANS. SIGNAL PROCESSING, IEEE TRANS. CIRC. SYST. VIDEO TECH., *Elsevier Signal Processing*, *Elsevier Sig. Proc. Image Comm.*, and others) and conferences (IEEE International Conference on Image Processing, IEEE MMSP, European Signal Processing Conference, and others).

BÃ©atrice Pesquet-Popescu BÃ©atrice Pesquet-Popescu (SM'06, F'13) received the engineering degree in telecommunications from the "Politehnica" Institute in Bucharest in 1995 (highest honours) and the Ph.D. degree from the Ecole Normale Supérieure de Cachan in 1998. In 1998, she was a Research and Teaching Assistant with Université Paris XI, Paris. In 1999, she joined Philips Research France, Suresnes, France, where she worked for two years as a Research Scientist, then as a Project Leader, in scalable video coding. Since Oct. 2000 she is with Télécom ParisTech (formerly, ENST), first as an Associate Professor, and since 2007 as a Professor, Head of the Multimedia Group. She is the Head of the UBIMEDIA common research laboratory between Alcatel-Lucent and Institut Télécom. Her current research interests are in source coding, scalable, robust and distributed video compression and sparse representations. Dr. Pesquet-Popescu was an EURASIP BoG member (2003-2010), and an IEEE Signal Processing Society IVMSP TC member and MMSP TC associate member. She serves as an Associate Editor for IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on CSVT, Elsevier Image Communication, and Hindawi Int. J. Digital Multimedia Broadcasting journals and was till 2010 an Associate Editor for Elsevier Signal Processing. She was a Technical Co-Chair for the PCS2004 conference, and General Co-Chair for IEEE SPS MMSP2010, EUSIPCO 2012, and IEEE SPS ICIP 2014 conferences. Beatrice Pesquet-Popescu is a recipient of the "Best Student Paper Award" in the IEEE Signal Processing Workshop on Higher-Order Statistics in 1997, of the Bronze Inventor Medal from Philips Research and in 1998 she received a "Young Investigator Award" granted by the French Physical Society. She holds 23 patents in wavelet-based video coding and has authored more than 290 book chapters, journal and conference papers in the field. In 2006, she was the recipient, together with D. Turaga and M. van der Schaar, of the IEEE Trans. on Circuits and Systems for Video Technology "Best Paper Award".