Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2010, Article ID 425641, 16 pages
doi:10.1155/2010/425641

Research Article

A Scalable Multiple Description Scheme for 3D Video Coding Based on the Interlayer Prediction Structure

Lorenzo Favalli and Marco Folli

Department of Electronics, University of Pavia, Via Ferrata 1, 27100 Pavia, Italy

Correspondence should be addressed to Lorenzo Favalli, lorenzo.favalli@unipv.it

Received 2 May 2009; Revised 8 September 2009; Accepted 11 December 2009

Academic Editor: Pietro Zanuttigh

The most recent literature indicates multiple description coding (MDC) as a promising coding approach to handle the problem of video transmission over unreliable networks with different quality and bandwidth constraints. Furthermore, following recent commercial availability of autostereoscopic 3D displays that allow 3D visual data to be viewed without the use of special headgear or glasses, it is anticipated that the applications of 3D video will increase rapidly in the near future. Moving from the concept of spatial MDC, in this paper we introduce some efficient algorithms to obtain 3D substreams that also exploit some form of scalability. These algorithms are then applied to both coded stereo sequences and to depth image-based rendering (DIBR). In these algorithms, we first generate four 3D subsequences by subsampling, and then two of these subsequences are jointly used to form each of the two descriptions. For each description, one of the original subsequences is predicted from the other one via some scalable algorithms, focusing on the inter layer prediction scheme. The proposed algorithms can be implemented as pre- and postprocessing of the standard H.264/SVC coder that remains fully compatible with any standard coder. The experimental results presented show that these algorithms provide excellent results.

1. Introduction

The research on stereoscopic video has received high interest over the past decade in order to provide viewers with more realistic vision than traditional 2D video. Stereoscopic video coding has been traditionally performed using two different views which are then merged to create a full set of different possible views. Other than using this left and right views, a more recent technique, known as depth image-based rendering (DIBR) [1], represents 3D video based on a monoscopic video and associated per-pixel depth information, simply called *color image* (or *video*) and *depth map,* respectively. While the color consists of three components, Y, U, and V as in the traditional video applications, the depth map video uses only one component to store the depth information of objects within the scene related to the camera position. The advantage of such a scheme is that it can capture the stereoscopic sequences more easily compared to the traditional left and right views techniques.

Even though 3D video can provide a more impressive experience than conventional 2D video, in the past, many factors such as huge bandwidth requirements and the discomfort due to special devices needed to observe 3D effect prevented widespread diffusion of 3D video in commercial services. Following the path of "traditional" video, it is easily foreseeable that, thanks to the improvements in coding techniques [2], also 3D applications will soon be available over both the Internet and wireless networks.

Reliable video transmission over unreliable or inefficient networks such as the Internet or wireless networks poses many challenges related to bandwidth variations and packet losses due to congestion on one side and to fading, interference, and mobility on the other one [3]. Traditionally, to cope with network and device heterogeneity, scalability techniques have been proposed.

A scalable video sequence is composed of a so-called *base*-layer and of one (or more) *enhancement*-layer(s): compared to a single-layer sequence, the base-layer is self-contained and fully decodable to a signal of lower quality and/or

lower resolution in terms of pixel or time. Enhancement layers, on the contrary, cannot be decoded if the base layer is lost or damaged and can only be used to improve the overall quality. Scalable coders not only allow a stream to be sent over channels with different bandwidth constraints or to devices having different capabilities, but also allow for different error protection schemes or even adaptive transmission techniques to be applied. Scalability is successfully introduced in the coding algorithms since the MPEG2 standard [4] up to the fine grain scalable option (FGS) in MPEG4 [5, 6] and H.264 [7].

A different approach in search for a solution to the problem of heterogeneous and unreliable networks is represented by multiple description coding (MDC). An MDC algorithm creates several substreams, all individually decodable, each at a lower quality than the original: receiving all the descriptions, ideally allows the full recovery of the single stream coded video [8, 9]. This approach is very attractive since it is possible to exploit the inherent protection provided by path diversity among the different descriptions [10].

Since scalability and multiple descriptions target the solution of different problems (bandwidth variations for scalability and robustness for multiple description coding), it is useful to exploit a combination of the two complementary methods in order to obtain a more efficient video coding algorithm. Previous works have already addressed the topic, mixing scalability and MDC thus creating scalable multiple descriptions (MDSC) algorithms. Approaches include exploitation of temporal segregation [11], a hybrid of spatial and temporal prediction [12], and wavelet coding [13, 14].

The starting point in this paper is to develop efficient mixes of scalability and multiple description and be compatible with a standard H.264/SVC coder. To this aim, we developed a method using simple pre and post processing schemes to generate substreams that can be used within the H.264/SVC coder. In the preprocessing part we down sample the original color and depth maps by rows and columns generating four subsequences that can be independently coded. To reduce redundancy, we propose to predict two of them by using some of the tools that guarantee scalability in the H.264/SVC coder. This method, called Interlayer Prediction Spatial Multiple Description Scalable Coding (ILPS-MDSC), takes advantage of the interlayer prediction method [7] introduced in H.264/SVC.

The first sections of the paper are devoted to a description of previous related work in 3D and multiple descriptions coding (Sections 2 and 3, resp.) and to the introduction of the scalable coding tools implemented in H.264/SVC (Section 4). The algorithm is described in Section 5, while the description of its implementation on top of the H.264/SVC coder and simulation results are provided in Section 6.

2. 3D Coding Algorithms

In this section we briefly recall the two basic techniques used to generate multiple views that can be used to generate 3D videos.

2.1. Stereoscopic (Left/Right) Video Coding. Coding of stereoscopic images has been the subject of research through several years. It basically mimics the human binocular view and consists of coding scenes captured by two slightly distant cameras. In stereoscopic viewing, two sets of images are used, one offered to the left eye and the other to the right eye of the viewer. Each of the views forms a normal two-dimensional sequence. Furthermore, the two views are taken under specific constraints, such that the human brain can fuse these two slightly different views resulting in the sensation of depth in the scene being viewed. Since this means doubling the information to be transmitted, the main focus in the research has been to exploit correlation between the two views. Since both cameras capture the same scene, the basic idea is to exploit the interview redundancy for compression. These redundancies can be of two types: interview similarity is across adjacent camera views, and temporal similarity between temporally successive images of each video. A coder that can efficiently explicit this kind of similarity is usually called *multiview coder* [15]. Unfortunately, although the two images are very similar, a *disparity* problem emerges as the system geometry may not be perfect. This disparity must be compensated and is generally described as an horizontal displacement between points collected by the two different sensors.

2.2. The DIBR Algorithm. The main advantage of DIBR technique compared to traditional representation of 3D video with left-right views is that it provides high-quality 3D video with smaller bandwidth. This is because the depth map can be coded more efficiently than the two streams of monoscopic views if correlations and properties of the depth map are properly identified. Conceptually, DIBR utilizes a depth frame to generate two virtual views from the same reference (original) view, one for the left eye and the other one for the right eye [16]. This process can be described by the following 2-steps procedure. Firstly, the original image points are re-projected into the 3D domain, utilizing the respective depth values. Thereafter, given the position of the left and right camera in the 3D domain, the views are obtained by the projection of the 3D intermediate points to the image plane of each camera. This procedure is usually referred to as "3D image warping" in the computer graphics literature. In this process the original image points at locations (x, y) are transferred to new locations (x_L, y) and (x_R, y) for left and right view, respectively according with the following formulas

$$x_L = x + \frac{\alpha_x \cdot t_c}{2}\left(\frac{1}{Z} - \frac{1}{Z_C}\right),$$
$$x_R = x - \frac{\alpha_x \cdot t_c}{2}\left(\frac{1}{Z} - \frac{1}{Z_C}\right), \quad (1)$$

where α_x is the focal length of the reference camera expressed in multiples of the pixel width and t_C is the distance between the left and right virtual cameras. Z_C is the convergence distance located at the zero parallax setting (ZPS) plane and Z denotes the depth value of each pixel in the reference view.

Notes that the y component is constant since the virtual cameras used to capture the virtual views (left-right) are assumed to be located at the same horizontal plane. The quality of virtual views depends on the quality of received color and depth map.

3. Multiple Descriptions Algorithms

Generally speaking, multiple description coding dates back to the mid 1970s when the first theoretical works appeared [17] soon followed by applications to voice coding [18]. Application to video was later introduced as described in [8]. The concept is simple and is based on splitting in some way the information to be transmitted which is then sent over independent channels. Being able to actually generate two (or more) distinct and independent information flows would optimally exploit transmission diversity. In the following we introduce some algorithms for MDC developed in the 2D video coding framework and some applications to 3D video.

3.1. Multiple Descriptions for 2D Video Transmission. A simple and efficient MDC scheme can be based on the temporal splitting of the odd and even frames of a video sequence into separate, individually decodable descriptions that can be decoded using standard receivers. Such a scheme, called Multiple Description Motion Compensation (MDMC), is described in [19] by designing temporal predictors that exploit not only the temporal correlation within a description, but also across the descriptions. Another simple method for MDC derives the descriptions as spatially subsampled copies of the original video sequence. In [20] a polyphase sub sampler along rows and columns is used to generate an arbitrary number of downsampled descriptions that are then independently coded. This scheme is then called Polyphase Spatial Subsampling multiple description coding (PSS-MDC).

The main problem of these techniques, as long as they are solely multiple description methods, is that they aim at increasing the robustness by exploiting link diversity and do not address other important transmission challenges, such as bandwidth variations or device heterogeneity, which require a scalable approach. On the other side, a traditional scalable approach does not guarantee the same robustness provided by MDC.

The complementarity of the two approaches has been exploited to implement an efficient solution that was called Multiple Description Scalable Coding (MDSC). A simple example of MDSC is the scalable extension of MDMC, proposed by [11]. A combination of motion compensation and spatial subsampling is described in [12].

A new type of MDSC in which the multiple description is not obtained only via spatial or temporal algorithms but also introducing quality metrics affecting signal-to-noise ratio (SNR) has been proposed by several authors using wavelet-based coding, in order to reduce temporal and spatial redundancy [13]. Another approach using the DWT is proposed in [14]. It is possible to combine MCTF and DWT in the so-called 3D (or 2D+t) discrete wavelet transform which may be used to generate a PSS-MD with an arbitrary number or descriptions by first subsampling the original sequence by rows and columns and then coding each of them [21].

Based on the scalable extension of the H.264 coder (H.264/SVC [22]), in [23] the authors proposed a further improvement of the MD-MCTF scheme, which separates each highpass frame generated by MCTF in two frames, the motion frame and the texture frame. Each of these frames is then handled separately, and the motion information is divided between the descriptions using a quincunx lattice, while the texture information is divided by sending the odd frames in one of the descriptions and the even frames in another with the exception of the intra coded macroblocks which are inserted in both descriptions.

3.2. Multiple Descriptions for 3D Video Coding. Depending on the stereoscopic coding approach used, different algorithms have been proposed to extend MDC to 3D video sequences.

There are primarily three approaches that can be used for coding of stereoscopic video. The first one, called *Simulcast stereoscopic coding,* involves independent coding of left and right views: either of the two views may be decoded and displayed as a compatible signal on 2D video receivers. In the second, called *Compatible stereoscopic coding*, one view, say the left view, is coded independently while the other one, in this case the right view, is then coded with respect to the independently coded view. Thus, only the independently coded view may be decoded and displayed as the compatible signal. Finally, the last method, called *Joint stereoscopic coding*, consists in coding both the left and right views together. Most of the common approaches tend to mix different views into one description. In these cases, one view is fully coded, while the others are usually time and/or spatially subsampled version of the originals [24]. This way of doing ensures that some information from which we reconstruct a three-dimensional vision is always available.

Considering the case of DIBR coding, descriptions may be formed using any of the MDC algorithms described for the case of 2D video while depth data are superimposed to any description and eventually represent the *enhancement layer* in an MDSC-like approach [25]. An example of this scheme is the one proposed in [25], where the temporal MDC of the base layer is produced by using the multistate video coding approach, which separates the even and odd frames of a sequence into two MDC streams. Hence, in the scalable MDC simulation, the even and odd frames are separated before the encoding process for both texture and depth and are used to generate the two descriptions: one contains the even frames of both color view and depth map, while the other description contains the odd frames. In both descriptions, frames from the color view are coded in the base layer and those from the depth map are coded in the enhancement layer.

4. Description of Scalable Coding Tools

H.264/AVC [22] is one of the latest international video coding standards. It uses state-of-the-art coding techniques and provides enhanced coding efficiency for a wide range of applications, including video telephony, video conferencing and video streaming. The scalable extension of H.264/AVC, named H.264/SVC, uses a layered approach to provide spatial, temporal, and SNR scalability such that only some specific combinations of spatial, temporal, and SNR data can be extracted and decoded from a global scalable video stream. The basic idea in designing the scalable extension of H.264/AVC is to extend the hybrid coding approach of this coder toward motion compensated temporal filtering. However, the algorithms that perform the prediction and update steps are similar to motion compensation techniques in the generalized B frame used in H.264/AVC, so that backward compatibility of the base layer of the scalable extension is guaranteed. Further detail of H.264/SVC can be found in [26].

In addition to the basic coding tools of H.264/AVC, SVC provides some so-called interlayer prediction methods, which allow the exploitation of the statistical dependencies between different layers to improve the coding efficiency of the enhancement layers. In the previous coder, the only supported interlayer prediction methods employ the reconstructed samples of the lower layer signal. The prediction signal is formed by motion-compensated prediction inside the enhancement layer, by upsampling the reconstructed lower layer signal, or by averaging such an upsampled signal with a temporal prediction signal. Although the reconstructed lower layer samples represent the complete lower layer information, they are not necessarily the most suitable data that can be used for interlayer prediction.

Usually, the interlayer predictor has to compete with the temporal predictor, and especially for sequences with slow motion and high spatial detail, the temporal prediction signal typically represents a better approximation of the original signal than the upsampled lower layer reconstruction. In order to improve the coding efficiency for spatial scalable coding, two additional interlayer prediction concepts have been added in SVC: prediction of macroblock modes and associated motion parameters and prediction of the residual signal. All interlayer prediction tools can be chosen on a macroblock or submacroblock basis allowing an encoder to select the coding mode that gives the highest coding efficiency and they are described below.

InterLayer Intra Prediction. for SVC enhancement layers, an additional macroblock coding mode is provided, in which the macroblock prediction signal is completely inferred from colocated blocks in the reference layer without transmitting any additional side information. When the colocated reference layer blocks are intra-coded, the prediction signal is built by the upsampled reconstructed intra signal of the reference layer: a prediction method also referred to as interlayer intra-prediction

InterLayer Macroblock Mode and Motion Prediction. when at least one of the colocated reference layer blocks is not intra-coded, the enhancement layer macroblock is inter-picture predicted as in single-layer H.264/AVC coding, but the macroblock partitioning—specifying the decomposition into smaller blocks with different motion parameters, and the associated motion parameters—is completely derived from the colocated blocks in the reference layer. This concept is also referred to as interlayer motion prediction. For the conventional intercoded macroblock types of H.264/AVC, the scaled motion vector of the reference layer blocks can also be used as replacement for usual spatial motion vector predictor.

InterLayer Residual Prediction. a further interlayer prediction tool referred to as interlayer residual prediction, targets a reduction of the bit rate required to transmit the residual signal of intercoded macroblocks. With the usage of residual prediction, the upsampled residual of the colocated reference layer blocks is subtracted from the enhancement layer residual (difference between the original and the inter-picture prediction signal) and only the resulting difference, which often has a smaller energy then the original residual signal, is encoded using transform coding as specified in H.264/AVC.

Other scalability techniques can be exploited by the H.264/SVC coder. In each layer, an independent hierarchical motion-compensated prediction structure with layer-specific motion parameters is employed. This hierarchical structure provides a temporal scalable representation of a sequence of input pictures that is also suitable for efficiently incorporating spatial and quality scalability. The reconstruction quality of a layer can be improved by additional coding of the so-called progressive refinement slices. These slices represent refinements of the texture data (intra- and residual data). Since the coding symbols of progressive refinement slices are ordered similarly to a coarse-to-fine description, these slices can be arbitrarily truncated in order to support fine granular scalability (FGS) or flexible bit-rate adaptation. Finally, Medium Grain Scalability (MGS) is introduced. This scalability allows to choose a partitioning of a block of 4×4 DCT coefficients as specified by the so called *MGSVectors*. A quality identifier (Q_{id}) which represents a priority inside the bitstream is then assigned to each packet of a slice. This type of scalability may lead to a number of layers that varies from coarse grain scalability to fine grain scalability. All these types of scalability are not specifically addressed in the present work.

5. Proposed Scheme

We have developed a scheme to generate multiple descriptions, based on the prediction algorithms of the scalable extension of H.264/AVC. In order to preserve the standard coder, a pre and post processing scheme is implemented. In the preprocessing part, we downsample the original sequence by rows and columns thus generating four different subframes, similarly to what is done in PSS-MD. These subframes could in principle be used to generate four independent descriptions. Since they are in general highly correlated, we can exploit this by reducing the number of descriptions to two so that each description will be formed

by two of the subframes arranged so that one subframe will be predicted form the other. By doing this, we still obtain two independent descriptions and, for each description, we remove a large part of the redundancy due to spatial correlation.

Scalability within each description is easily obtained exploiting the options provided by the scalable extension of the H.264 standard introduced in Section 4. Specifically, a coarse grain scalable (CGS) stream (i.e., a stream formed by a base layer and a single enhancement layer) is obtained by forcing the coder to accept one subframe as the base layer and the other as the enhancement layer. This process is depicted in Figure 3 and repeated for both the descriptions to be generated. Prediction is performed by the coder itself using its internal *interlayer* prediction *algorithms*. No prediction is performed across the descriptions.

In case of bandwidth variations or devices requiring a lower-grade signal, scalability will be simply obtained by discarding the enhancement layer(s). Note that a more radical form of scalability can be devised by allowing the receiver to decode only the base layer of one description, which means in practice that only 1/4th of the information will be delivered. In other words the following scenarios can be seen.

(1) Both descriptions are completely received. In this case all the information is recovered,

(2) One of the enhancement layers is dropped (regardless which one). In this case the receiver only looses 1/4th of the information as a description will be fully received together with the base layer of the other one,

(3) Both enhancement layers are lost. Now we have only half of the information available corresponding to the base layers of both descriptions,

(4) One description is completely lost as well as the enhancement layer of the other. Still something can be displayed to the user although corresponding to only one of the original image pixels.

In the postprocessing part, the original sequence is obtained by merging the descriptions. In case of lost description or discarded enhancement layers, the missing pixels are reconstructed by interpolation from the received ones.

Two different schemes are applied to group the subsequences obtained after the polyphase spatial subsampling. In the first one, called *by rows*, we group the subsequences in which the pixels form even or odd rows of the original sequence. In the other one, called *quincunx*, we group the subsequence so that the pixels form a quincunx lattice of the original sequence.

In other words, referring to Figures 1 and 2, if we number the four subsequences as in the raster scan from the top left corner, in the *by rows* scheme, we form the first description with subsequences one and two, and the other one with subsequences three and four. Instead, in the *quincunx* scheme, we group the subsequences one and three to form the first description, and the subsequences two and four for the other one.

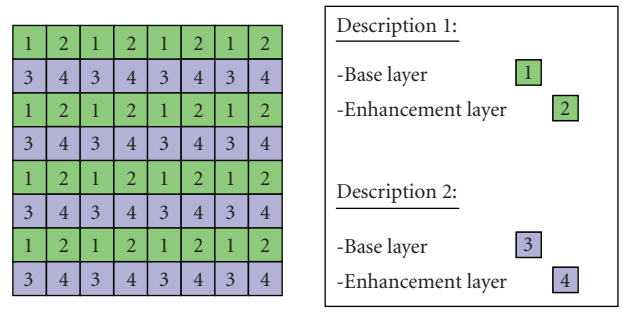

Figure 1: Subsampling patterns and descriptions composition for the case of by rows sampling.

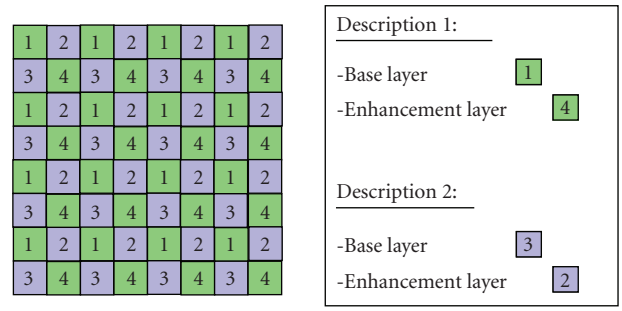

Figure 2: Subsampling patterns and descriptions composition for the case of quincunx sampling.

In case of lost information (either a whole description or the subframe corresponding to an enhancement layer), we use two different interpolation methods according to the different transmission scheme applied. In case we receive only one *by rows* description; then we recover the missing information taking the mean of the two nearest pixels. Otherwise, if we receive only one *quincunx* description, we recover the missing information simply taking the mean of the four nearest pixels.

Given this basic description of the algorithm, there are two different way to apply it to 3D video sequences. The first scheme, that takes into account the *Simulcast stereoscopic coding*, consists in coding independently the two different views (both Left/Right or Color/Depth), by applying the interlayer prediction algorithm to two subsequences of the same view, similar to what was done for the 2D sequences in previous works [27]. We call this method *Interlayer Prediction Structure (ILPS)*. In the second one, we try to create a *Compatible stereoscopic coding*, by subsampling the views in four subsequences, as done in PSS, and using the interlayer prediction to predict all the subsequences of one view (resp., right or depth), from the same subsequence of the other view (left or color). We call this scheme 3D-ILPS.

3D-ILPS is somehow similar to what is done in [24] where the authors mix a spatially scaled copy of one view with the full version of the other to form a description. By doing this, the authors do not introduce scalability in the strict sense in each description but exploit prediction to remove redundancy between the two views in each

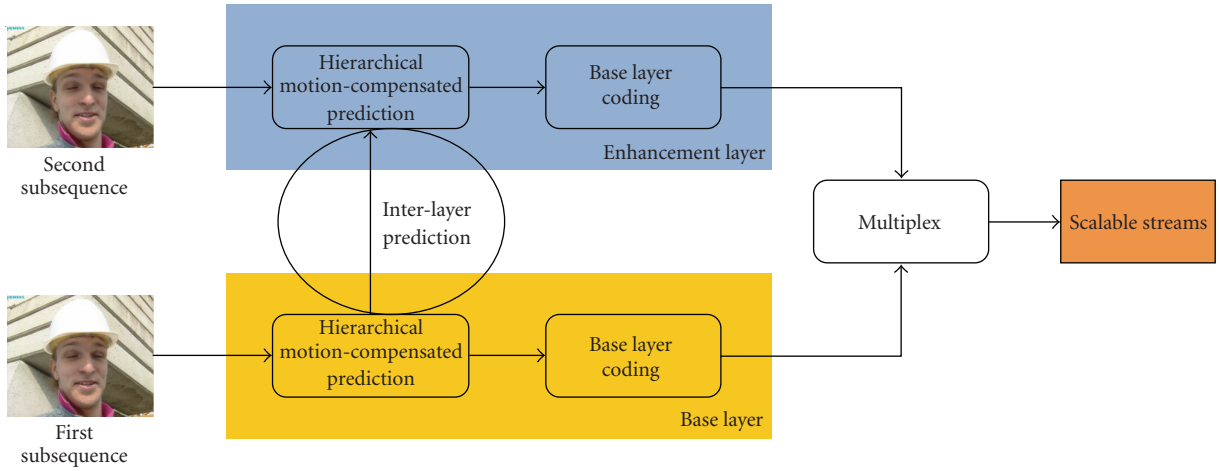

Figure 3: ILPS-MDSC with the interlayer prediction highlighted.

description. The redundancy introduced by this algorithm is then driven by the subsampling factor and the *QP* parameter applied to the predicted sequence.

The two proposed algorithms differ significantly and their characteristics can be exploited in different networking environments. In ILPS, each description carries a subsampled version of one of the two views, either left-right or color-depth, in a way that introduces a SNR scalability. In this method, the redundancy introduced is driven only by the header needed to correctly transmit all the descriptions. For those characteristics, we can say that the ILPS method can be very useful in case of transmission over unreliable channels in which all the receivers can decode a three-dimensional sequence, but have heterogeneous devices that require some sort of scalability. In the 3D-ILPS algorithm, instead, we propose a slightly different way to form the description, while keeping the same subsampling method as ILPS. By forming the description with, respectively, a subsampled version of the left/color and right/depth sequence, we introduce a new dimensional scalability. With this algorithm, it is possible to easily extract a 2D view from the stereoscopic sequence by discarding the enhancement layer. Considering this, we can apply the 3D-ILPS algorithm to network in which coexist users that can decode and view the 3D-videos and user that can only view monoscopic sequences.

6. Results

The software used in our experiments is H.264/SVC rel. 8.1, but the algorithm can be easily extended to virtually every version of the coder. Results are reported using the YUV 4 : 2 : 0 sequences *Breakdance* and *Ballet* (100 frames each), in order to determine the performance of the method under several different conditions. All sequences are at CIF resolution, 30 fps, single reference frame, GOP size 8. An 1 frame is only inserted at the beginning, prediction is performed at 1/4 pixel accuracy over 16×16, 16×8, 8×16, 8x8 blocks with SAD metric. Finally, CABAC is applied. Both sequences are natively at a resolution of 1024×768 and have been converted to CIF format by means of a non-normative downsampling, described in JVT-R006.

The results are shown over a rate span from 100 kbit/s to 1900 kbit/s in steps of 200 kbit/s, with or without random losses, to evaluate the pure performances of the coder with respect to other coding methods. In particular we plot the quality-rate curves for the color and left views, and we show some tables of the performances of the depth view, from a rate span of 100 kbit/s to 900 kbit/s, since after that rate the differences of PSNR are very minimal. Regarding the right view, no results are reported, since it can be easily extended from the ones showed for the left view.

Comparisons are mainly carried out using peak signal-to-noise ratio (PSNR) evaluation. Although it is known that it is unable to capture some 3D specific artifacts, PSNR is anyway widely used in literature to asses the quality of a 3D sequence from the quality of its components [25, 28, 29] and it has been shown that, though far from being perfect, it shows a good relationship with visual evaluation [30]. We also present some limited subjective quality measurement with the aid of the Mean Opinion Score (MOS), by playing the two rendered views (for the left and right eye) independently to the viewers.

In order to evaluate the robustness of the proposed algorithm with respect to the simple PSS-MD algorithm and a Single Description Coding (SDC), we have considered a random packet losses of about 10% of the overall transmitted packet. In case we receive only one description, we considered a packet loss of 5%, since in this cases we lost about the 55% of the information. In order to compare the performance of the single description coding with the case when only one description is received, we have simulated in such sequences a packet loss of about 50%, equal to the loss of information obtained when only one description is received.

A set of simulations has also been run considering a multiview video coder (MVC), specifically the H.264/MVC video coder version 5.0, where we predict the right sequence

Table 1: Performance of depth view, Ballet, without packet loss.

	One subsequence received				Two subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	39,46	33,22	33,22	33,50	39,46	34,28	35,77	37,17
300	44,58	34,47	34,47	34,45	44,58	36,49	38,88	39,57
500	47,19	34,74	34,74	34,73	47,19	37,02	39,85	40,32
700	48,95	34,86	34,86	34,85	48,95	37,28	40,32	40,67
900	50,38	34,92	34,92	34,92	50,38	37,41	40,61	40,89
	Three subsequence received				Four subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	39,46	35,16	35,58	35,81	39,46	35,78	36,61	36,70
300	44,58	38,75	38,96	38,85	44,58	41,55	42,15	41,86
500	47,19	39,81	39,95	39,87	47,19	44,20	44,84	44,48
700	48,95	40,37	40,47	40,37	48,95	46,15	46,82	46,31
900	50,38	40,69	40,77	40,70	50,38	47,69	48,35	47,90

Table 2: Performance of depth view, Ballet, 10% packet loss.

	One subsequence received				Two subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	34,64	32,62	32,38	33,24	34,64	33,48	34,90	36,50
300	36,97	33,92	33,76	33,92	36,97	35,59	37,54	38,41
500	39,13	34,30	33,99	34,11	39,13	35,90	37,93	39,02
700	39,53	34,19	34,36	34,43	39,53	36,22	38,52	39,67
900	40,75	34,21	34,18	34,54	40,75	36,09	38,85	39,55
	Three subsequence received				Four subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	34,64	37,07	34,42	35,40	37,69	34,07	35,49	35,89
300	36,97	35,65	37,18	38,14	40,61	37,62	39,36	41,22
500	39,13	38,16	38,29	39,24	43,78	38,42	42,54	42,24
700	39,53	38,88	39,14	39,18	44,70	38,52	43,83	44,63
900	40,75	38,84	38,96	39,64	44,39	39,15	45,18	46,14

from the left one. Unfortunately, this kind of coder does not support efficiently packet losses, so it was not possible to evaluate the performances of this algorithm in an error prone environment.

One final notation remark is that in all the tables, FD means that both base layer and enhancement layer are received for the specific description, BL refers to the case when only the base layer is received, and SVC and MVC are used to indicate single view and multiview video coding, respectively.

6.1. Algorithm Performance. We first present results for the cases when a subsequence is either completely received or it is completely lost. In case of loss of a subsequence, in order for the matching description to be decodable, we assume it corresponds to the subsequence used as the enhancement layer of the scalable description. Figures 4 and 5 show the performance when only one subsequence, corresponding to the base layer of one description, is received. As expected, all the proposed algorithms achieve the same performances since it is not possible, in this situation, to get benefits from our algorithms. In Figures 6 and 7 we consider the case when two subsequences, corresponding to the two base layers, are received. In this case, the performances of the *quincunx* and the *rows* schemes are very dependent from the specific sequence, so that it is not possible to select the best overall scheme. Instead, we can see that the ILPS algorithm seems to give better performances than the 3D-ILPS only in the *rows* scheme. Figures 8 and 9 show the performances when three subsequences are received. In these cases, all the considered algorithms seem to give similar performances, with only the 3D-ILPS algorithm that gives slightly worse performances. Finally, Figures 10 and 11 show the performances when all the subsequences are received. It is easy to see that, for the *Breakdance* sequence, the performances of the proposed method are comparable with the SDC and MDC, while, for the *Ballet* sequence, they are slightly worse.

Table 3: Performance of depth view, Breakdance, without packet loss.

	One subsequence received				Two subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	38,58	33,41	33,41	33,66	38,58	33,72	36,24	38,29
300	44,41	35,00	35,00	34,99	44,41	35,52	40,19	42,28
500	47,26	35,40	35,40	35,40	47,26	36,03	41,68	43,97
700	49,20	35,56	35,56	35,55	49,20	36,26	42,46	44,94
900	50,70	35,63	35,63	35,63	50,70	36,36	42,91	45,51
	Three subsequence received				Four subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	38,58	35,92	36,44	36,42	38,58	35,86	36,65	36,42
300	44,41	40,73	41,17	41,58	44,41	41,26	42,06	41,58
500	47,26	42,88	43,26	44,35	47,26	44,15	44,98	44,35
700	49,20	44,19	44,48	46,40	49,20	46,26	47,05	46,40
900	50,70	45,01	45,25	47,97	50,70	47,85	48,66	47,97

Regarding the depth view, Tables 1 and 3 report the PSNR value for the proposed algorithm, without packet losses. It is easy to see that the proposed algorithm gives better results than the PSS that and are not so far from the SDC. It is worth noting that all simulations are run with the coder generating the same total amount of bits for all the techniques. This means that the the MDC algorithms suffer from the fact that some redundancy is intrinsically introduced for two main reasons. First of all, since no prediction is performed across the descriptions, some correlation is anyway present and this reduces coding efficiency. Second, for each description to be independently (de)coded, we need headers to be duplicated. As a consequence, in the simulation conditions proposed, SDC will always perform better than MDC when all the information is received. The fact that ILPS and 3D-ILPS provide improvements with respect to PSS can be seen as a better coding efficiency.

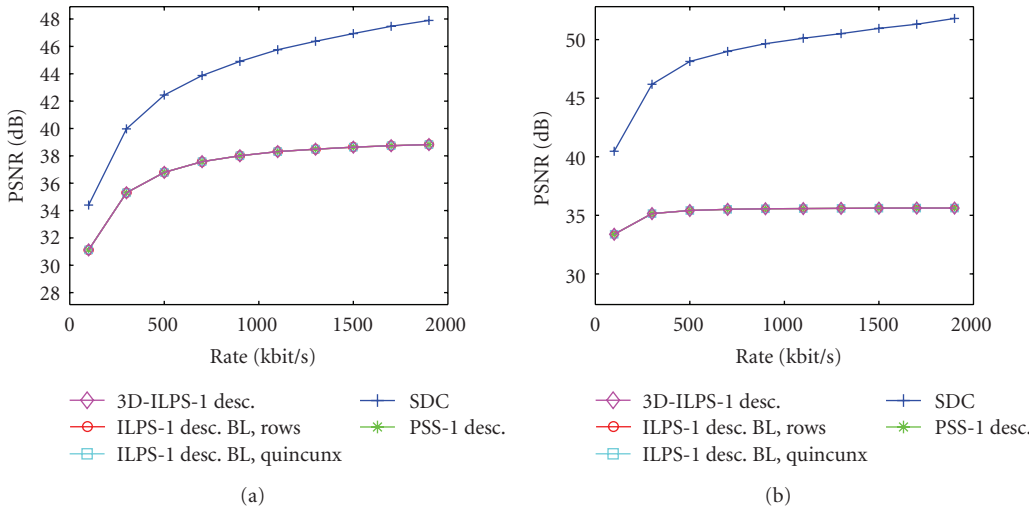

Figure 4: Performance of color view, one subsequence received. (a) Breakdance. (b) Ballet.

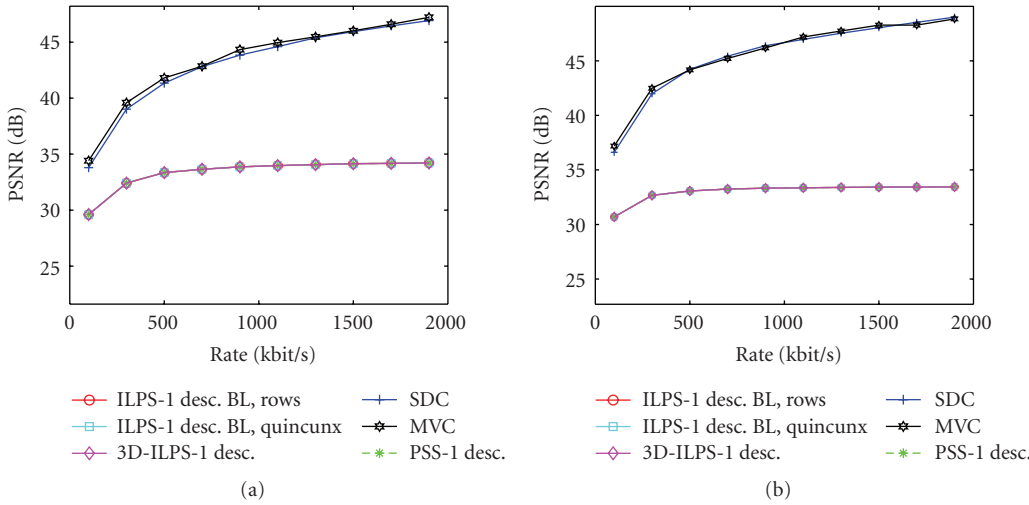

Figure 5: Performance of left view, one subsequence received. (a) Breakdance. (b) Ballet.

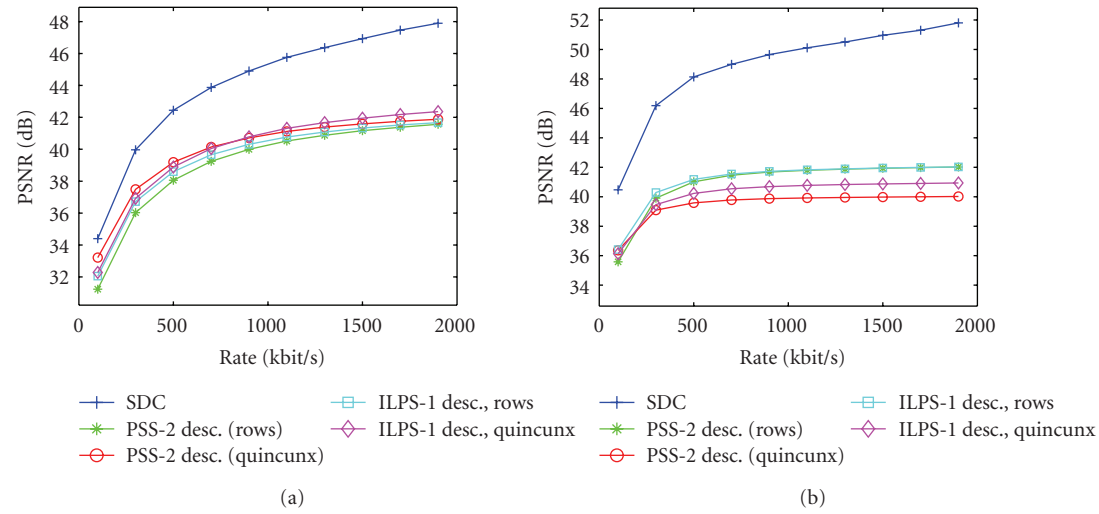

Figure 6: Performance of color view, two subsequences received. (a) Breakdance. (b) Ballet.

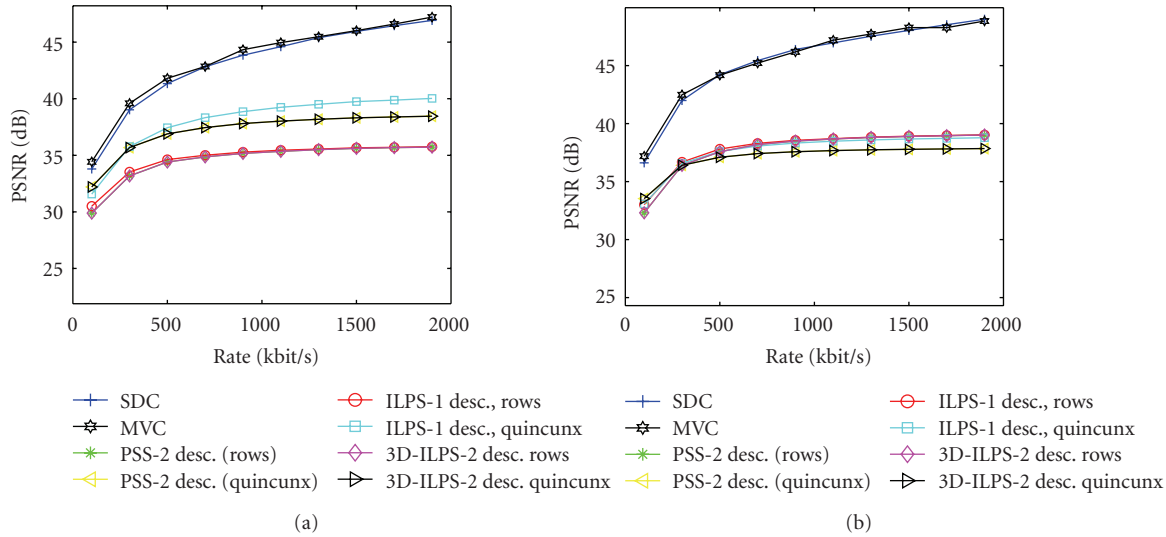

Figure 7: Performance of left view, two subsequences received. (a) Breakdance. (b) Ballet.

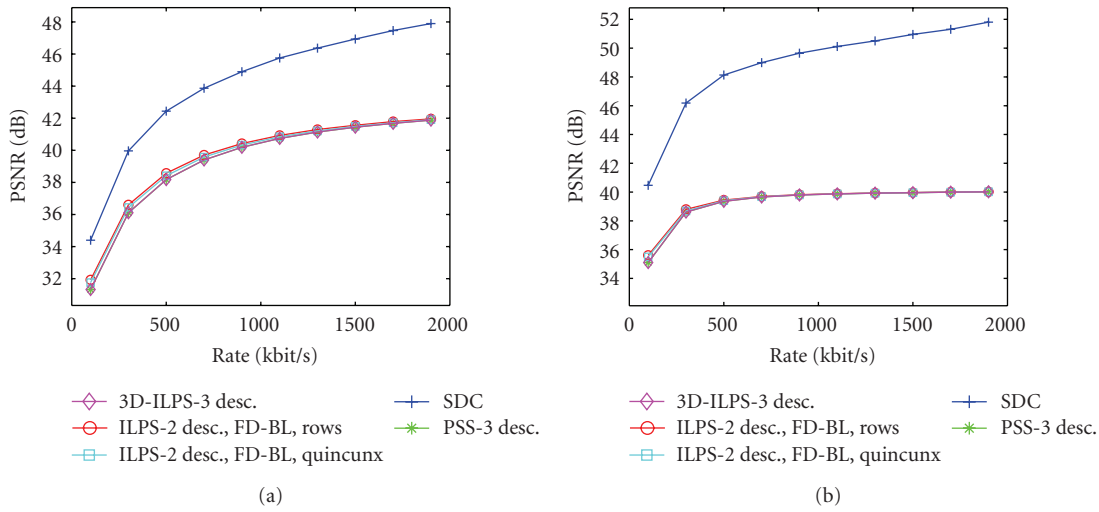

Figure 8: Performance of color view, three subsequences received. (a) Breakdance. (b) Ballet.

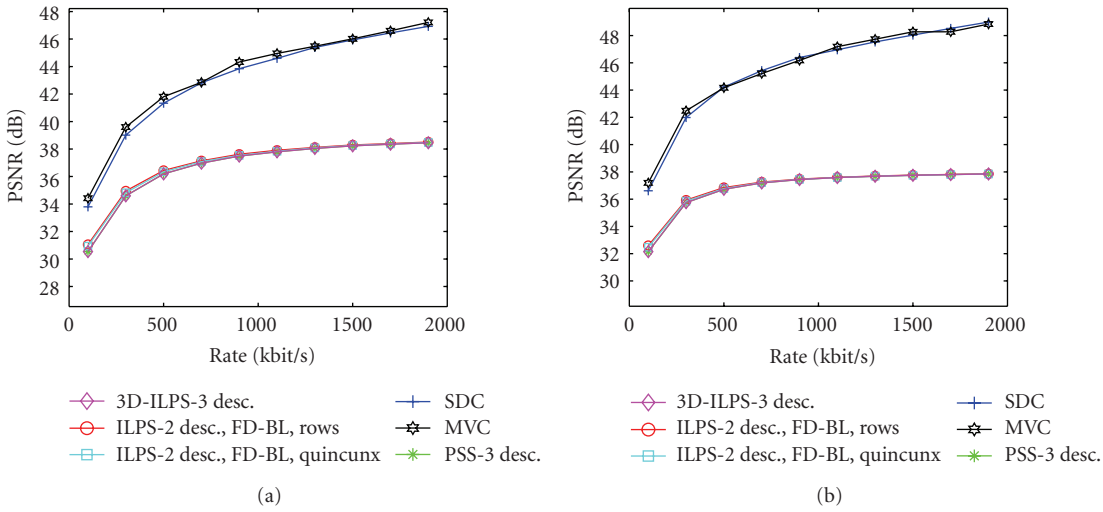

Figure 9: Performance of left view, three subsequences received. (a) Breakdance. (b) Ballet.

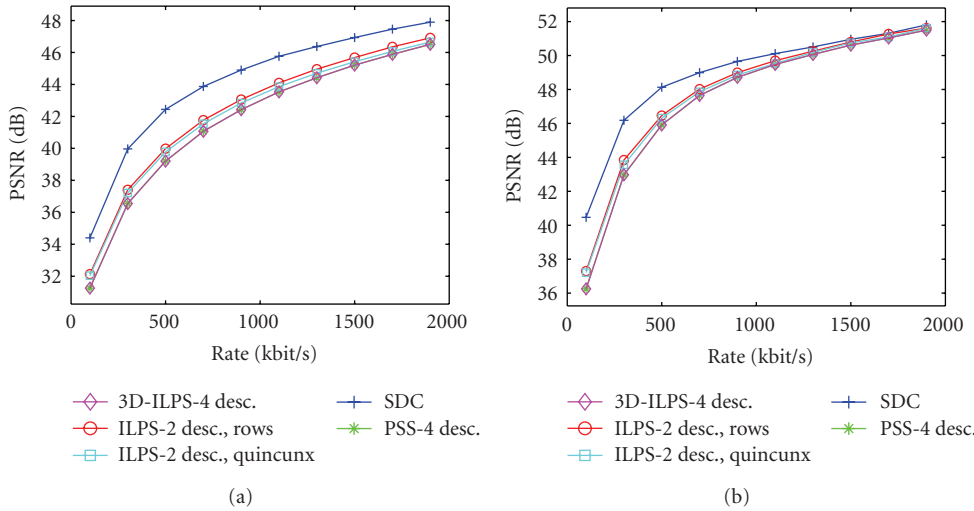

Figure 10: Performance of color view, all subsequences received. (a) Breakdance. (b) Ballet.

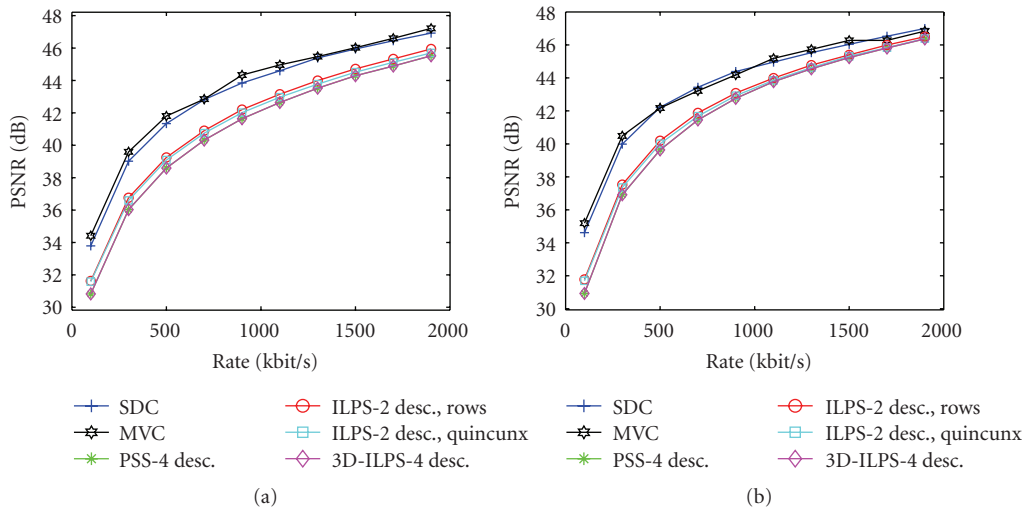

Figure 11: Performance of left view, four subsequences received. (a) Breakdance. (b) Ballet.

6.2. Performance with Random Losses. We now consider the case when random losses are superimposed to the previous cases. Random packet losses are generated as outcomes of a uniform distribution with loss probability $P = 10\%$. More detailed loss mechanisms should be considered in specific environments also taking into account the application scenario (real time versus streaming) and specific error control/retransmission strategies at medium access control layer. Furthermore, no specific error concealment strategies have been implemented so that it can be reasonably assumed that more realistic loss distributions would affect all different proposed methods in similar ways. Random losses on the contrary are less aggressive with single description streams.

Several simulations have been run for each of the possible cases using frame replacement as only error concealment technique. Despite of this, in some cases the decoder could not proceed: the reported results are given averaging over 10 different complete realizations. As expected, to obtain this number, SDC usually required a number of runs between two and three times larger than that of the MDC algorithms. This ratio may also be seen as a robustness gain provided by the path diversity approach. Note that no bandwidth constraints were given so that packet losses were only due to the random process.

Again, results are presented for increasing number of subsequences received and for color and left coding.

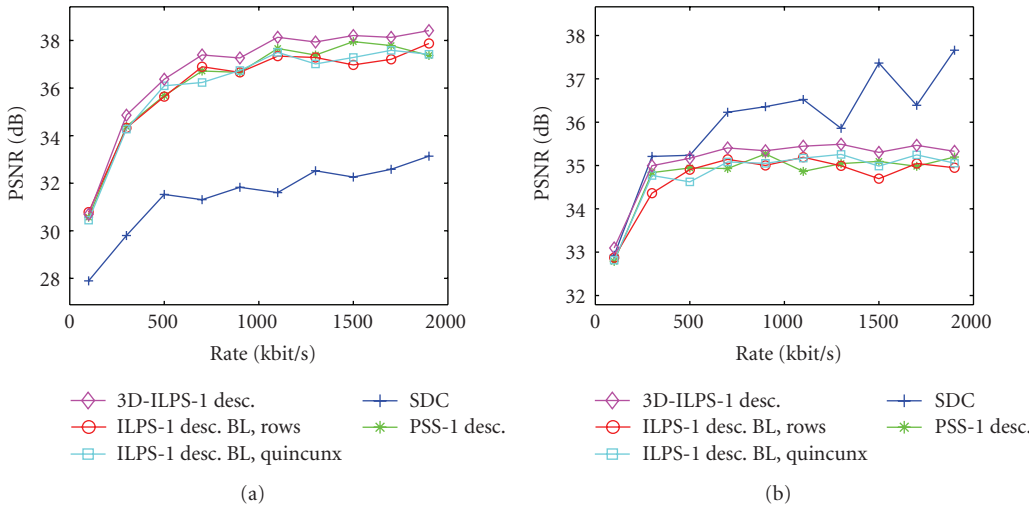

Figure 12: Performance of color view, one subsequence received, 10% packet loss. (a) Breakdance. (b) Ballet.

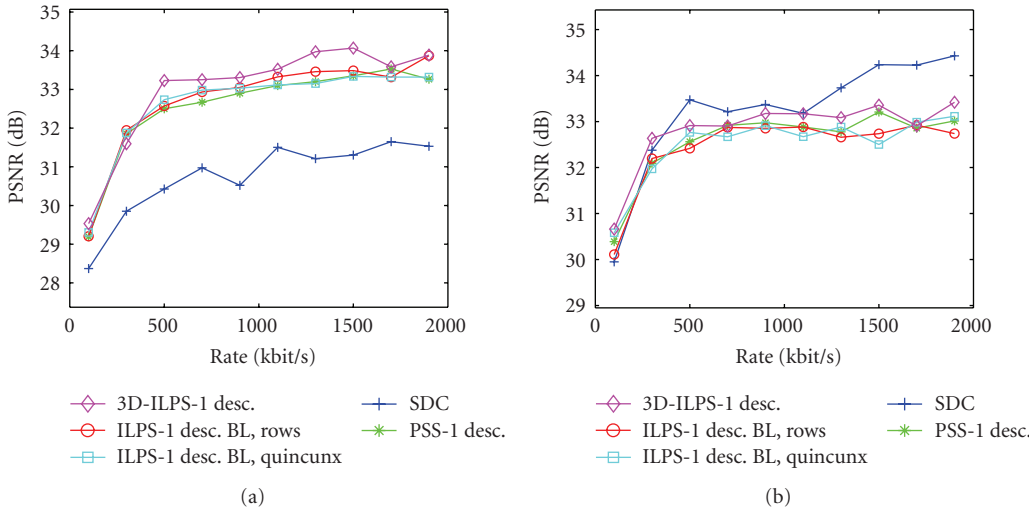

Figure 13: Performance of left view, one subsequence received, 10% packet loss. (a) Breakdance. (b) Ballet.

Figures 12 and 13 show the performance when only one subsequence is received, with random packet losses. In this case, the 3D-ILPS algorithm seems to give the better robustness to the packet losses. In Figures 14 and 15 it is possible to see the performance when two subsequences are received with random packet losses. As said before, the performances of the *quincunx* and the *rows* schemes are very dependent from the specific sequence, but we can easily view that the 3D-ILPS scheme again gives the best robustness to packet losses. Figures 16 and 17 show the performances when three subsequences are received. In these cases, all the considered algorithms seems to give similar performances, with only the 3D-ILPS algorithm that gives slightly lesser performances. Figures 18 and 19(b) sequences are received. In these cases, we can note a similar performance as the previous cases. Considering the depth view, Tables 2 and 4 report the robustness of the algorithms to a random packet loss of about 10%. These results indicate that both algorithms perform better than the PSS and the SDC coding. In particular, the 3D-ILPS algorithm seems to give the best performances among the considered schemes.

These results confirm the superior performances of the ILPS algorithm with respect to PSS, both in terms of coding efficiency and error robustness. This is not surprise since the 3D coding depicted scenario is an extension of the 2D coding scheme evaluated in the previous work [27].

The new algorithm proposed, the 3D-ILPS, further improves the performance of the ILPS scheme since it has been specifically studied for the 3D environment. Predicting the depth view from the same color view with the aid of the DIBR algorithm, this new algorithm has a better coding efficiency. Furthermore it also exhibits a strong robustness to transmission errors, since we can obtain four independent

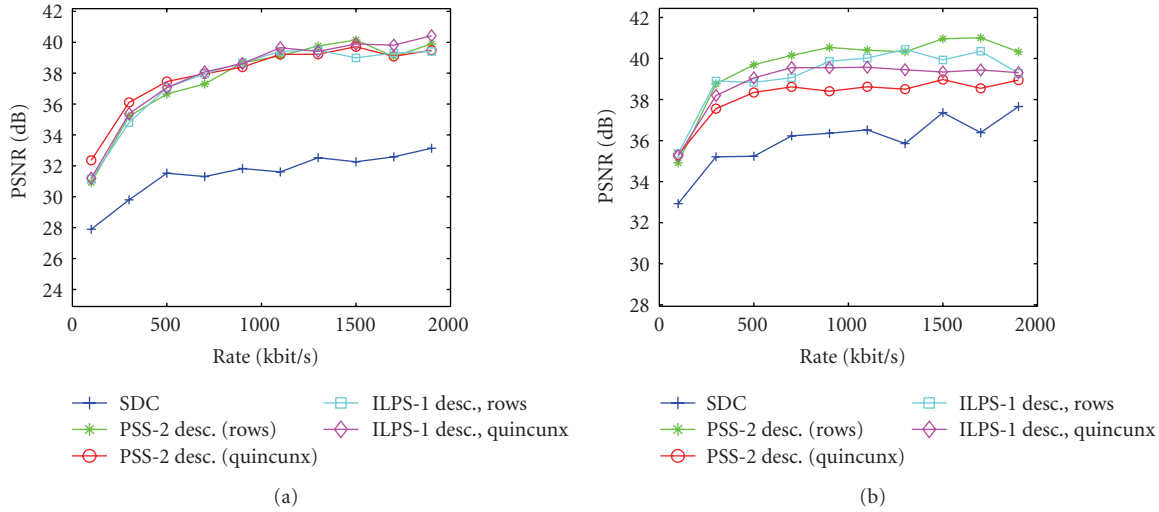

Figure 14: Performance of color view, two subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

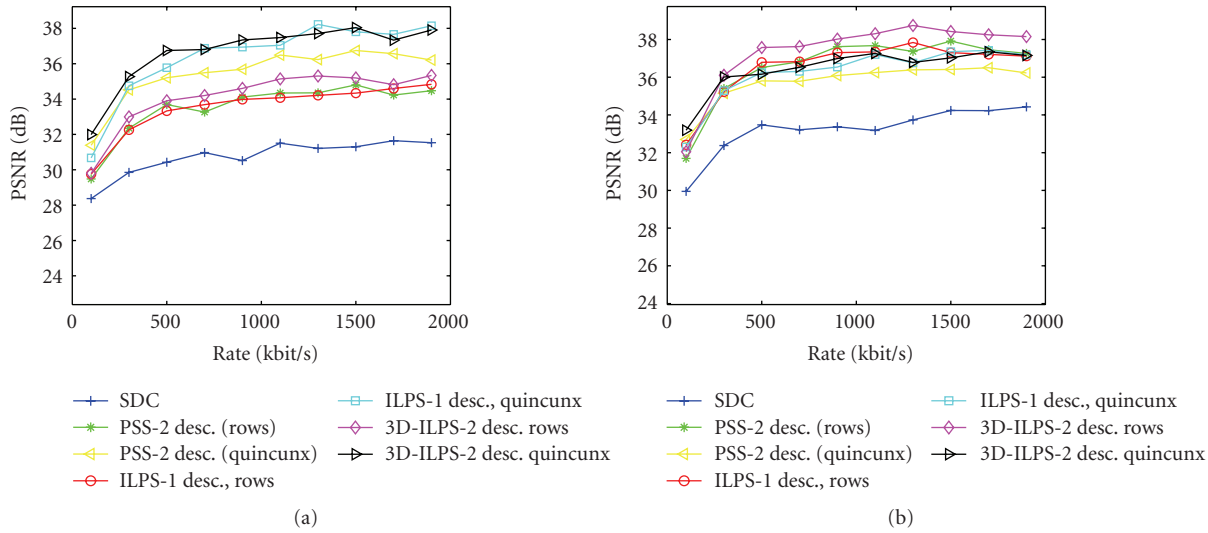

Figure 15: Performance of left view, two subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

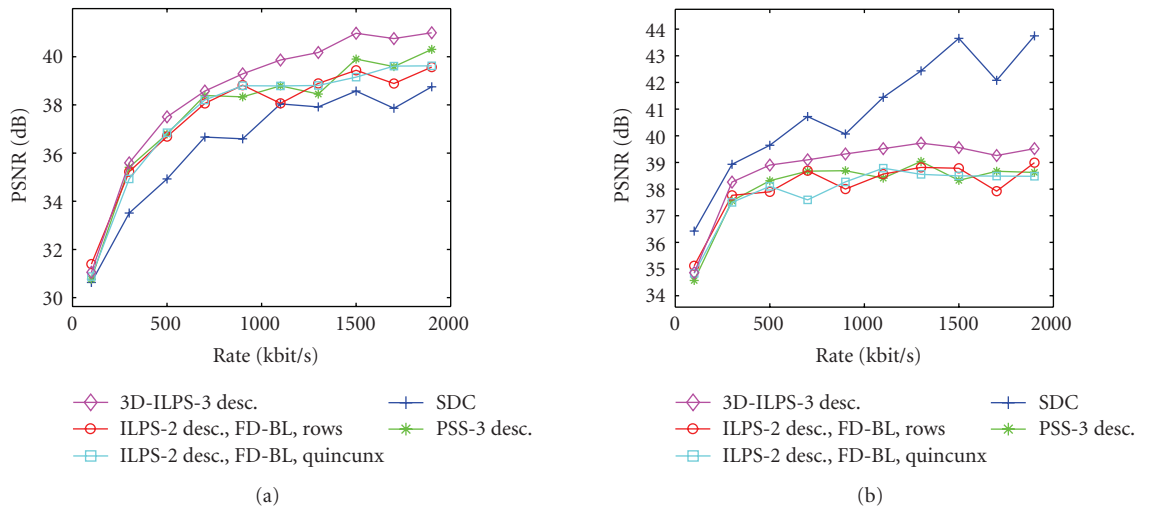

Figure 16: Performance of color view, three subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

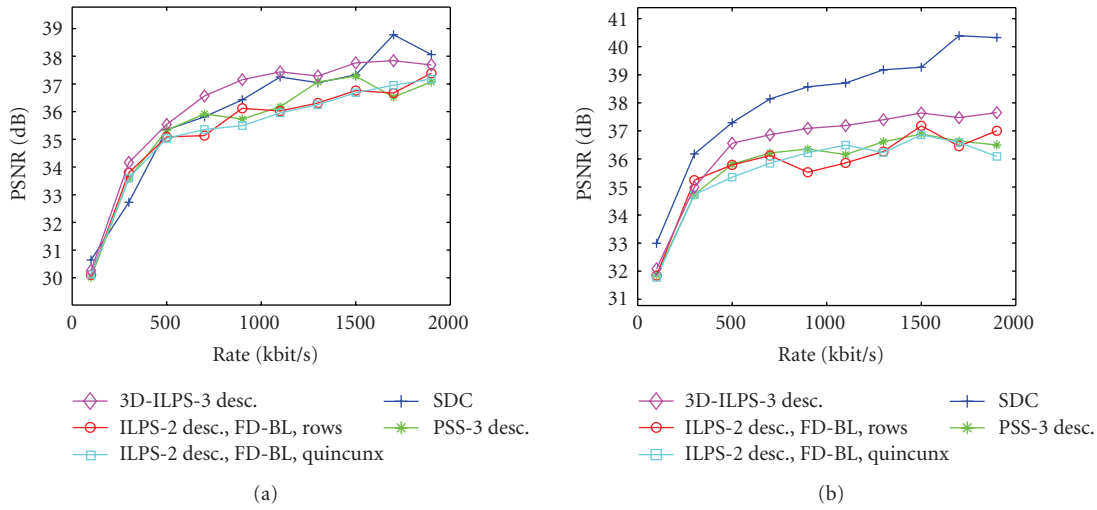

Figure 17: Performance of left view, three subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

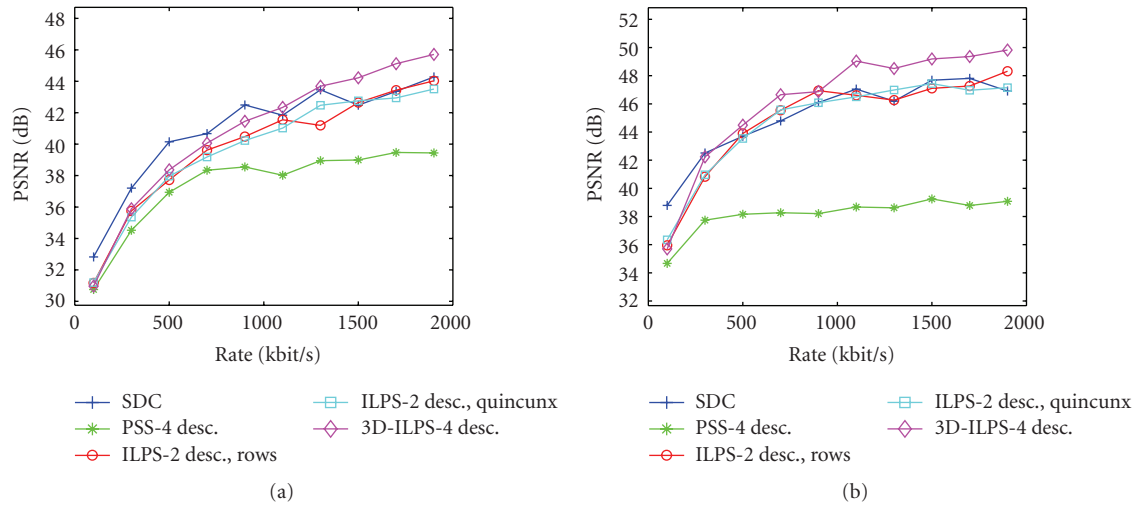

Figure 18: Performance of color view, all subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

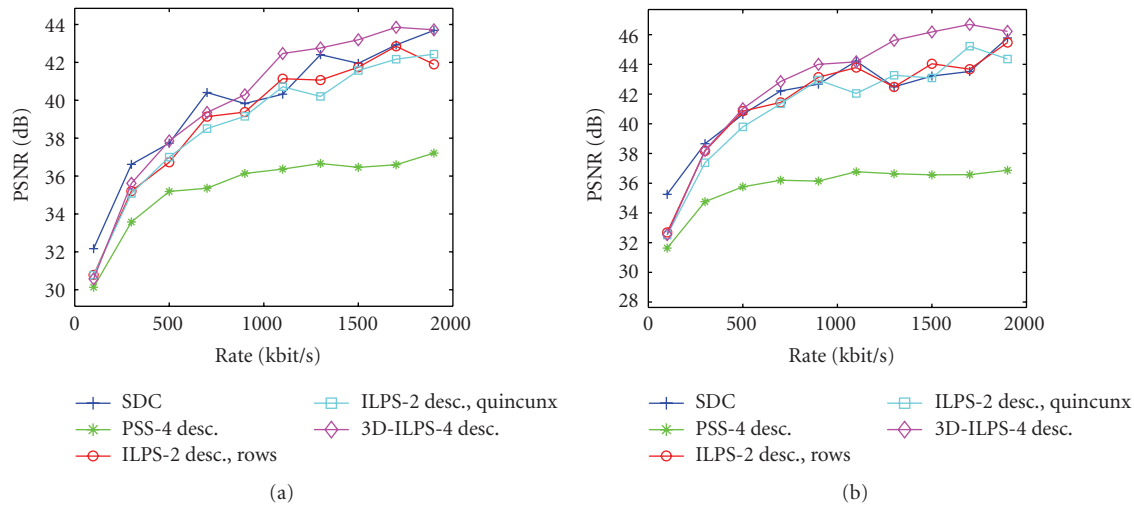

Figure 19: Performance of left view, four subsequences received, 10% packet loss. (a) Breakdance. (b) Ballet.

Figure 20: Image from Breakdance sequence. Left to right: color view, depth map, right view, left view. to bottom: SVC, PSS, ILPS, 3D-ILPS. All information received.

Figure 21: Image from Breakdance sequence. Left to right: color view, depth map, right view, left view. Top to bottom: SVC, PSS, ILPS, 3D-ILPS. All information received.

descriptions from the original sequence that can be sent over different channels or with different error protection schemes.

6.3. Subjective Tests. The "objective" results provided by PSNR are complemented by a set of subjective results. We first report, in Figure 20, images from the Breakdance sequence coded at an overall rate of 900 Kbit/sec for the proposed methods. It is supposed that no losses have occurred and it can be seen that the quality of all the algorithms cannot be discriminated (and this is true also at regular size). Figure 21 reports, on the contrary, the same cases when only the base layer of the two descriptions is received.

Visual experiments have been repeated to obtain subjective measures based on the mean opinion score (MOS) that can be very useful in order to produce some sort of metric for the Quality of Experience (QoE) [31]. Although the MOS evaluation could not be performed over a large number of viewers, we have been able to select a relatively small number of people, from our and neighboring laboratories, divided in two groups of about 10 persons each: with and without experience in video coding. We report only the results for the *Color* view, as the results for the other views are very

Table 4: Performance of depth view, Breakdance, 10% packet loss.

	One subsequence received				Two subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	30,92	33,01	32,96	33,26	36,86	33,21	35,45	37,190
300	32,93	34,40	34,18	34,52	40,76	35,03	38,46	40,45
500	33,34	34,63	34,63	34,56	43,71	35,21	39,91	41,82
700	35,09	34,58	34,79	34,75	44,63	35,36	41,19	42,43
900	34,42	34,72	34,90	34,96	45,74	35,73	40,93	42,77
	Three subsequence received				Four subsequence received			
Rate	SDC	PSS	ILPS	3D-ILPS	SDC	PSS	ILPS	3D-ILPS
100	34,25	35,17	35,34	35,80	36,86	34,89	35,02	35,58
300	37,78	38,84	39,36	39,79	40,76	39,14	39,61	40,17
500	39,32	41,84	40,92	41,61	43,71	40,80	41,47	42,38
700	39,28	41,68	41,88	43,02	44,63	42,57	43,68	43,43
900	40,63	43,49	43,61	43,37	45,74	42,33	45,96	45,73

Table 5: Mean opinion Score without packet loss, Breakdance/Ballet, color view.

	One description received				
	PSS-MD rows	PSS-MD quinc.	ILPS rows	ILPS quinc.	3D-ILPS
Viewer A	3/3.5	3.5/3	3.5/4	4/3.5	4/4.5
Viewer B	3.5/3	3.5/3.5	4/4.5	3.5/3	4.5/4.5
	Two descriptions received				
	SDC	MVC	PSS-MD	ILPS rows	ILPS quinc.	3D-ILPS
Viewer A	4.5/5	5/5	4/4.5	4.5/5	4.5/5	5/5
Viewer B	5/5	5/5	4.5/4.5	4.5/5	4/4.5	5/5

Table 6: Mean Opinion Score with 10% packet loss, Breakdance/Ballet, color view.

	One description received				
	PSS-MD rows	PSS-MD quinc.	ILPS rows	ILPS quinc.	3D-ILPS
Viewer A	2.5/2	2/2	3/2.5	2.5/2	3.5/3.5
Viewer B	3/2.5	2/2.5	2.5/3	3/3.5	3/4
	Two descriptions received				
	SDC	PSS-MD	ILPS rows	ILPS quinc.	3D-ILPS
Viewer A	4/3.5	3/3.5	4/3.5	3.5/4	4/4.5
Viewer B	3.5/4	3.5/3.5	3.5/3.5	3.5/3.5	4.5/5

similar to the ones reported. Results are in Tables 5 and 6. It is possible to see that the proposed method seems to achieve the same or even slightly better performances than PSS-MD, and at least the same performance than the SDC and MVC, in particular with packet losses. It is important to highlight that the 3D-ILPS method seems to give the best overall subjective quality among the proposed methods.

7. Conclusions

In this paper we have introduced an novel algorithm to generate 3D multiple descriptions in a H.264/SVC coder, with the use of *depth* and *color* views, as well *left* and *right* views, and we have shown its performance. Provided results show its effectiveness in terms of both coding efficiency, robustness, and flexibility. Work is in progress to improve these algorithms and to introduce them in more realistic network scenarios. In addition, future works will be related to introduce Medium Grain Scalability in each layer in order to be more flexible in transmission environments with varying effective bandwidth. Work is also planned to study more specific quality measures and models to determine rate distortion functions suitable to further adapt the proposed schemes and to be used in a more extensive QoE framework.

Acknowledgments

This work is partially supported by the Italian Ministry of Research (MUR) under Grants FIRB-RBIN043TKY "Software and Communication Platforms for High-Performance Collaborative Grid."

References

[1] C. Fehn, "A 3D-TV system based on video plus depth information," in *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers (ACSSC '03)*, vol. 2, pp. 1529–1533, Pacific Grove, Calif, USA, November 2003.

[2] B. Kamolrat, W. A. C. Fernando, M. Mrak, and A. Kondoz, "Joint source and channel coding for 3D video with depth image-based rendering," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 887–894, 2008.

[3] M. van der Schaar and P. A. Chou, Eds., *Multimedia over IP and Wireless Networks*, Elsevier, New York, NY, USA, 2007.

[4] ISO/IEC/JTC/SC29/WG11—ISO/IEC 13818.

[5] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301–317, 2001.

[6] H. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streamingover IP," *IEEE Transactions on Multimedia*, vol. 3, pp. 53–68, 2001.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Basic concepts for supporting spatial and SNR scalability in the scalable H.264/MPEG4-AVC extension," in *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP '05)*, Chalkida, Greece, September 2005.

[8] V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, 2001.

[9] Y. Wang, A. R. Reibman, and S. N. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–69, 2005.

[10] E. Setton, P. Baccichet, and B. Girod, "Peer-to-peer live multicast: a video perspective," *Proceedings of the IEEE*, vol. 96, no. 1, pp. 25–38, 2008.

[11] M. Liu and C. Zhu, "Multiple description video coding using hierarchical B pictures," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '07)*, pp. 1367–1370, Beijing, China, July 2007.

[12] N. Franchi, M. Fumagalli, R. Lancini, and S. Tubaro, "Multiple description video coding for scalable and robust transmission over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 3, pp. 321–334, 2005.

[13] M. van der Schaar and D. S. Turaga, "Multiple description scalable coding using wavelet-based motion compensated temporal filtering," in *Proceedings of the International Conference on Image Processing (ICIP '03)*, vol. 3, pp. 489–492, Barcelona, Spain, September 2003.

[14] H. Bai and Y. Zhao, "Multiple description video coding based on lattice vector quantization," in *Proceedings of the 1st International Conference on Innovative Computing, Information and Control (ICICIC '06)*, vol. 2, pp. 241–244, Beijing, China, August 2006.

[15] M. Flierl and B. Girod, "Multi-view video compression—exploiting inter-image similarities," *IEEE Signal Processing Magazine*, vol. 24, no. 7, 2007.

[16] C. Fehn, R. Rarre, and S. Pastoor, "Interactive 3D-TV-concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006.

[17] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple descriptions," *The Bell System Technical Journal*, vol. 59, no. 8, pp. 1417–1426, 1980.

[18] N. S. Jayant, "Subsampling of a DPCM speech channel to provide two 'self-contained' half-rate channels," *The Bell System Technical Journal*, vol. 60, no. 4, pp. 501–509, 1981.

[19] Y. Wang and S. Lin, "Error-resilient video coding using multiple description motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 438–452, 2002.

[20] M. Caramma, M. Fumagalli, and R. Lancini, "Polyphase downsampling multiple description coding for IP transmission," in *Visual Communications and Image Processing*, vol. 4310 of *Proceedings of SPIE*, pp. 545–552, San Jose, Calif, USA, January 2001.

[21] M. Yu, Z. Wenqin, G. Jiang, and Z. Yin, "An approach to 3D scalable multiple description video coding with content delivery networks," in *Proceedings of the IEEE International Workshop on VLSI Design and Video Technology (IWVDVT '05)*, pp. 191–194, Suzhou, China, May 2005.

[22] H. Schwarz, T. Hinz, H. Kirchhoffer, D. Marpe, and T. Wiegand, "Technical description of the HHI proposal for SVC CE1," *ISO/IEC JTC1/SC29/WG11*, Doc. m11244, Palma de Mallorca, Spain, October 2004.

[23] H. Mansour, P. Nasiopoulos, and V. Leung, "An efficient multiple description coding scheme for the scalable extension of H.264/AVC (SVC)," in *Proceedings of the 6th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '07)*, pp. 519–523, Vancouver, Canada, August 2007.

[24] A. Norkin, A. Aksay, C. Bilen, G. Bozdagi Akar, A. Gotchev, and J. Astola, "Schemes for multiple description coding of stereoscopic video," in *Proceedings of the International Workshop on Multimedia Content Representation, Classification and Security (MRCS '06)*, vol. 4105, pp. 730–737, Istanbul, Turkey, September 2006.

[25] H. A. Karim, C. T. E. R. Hewage, S. Worrall, and A. M. Kondoz, "Scalable multiple description video coding for stereoscopic 3D," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 745–752, 2008.

[26] R. Schäfer, H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "MCTF and scalability extension of H.264/AVC and its application to video transmission, storage, and surveillance," in *Visual Communications and Image Processing*, vol. 5960 of *Proceedings of SPIE*, no. 1, pp. 343–354, Beijing, China, July 2005.

[27] M. Folli, L. Favalli, and M. Lanati, "Parameter optimization for a scalable multiple description coding scheme based on spatial subsampling," in *Proceedings of the International Mobile Multimedia Communications Conference (MobiMedia '08)*, Oulu, Finland, July 2008.

[28] N. Ozbek and A. M. Tekalp, "Scalable multi-view video coding for interactive 3DTV," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 213–216, Toronto, Canada, July 2006.

[29] C. T. E. R. Hewage, H. A. Karim, S. Worrall, S. Dogan, and A. M. Kondoz, "Comparison of stereo video coding support in MPEG-4 MAC, H.264/AVC and H.264/SVC," in *Proceedings of the 4th IET International Conference on Visual Information Engineering (VIE '07)*, London, UK, July 2007.

[30] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1969–1976, 2008.

[31] A. van Moorsel, "Metrics for the internet age: quality of experience and quality of business," in *Proceedings of the 5th Performability Workshop*, Erlangen, Germany, September 2001.