Multiple Description Coding for Stereoscopic Videos With Stagger Frame Order

Chunyu Lin, Yao Zhao, *Senior Member, IEEE*, Tammam Tillo, *Senior Member, IEEE*, and Jimin Xiao

***Abstract*— Due to the prediction structures employed in video coding, the loss of one packet will affect many following frames. In this paper, a multiple description coding scheme with stagger frame order is proposed for stereoscopic 3-D videos. First, the reference and auxiliary views in stereoscopic sequences will be asymmetrically encoded into one description, whereas the other description will be formed in the same way with one dumb frame delay. Because of the stagger frame order, the coarsely encoded B frames will be inserted into different positions of the two descriptions. If a certain frame encoded with I/P mode is lost, then its corresponding B-frame version will be employed to compensate for the loss. In each description, the quantization steps of B frames are tuned based on a closed-form solution that considers the video contents, network status, frame positions in the group of picture, and the layer of the views. For further improvement, a fusing scheme is provided. The experimental results demonstrate that the proposed scheme outperforms state-of-the-art schemes. Specifically, up to 1.3-dB gain is achieved in the case of packet loss, and 2-dB gain is obtained for the side/central performance.**

***Index Terms*— Multiple description coding (MDC), stereoscopic video coding, video coding.**

I. Introduction

3-D VIDEO has attracted a lot of attentions both from academia and industry. Various types of 3-D formats, such as stereoscopic, 2-D + depth, and multiview videos appeared consequently [1]. To support the greater number of views compared with traditional 2-D formats, 3-D formats require increased data storage and incur a higher transmission burden. To efficiently compress 3-D videos, interview prediction, as well as intra-prediction and temporal inter-prediction has been adopted. Although motion prediction allows for higher compression efficiency, it also makes the compressed video stream vulnerable to transmission errors. When compressed 3-D videos are transmitted over error-prone channels, error propagations caused by packet losses lead to low video quality, thus causing a poor 3-D viewing experience. Hence, error resilience techniques for 3-D videos are in high demand. Stereoscopic 3-D is still the dominant 3-D format on the market, hence the majority of research, including ours, focuses on it.

To combat packet losses, forward error correction (FEC) is one possible solution. In [2], two different FEC codes are utilized to protect the stereoscopic video data against transmission errors. In [3] and [4], an approximate analytical model of the rate-distortion (RD) curve is proposed for three separate layers of 2-D video data to calculate the optimal source encoding bitrate for a given bandwidth. In [5], video packets that arrive after the display deadline of their frames are employed in combination with FEC code to improve error resilience. In [6], the expanding-window-based FEC scheme is proposed to improve the FEC recovery performance without introducing encoding/decoding dependency to avoid any delay. Joint source-channel decoding schemes are introduced with turbo code in [7] and [8]. Because of the FEC protection, some extra delay and high computation are introduced. Moreover, in conventional FEC systems, when the packet loss rate (PLR) of a channel exceeds the correction capability of the FEC codes, a cliff effect is observed, which results in highly unacceptable video quality.

Multiple description coding (MDC) is another effective way to combat packet loss over unreliable and nonprioritized networks. It provides a promising framework for video applications in which retransmission is unacceptable [9]. In MDC, one source is encoded into two or more representations (descriptions), which can then be transmitted over separate channels. If only one channel works, the side decoder can reconstruct the source with a certain desired fidelity, associated with a so-called side distortion. When all the channels work, the reconstruction quality can be enhanced up to the central distortion. The side distortion and central distortion cannot be simultaneously minimized; hence, the tradeoff between them should be tuned according to the network status. To achieve the same performance as that of single description scheme with error free transmission, there is always some redundant bitrate for MDC. Thus, the flexible tuning of redundancy is the key task of MDC schemes.

Many MDC schemes have been proposed for robust 2-D video coding [10], among which some have also been explored for stereoscopic 3-D video coding. In [11], the spatial scaling MDC (SS-MDC) and the multistate MDC (MS-MDC) schemes are proposed for stereoscopic videos. For SS-MDC, an asymmetric stereo pair is used to form descriptions, such that one view is at full resolution and the other view

Manuscript received March 31, 2014; revised July 20, 2014, September 3, 2014, and October 19, 2014; accepted October 29, 2014. Date of publication November 5, 2014; date of current version June 2, 2015. This work was supported by the 973 program (No.2012CB316400), supported by National Natural Science Foundation of China (No.61402034, No.61210006, and 61202240), supported by Beijing Natural Science Foundation (4154082), and supported by the specialized research fund for higher education under Grant 20130009120038). This paper was recommended by Associate Editor R. Hamzaoui.

C. Lin and Y. Zhao are with the Beijing Key Laboratory of Advanced Information Science and Network, Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: cylin@bjtu.edu.cn; yzhao@bjtu.edu.cn).

T. Tillo and J. Xiao are with Xi'an Jiaotong–Liverpool University, Suzhou 215123, China (e-mail: tammam.tillo@xjtlu.edu.cn; Jimin.Xiao@xjtlu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2014.2367391

is downsampled. For MS-MDC, temporal downsampling is applied. For example, the odd frames of both the left and right views are grouped to form one description, whereas the other description contains the information for the even frames. In [12], multiview videos are subsampled in both the horizontal and vertical directions to form four subsequences. Then, these four subsequences are paired to form two descriptions. In each description, one subsequence is directly encoded, whereas the other uses mode duplication (MD) based on the mode of the subsequence in the other description. This scheme is simple and efficient; however, its redundancy allocation is not flexible. Saliency detection for stereoscopic images has been studied in [13] and [14], in which the depth feature is extracted and employed. From the point of view of error resilience, the detected saliency region should be more strongly protected because of its importance. In [15], an MDC scheme with even and odd frames is also employed for stereoscopic videos by adding a controllable amount of side information to improve frame interpolation. Most existing MDC schemes for stereoscopic 3-D videos are simple extensions of 2-D video schemes, such that the characteristics of stereoscopic videos are not employed efficiently.

In [16], it has been shown that blocking effects produced by packet loss will affect the left and right views differently, thus causing binocular rivalry and considerable visual discomfort. Video freezing and frame rate dropping in stereoscopic 3-D also result in a low quality of experience. Instead, asymmetric compression on the left and right views can achieve the best perceived quality at a fixed total bitrate [17], [18]. In [19], it is concluded that asymmetric coding with SNR scaling achieves the best perceived quality at a high bitrate. All these factors should be considered when designing MDC schemes for stereoscopic videos.

In this paper, we propose an MDC scheme with stagger frame order that considers features of stereoscopic videos. The reference and auxiliary views in a stereoscopic video sequence will be asymmetrically encoded into one description, whereas the other description will be formed in the same way with one dumb frame delay. Because of the IBP structure and the introduction of one frame delay, a stagger frame order for B-frame positions is created. With this stagger frame order, the B frames will serve to reduce error propagation when their corresponding I/P versions are lost. Hence, the bitrate of the B frames is important for redundancy allocation. In the proposed scheme, the redundancy will be tuned based on the channel status, video contents, frame positions in the group of pictures (GOPs), and the layer of the views. Compared with classical schemes, gains of up to 1.3 and 2 dB can be achieved in the case of packet loss and in side/central performance, respectively.

The remainder of this paper is organized as follows. In Section II, an outline of the proposed scheme is sketched. First, the redundancy allocation is analyzed and modeled in Section II-A. For further improvement, Section II-B proposes a fusion scheme to reconstruct the description and implementation details are described in Section II-C. Experimental results and analysis are presented in Section III. Finally, the conclusion is drawn in Section IV.

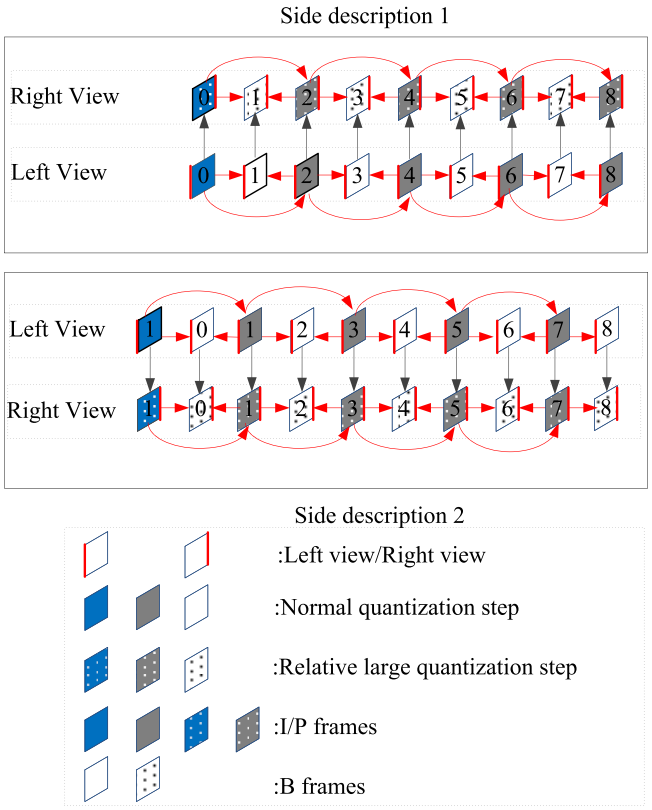

Fig. 1. Proposed multiple description scheme.

II. Proposed Scheme

The proposed multiple description scheme is shown in Fig. 1. In this scheme, a staggered frame order is created by inserting one dumb frame into one description. For side description 1, the stereoscopic video pair is encoded with asymmetric left and right quality. In our case, the left view will be used as the reference view, and the right view will be the auxiliary view. For side description 2, dumb frame 1 is inserted into the same stereoscopic video and all frames are still encoded with asymmetric left and right quality. According to the binocular suppression theory [19], [20], the reference and auxiliary views are encoded with asymmetric performance to ensure a good tradeoff between the perceived quality and the bitrate; the reference view will be the dominant perceived view, and hence, the auxiliary view (the right view in Fig. 1) is encoded at relatively low quality. Because descriptions 1 and 2 are encoded in a similar manner, only the encoding process for description 1 will be described hereafter unless otherwise noted.

Because of the proposed encoding structure, a stagger frame order for the B-frame positions will be created. Considering description 1 as an example, the left view is encoded at high quality, whereas the right view is encoded at relatively low quality; this also applies to description 2. If only description 1 is received, then the high-quality left view will establish the dominant perceived image quality. The block full of dots in Fig. 1 denotes that the corresponding frames will be encoded with a relatively large quantization step, thus resulting in low bitrate and quality.

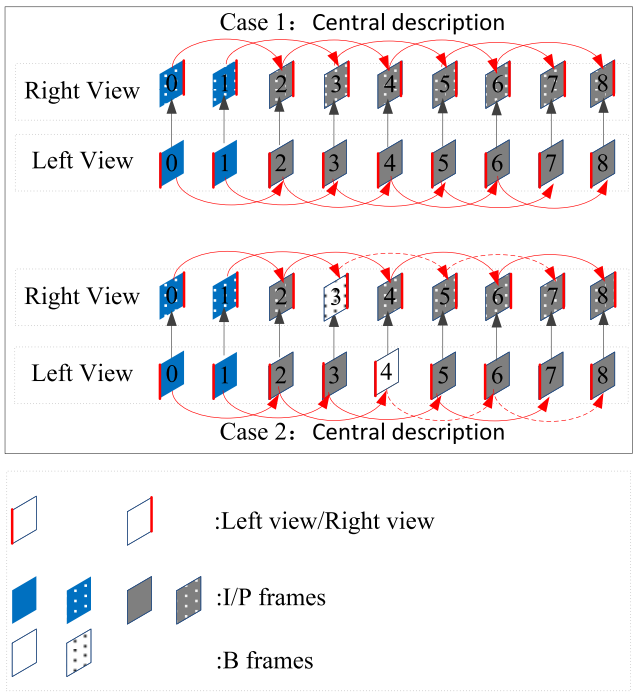

Fig. 2. Two cases with respect to the central performance.

The advantage of the stagger frame order can be described as follows. If both of the two descriptions are received, then none of the B frames in the two views will be used; instead, their corresponding I/P frames will be employed to achieve good central performance. However, if a certain I/P frame in one description is lost, then the corresponding B-frame version of the same content will be used to reduce the mismatch distortion. The quantization steps of the B frames will affect the mismatch distortion and the inserted redundancy, which should be tuned according to the network status, video contents, frame positions in the GOP, and the layer of views. Furthermore, we will also demonstrate that the fusion of the B-frame version and the P-frame version with propagated error can always achieve some gain.

When the channel condition is poor, there is a high probability that only one description will be received. In such a case, better side performance is preferred, and all frames, including the B frames in the reference view, should be of better quality. The redundancy is high in this case because the B frames in both of the two views will be coded with a quantization step close to that of the P frames. If the channel status is good, then the probability of receiving both of the two descriptions increases. In this case, the central performance is of greater importance, and the B frames in both of the two descriptions will be encoded at a low bitrate, meaning that the redundancy is low. Case 1 of Fig. 2 shows such an example, in which the central performance is reconstructed with the I/P frames, while the B frames encoded at a low bitrate are discarded. The above discussion represents two opposing cases: the first is optimized for good side performance, and the second is optimized for good central performance.

Generally, some frames/packets in one or the other description will get lost. If the lost frames/packets happen to be I/P frames, then subsequent frames will be affected. Concealment techniques can reduce some errors; however, the reconstructed quality is typically insufficient, which causes an uncomfortable viewing experience. In the proposed scheme, we will use the corresponding B frames to compensate for the lost I/P frames. Case 2 of Fig. 2 shows such an example, in which the P version of frame 4 in the left view is replaced with the B version from the other description. The figure also presents another example, in which the P version of frame 3 in the right view is lost and replaced with its corresponding B frame. To reduce mismatch and error propagation using B frames instead, the quantization parameters (QPs) of the B frames in each view should be assigned appropriately. In general, if the PLR is high, the B frames will be coded at a similar quality as that of the I/P frames. If PLR is low, then relatively high mismatch and error propagation can be tolerated. Hence, the key task is the redundancy allocation according to the network status.

A. Redundancy Allocation

In this section, we propose a mathematical framework for the redundancy allocation problem. The GOP structure is shown in Fig. 1, and the IBP structure is employed. Because the two descriptions are treated with similar coding processes, only description 1 will be analyzed. In addition, the reference view, that is, the left view in our scheme, will be analyzed first; corresponding conclusions can then be drawn for the auxiliary right view. For convenience, the terms reference view and auxiliary view will be used interchangeably with left view and right view, respectively.

Note that the B frames in description 1 will not be used unless one or more corresponding P-frame versions in description 2 are lost; the same also applies to the B frames in description 2. This means the quantization of the B frames in description 2 will be considered during the encoding process of description 1, whereas the quantization of the B frames in description 1 will be considered during the encoding process of description 2.

When no loss occurs, the I/P frames in both of the two descriptions will be combined in the temporal domain to reconstruct the video, which corresponds to Case 1 of Fig. 2. This process will not generate any mismatch. If a certain I/P frame in description 1 is lost, its corresponding B frame in description 2 will be employed to compensate for the loss error. Because the B-frame version is not exactly the same as that of the I/P frame version, a mismatch error will arise, which will propagate to subsequent frames. It is worth indicating that this type of mismatch error does not produce disturbing visual effect, such as those caused by concealment, which is attributable to the fact that the lost content is replaced with a low-quality version of the same content. The propagation is dependent on the position in which the loss occurs in the GOP, the layer of the view, and the PLR of the channel status.

Let $d_P[i]$ and $d_B[i]$ represent the distortions that arise when the ith frame is encoded with the P and B modes, respectively.

If the B version of frame i is used to compensate for a corresponding lost I/P frame, then the total distortion $d_B^t[i]$ will include the encoding distortion, the mismatch and the propagated distortions

$$d_B^t[i] = d_B[i] + \sum_{j=(i+2):2}^{N} d_{B,i}^m[j] + \sum_{k=i:2}^{N} d_{B,i}^{\prime m}[k] \quad (1)$$

where superscripts t and m denote the total distortion and the mismatch distortion, respectively. $d_B[i]$ is the distortion caused using B-frame version. $d_{B,i}^m[j]$ denotes the mismatch distortion at frame j caused by the loss of frame i. It will propagate from frame $i+2$, a P frame, through frame N, that is, the end of the GOP. $d_{B,i}^{\prime m}[k]$ represents the mismatch distortion in the auxiliary view, which is caused by the frame loss in the reference view. Due to the prediction structure used in the auxiliary view, the mismatch error $d_B^{\prime m}[k]$ will propagate from frame i, instead of $i+2$, through the end of the GOP. In addition, B-frame version will be replaced with its corresponding I/P frame version; hence, the error propagation to B frames is not considered here, which is the reason that the propagation applies only to every second frame, as indicated by $:2$ notation.

Equation (1) just applies to the case that loss occurs in the reference view. When loss occurs in the auxiliary view, the total distortion can be estimated as

$$d_B^{\prime t}[i] = d_B[i] + \sum_{j=(i+2):2}^{N} d_{B,i}^{\prime\prime m}[j] \quad (2)$$

where the total distortion $d_B^{\prime t}[i]$ includes the distortion caused using the B version and its propagated distortion $d_{B,i}^{\prime\prime m}$. Due to the prediction structure, the mismatch error will propagate from the current frame $i+2$ to the subsequent frames in the auxiliary view.

From (1) and (2), it can be seen that the key step is the estimation of the propagated distortion. In stereoscopic or multiview video coding, most studies have found that temporal prediction is adopted in regions with homogeneous motion or relatively static backgrounds, whereas the interview prediction is employed only in the regions with complex motion. In [21], the conclusion is reported that 13.1% of the macroblocks use interview prediction. In [22], the percentage for interview prediction is only 8% on average. Therefore, it is reasonable to assume that packet losses in the reference view will generate relatively small error propagations to the auxiliary view.

Let $X^R[i]$ and $X^A[i]$ denote the ith frame X in the reference view and auxiliary view, respectively. When one packet is lost in the reference view, the mismatch distortion will propagate to the auxiliary view through two paths. First, a packet loss on frame $X^R[i]$ will directly affect frame $X^A[i]$. The affected frame $X^A[i]$ will then influence its temporally subsequent P frames, such as $X^A[i+2]$. Second, the packet loss on frame $X^R[i]$ will affect its following P frames, such as frame $X^R[i+2]$. Because of interview prediction, frame $X^R[i+2]$ will influence its auxiliary view $X^A[i+2]$. These two propagation paths also apply to the remaining subsequent P frames in the auxiliary view. However, generally, the regions for which interview prediction is adopted contain complex motion [22]. These regions are generally complex and nonhomogeneous, so they are rarely used as reference because there is a high probability that their corresponding regions in subsequent frames will still use interview prediction instead of temporal prediction. Therefore, the propagation path from $X^A[i]$ to $X^A[i+2]$ caused by packet loss in the reference view is ignored here; only the second path, from $X^R[i+2]$ to $X^A[i+2]$, is considered.

Under the assumption stated above, $d_B^{\prime m}[k]$ can be estimated as $\beta d_B^m[k]$, where β is the percentage of frame k for which interview prediction is adopted. Then, (1) can be approximated as

$$d_B^t[i] = d_B[i] + \sum_{j=(i+2):2}^{N} d_{B,i}^m[j] + \sum_{k=i:2}^{N} \beta d_{B,i}^m[k]. \quad (3)$$

If a loss occurs on frame i in the auxiliary view, then β percent of that one frame will not be affected because that portion is predicted from the reference view. Then, (2) can be approximated as

$$d_B^{\prime t}[i] = d_B[i] + \sum_{j=(i+2):2}^{N} (1-\beta) d_{B,i}^m[j]. \quad (4)$$

Assume that a P frame is predicted only from its immediately preceding I/P frame, whereas a B frame uses the immediately preceding frame and the next frame as references. Then, a P frame in description 1 can be reconstructed as

$$\hat{X}_P[n] = P(\hat{X}_P[n-2]) + Q(X[n] - P(\hat{X}_P[n-2])) \quad (5)$$

where $X[n]$ denotes the nth frame and the subscript P represents the P mode. Then, the current frame $\hat{X}_P[n]$ uses its preceding P frame $\hat{X}_P[n-2]$ for prediction. Note that $X[n-1]$ is a B frame and will not be used as reference. $P(\cdot)$ denotes the prediction function, and $Q(\cdot)$ represents the quantization and inverse quantization process. In the other description, frame $X[n]$ is coded as a B frame and can be reconstructed as

$$\begin{aligned}\hat{X}_B[n] = {} & P(\hat{X}_P[n-1], \hat{X}_P[n+1]) \\ & + Q(X[n] - P(\hat{X}_P[n-1], \hat{X}_P[n+1])). \end{aligned} \quad (6)$$

Because of the more accurate prediction, B frames generally cost fewer bits than P frames. However, under the assumption that QP is sufficiently small and the residue is sufficiently large, the reconstructed distortion will depend on the QP, i.e., the same QP for P and B frames will produce similar distortion ranges. Let e_P and e_B denote the coding errors of the P frame and the B frame, respectively. Then, distortions $D(e_P)$ and $D(e_B)$ will have similar values. Even with a similar distortion range, it is not necessarily true that the P-frame version and the B-frame version will have the same reconstructed pixel values at all positions in the frame. Hence, when using a B frame $\hat{X}_B$ for substitution, the subsequent frames that use the P-frame version as reference will be affected by mismatch error. To obtain a closed solution, we must first estimate the mismatch distortion. Suppose that a P frame in one description is lost and replaced with its B-frame

version from the other description; the mismatch distortion can be calculated as

$$\begin{aligned} d^m &= D\{\hat{X}_P - \hat{X}_B\} \\ &= D\{(X - e_P) - (X - e_B)\} = D\{e_B - e_P\} \\ &= D\{e_B\} + D\{e_P\} - 2E\{e_B, e_P\} \\ &= \delta_B^2 + \delta_P^2 - 2\rho\delta_B\delta_P \end{aligned} \tag{7}$$

where D represents the variance function, E represents the expectation function, and $\hat{X}_P$ and $\hat{X}_B$ are the reconstructed values of X generated using the P- and B-frame versions, respectively. e_P and e_B are the coding errors of the P- and B-frame versions, respectively, whereas δ_P^2 and δ_B^2 are the variances of e_P and e_B, respectively. ρ is the correlation coefficient between e_P and e_B. For simplicity, the frame number n is omitted here. When the QP_P for the P frame and the QP_B for the B frame are equal, δ_B^2 and δ_P^2 will have similar values. If the quantization step of the B frame is larger than that of its P frame equivalent, δ_P will be smaller than δ_B and can be represented as $r\delta_B$. Then, (7) can be approximated as

$$\begin{aligned} d^m &= \delta_B^2 + r^2\delta_B^2 - 2r\rho\delta_B^2 = \delta_B^2(1 + r^2 - 2r\rho) \\ &= d_B(1 + r^2 - 2r\rho). \end{aligned} \tag{8}$$

The ratio r between δ_P and δ_B depends on the quantization steps. If the QP_B of the B frame is much larger than the QP_P of the P frame, then r will be very small and δ_P^2 will be negligible compared with δ_B^2.

The correlation coefficient ρ ranges from 0 to 1 and depends on the content of the current frame and that of its neighbor frames. The procedure for estimating ρ is presented in the next section. Under the assumption that a B frame has much larger distortion than its corresponding P version, (8) can be approximated as

$$d^m \approx d_B. \tag{9}$$

The mismatch distortion on the current frame X_n will propagate to the subsequent frames, through the end of the GOP.

Because of the deblocking filter and interpolation filter, the mismatch error will generally decay. In [23], the propagation function is simplified to $f[n] = e^{-\alpha n}$; this functional form is employed in our scheme because of its simplicity and effectiveness. For a packet loss in the reference view, the total distortion $d_B^t[i]$ in (1) can be calculated as

$$\begin{aligned} d_B^t[i] &= d_B[i] + d_B[i](1 + r^2 - 2r\rho) \\ &\quad \times \left(\sum_{j=(i+2):2}^{N} f[(j-i)/2] + \beta \sum_{k=i:2}^{N} f[(k-i+2)/2)] \right) \\ &= d_B[i] + d_B[i](1 + r^2 - 2r\rho)(\psi[i] + \beta\psi[i-2]) \end{aligned} \tag{10}$$

where $\psi[i] = \sum_{j=(i+2):2}^{N} f[(j-i)/2] = \sum_{j=2:2}^{N-i} f[j/2]$. For a packet loss in the auxiliary view, the total distortion $d_B'^t[i]$ in (4) can be approximated as

$$\begin{aligned} d_B'^t[i] &= d_B[i] + d_B[i](1 + r^2 - 2r\rho) \\ &\quad \times \sum_{j=(i+2):2}^{N} (1-\beta) f[(j-i)/2] \\ &= d_B[i] + d_B[i](1 + r^2 - 2r\rho)(1-\beta)\psi[i]. \end{aligned} \tag{11}$$

The term $(1 - \beta)$ reflects the error propagations within the same view, as β represents the percentage corresponding to interview prediction. If a certain packet in the auxiliary view is lost, it will not affect this β percent of the area because this portion is predicted from the reference view instead of its temporal frame.

Finally, for the reference view, the expected total distortion caused by frame i can be approximated as

$$\begin{aligned} \bar{d}[i] &= (1-p)d_P[i] + p(1-p)d_B^t[i] + p^2 d_0[i] \\ &= (1-p)d_P[i] + p(1-p)(1 + (1 + r^2 - 2r\rho)\psi[i] \\ &\quad + \beta\psi[i-2]))d_B[i] + p^2 d_0[i] \end{aligned} \tag{12}$$

where $d_P[i]$ represents the distortion generated with the P mode, whereas $d_B^t[i]$ denotes the total distortion, including the B mode distortion and its subsequent mismatch distortion. d_0 is the concealment distortion, which applies if neither the P nor the B frame version is received. To simplify the expression, let us define a weight parameter as

$$w[i] = (1 + r^2 - 2r\rho)(\psi[i] + \beta\psi[i-2]). \tag{13}$$

Then, (12) can be rewritten as

$$\bar{d}[i] = (1-p)d_P[i] + p(1-p)(1 + w[i])d_B[i] + p^2 d_0[i]. \tag{14}$$

On the one hand, $d_0[i]$ is less correlated with bitrate. On the other hand, $p^2 d_0[i]$ is quite small if PLR is low. Hence, this term can be ignored generally. The optimization problem for the reference view can be formulated as a constrained minimization one

$$\min \bar{D} = \sum_{i=1}^{N} \bar{d}[i] \tag{15}$$

$$\text{s.t.} \sum_{i=1}^{N} (R_P[i] + R_B[i]) = R_t \tag{16}$$

where $\bar{D}$ is the total expected distortion and R_t is the total bitrate of the two descriptions. $R_P[i]$ and $R_B[i]$ are the bitrates of frame i encoded with the P and B modes, respectively. To simplify the expression, here the I frame bitrate is not shown here. With standard Lagrangian approach, the constrained minimization problem can be solved as

$$L = \bar{D} + \lambda \sum_{i=1}^{N} (R_P[i] + R_B[i]) \tag{17}$$

where λ is the Lagrangian multiplier. By imposing $\nabla L = 0$, we obtain

$$\frac{\partial L}{\partial R_P[i]} = (1-p)\frac{\partial d_P[i]}{\partial R_P[i]} + \lambda = 0 \tag{18}$$

$$\frac{\partial L}{\partial R_B[i]} = p(1-p)(1 + w[i])\frac{\partial d_B[i]}{\partial R_B[i]} + \lambda = 0. \tag{19}$$

To minimize the expected total distortion $\bar{D}$, the two previous formulas can be jointly solved as

$$\frac{\partial d_P[i]}{\partial R_P[i]} = p(1 + w[i])\frac{\partial d_B[i]}{\partial R_B[i]}. \tag{20}$$

To select the optimum QP, the RD function for H.264/AVC is employed

$$\frac{\partial D}{\partial R} = -0.85 * 2^{\frac{\mathrm{QP}-12}{3}}. \tag{21}$$

Using (21), the relation between QP_P and QP_B can be obtained as

$$\mathrm{QP}_B[i] = \mathrm{QP}_P[i] - 3\log(p(1+w[i])). \tag{22}$$

This equation indicates that QP_B depends on the PLR p and the weight of error propagation $w[i]$. Furthermore, the weight $w[i]$ is determined by the video contents, the frame positions, and the layer of views. In particular, the higher p is, the smaller $\mathrm{QP}_B[i]$ is and the lower the propagated distortion will be. The expected distortion for the auxiliary view can be obtained similarly with a different weight, namely, $w'[i] = (1 + r^2 - 2r\rho)\ (1-\beta)\psi[i]$.

B. Further Improvement With a Fusion Scheme

When both P frames in description 1 and B frames in description 2 are encoded with the same quantization step and both of the two descriptions are received, we can average and fuse the two descriptions to obtain a better reconstruction, instead of simply discarding the B frames [24]. The new version reconstructed via fusion can be described as

$$\begin{aligned}\hat{X}' &= 0.5(\hat{X}_B + \hat{X}_P) \\ &= 0.5((X - e_B) + (X - e_P)) = X - 0.5(e_B + e_P).\end{aligned} \tag{23}$$

The distortion for the new reconstructed $\hat{X}'$ is

$$\begin{aligned}D(X - \hat{X}') &= D(0.5(e_B + e_P)) \\ &= 0.25\sigma_B^2 + 0.25\sigma_P^2 + 0.5\rho\sigma_B\sigma_P \\ &= 0.5\sigma_B^2(1+\rho).\end{aligned} \tag{24}$$

The final equation is obtained based on the assumption that $\sigma_B = \sigma_P$, which is reasonable when $\mathrm{QP}_P = \mathrm{QP}_B$. From (24), it is evident that there will always be some gain as long as $\rho < 1$. Hence, the expected distortion should be reevaluated when both the P frames and B frames use the same quantization step.

The above analysis is suitable for the case in which good side performance is desired. In the extreme case, each side description can achieve the same performance as a single description, but there is still some gain when both of the two side descriptions are received and fused. Notably, for most other MDC schemes, either it is not possible to achieve side performance similar to that of single description coding (SDC) or there is no gain when both of the two side descriptions are received.

Side/central performance applies predominantly to the ON–OFF channel. A certain amount of gain can also be achieved for the other types of channels. In our MDC scheme, P frames will be selected when both side descriptions are received, as shown in Case 1 of Fig. 2. However, if the P version is lost in one description, then its corresponding B frame in the other description will be used to compensate for the loss. Fig. 3 provides such an example, in which frame 2 in the left view is lost. Due to the mismatch between the P and B frames, the mismatch error will propagate from

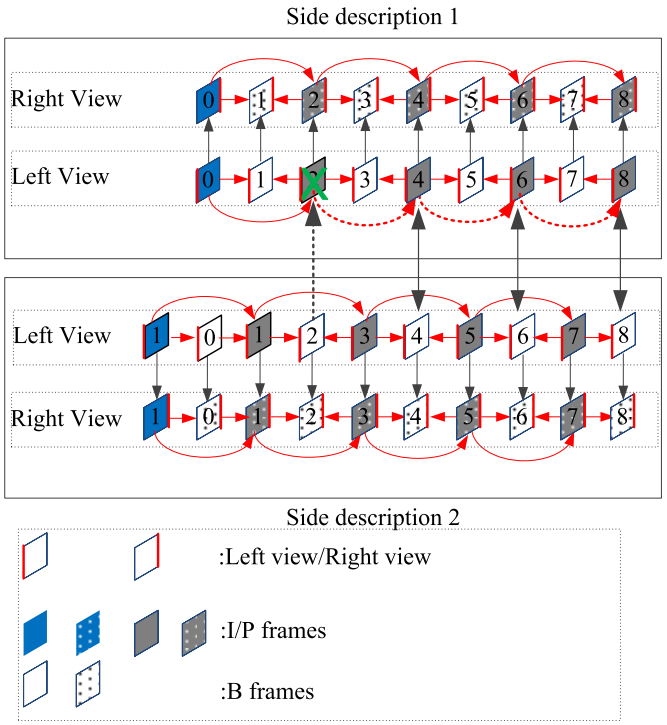

Fig. 3. Further improvement with a fusion scheme.

frame 2 to the subsequent frames. Typically, the subsequent P frames, such as frames 4, 6, and 8, will be selected to reconstruct the central description. However, because of the mismatch error propagation, the average of the P-frame version and the B-frame version will be employed instead as the new reconstruction.

The fusion procedure in the case of packet loss is described below. For frame $X[2]$ in description 1, its B and P reconstructions are

$$\hat{X}_P[2] = P(\hat{X}_I[0]) + Q_P(X[2] - P(\hat{X}_I[0])) \tag{25}$$

$$\hat{X}_B[2] = P(\hat{X}_P[1], \hat{X}_P[3]) + Q_B(X[2] - P(\hat{X}_P[1], \hat{X}_P[3])) \tag{26}$$

where X represents the frame; subscripts I, P, and B indicate the frame version; and the numerals indicate the frame positions. $P(\cdot)$ is used to represent the prediction process, whereas Q_P and Q_B represent the quantization and inverse quantization processes for P and B frames, respectively. Without packet loss, frame 4 in description 1 can be represented as

$$\hat{X}_P[4] = P(\hat{X}_P[2]) + Q_P(X[4] - P(\hat{X}_P[2])). \tag{27}$$

Due to the mismatch error, frame 4 in description 1 will be reconstructed as

$$\begin{aligned}\tilde{X}_P[4] &= P(\hat{X}_B[2]) + Q_P(X[4] - P(\hat{X}_P[2])) \\ &= P(\hat{X}_P[2] + \underbrace{(\hat{X}_B[2] - \hat{X}_P[2])}) + Q_P(X[4] - P(\hat{X}_P[2])).\end{aligned} \tag{28}$$

The term indicated by the bracket represents the mismatch distortion. Because of the relatively larger quantization step of the B frames compared with that of the P frames, this term

can be approximated as

$$\begin{aligned}(\hat{X}_B[2]-\hat{X}_P[2]) &= (X[2]-e_B[2])-(X[2]-e_P[2])\\ &= e_P[2]-e_B[2]\\ &\approx -e_B[2]. \end{aligned} \tag{29}$$

Using this approximation, (28) can be rewritten as

$$\tilde{X}_P[4] = P(\hat{X}_P[2]-e_B[2])+Q_P(X[4]-P(\hat{X}_P[2])). \tag{30}$$

Meanwhile, frame 4 in description 2 is encoded in the B mode and it can be reconstructed as

$$\hat{X}_B[4] = P(\hat{X}_P[3],\hat{X}_P[5])+Q_B(X[4]-P(\hat{X}_P[3],\hat{X}_P[5])). \tag{31}$$

In video coding, errors arise from the quantization process. Because of the generally small quantization step of the P frames, Q_P is smaller than QP_B, which is the reason that the P frame will be selected for the final reconstruction. However, with the introduction of mismatch error, the errors from $\tilde{X}_P[4]$ and $\hat{X}_B[4]$ will be in a similar range, as proven below.

Using (27), (30), and (31), the three different types of reconstructed error for X_4 can be represented as

$$e_P[4] = X[4]-\hat{X}_P[4] \tag{32}$$

$$\tilde{e_P}[4] = X[4]-\tilde{X}_P[4] = e_P[4]-e_B[2] \tag{33}$$

$$e_B[4] = X[4]-\hat{X}_B[4]. \tag{34}$$

Because prediction will not change the value of $e_B[2]$, $P(\cdot)$ can be ignored here and (33) can be obtained. Due to the relatively small quantization step, e_P is smaller than e_B. If e_P can be neglected, then the errors in $\tilde{X}_P[4]$ and $\hat{X}_B[4]$ will be in a similar range. The fusion of $\tilde{X}_P[4]$ and $\hat{X}_B[4]$ is represented by

$$\begin{aligned}&0.5(\tilde{X}_P[4]+\hat{X}_B[4])\\ &\quad= 0.5((X[4]+e_B[2])+(X[4]-e_B[4]))\\ &\quad= X[4]-0.5(e_B[4]-e_B[2]). \end{aligned} \tag{35}$$

From (24), we can draw a similar conclusion for (35), namely, the distortion that arises when the average of $\tilde{X}_P[4]$ and $\hat{X}_B[4]$ is used will always be smaller than that of $\tilde{X}_P$ alone, unless the residual values of $-e_B[2]$ and $e_B[4]$ are precisely identical. Hence, when there is packet loss on frame X_2 in description 1, the current P frame and subsequent P frames can be better reconstructed using the average of their $\tilde{\mathrm{P}}$ versions in one description and $\hat{\mathrm{B}}$ versions in the other description. In conclusion, we can always achieve some gain using our fusing scheme, whether QP_P and QP_B are similar or different, which is one of the advantages of the proposed scheme.

C. Implementation Details

By virtue of the asymmetric compression on the left and right views, the total bitrate can be reduced. However, the peak signal to noise ratio of the auxiliary view must be higher than a certain threshold, otherwise, symmetric compression on both views should be employed. This threshold depends on the display technology. For example, the threshold is 31 dB for a parallax barrier display and 33 dB for a full-resolution projection display [19]. In the proposed scheme, the QP offset between the reference and auxiliary views is fixed to 2, although better offset values can be achieved through subjective tests.

The QP_B for the entire reference view will be determined by (22), whereas the QP_B for the auxiliary view can be obtained in a similar manner as the weight $w'[i]$. To determine the weights $w[i]$ and $w'[i]$, the parameters ρ, r, and β should first be determined. β denotes the percentage of interview prediction, which depends on the camera distance used to record the two views, the video contents, and the correlation between neighbor frames. The majority of the frame in the auxiliary view will use the frame in the same layer for prediction, although some portion of the frame could use the frame in the reference view for prediction. In our case, the percentage β can be obtained when the first frame is encoded and tuned based on the last encoded frame. For example, the percentage of the frame for which interview prediction is employed can be obtained after the encoding of the current frame, and the next frame will update β using this new value. The parameter ρ can also be approximated from the last encoding process.

r depends on the quantization step. Although the quantization errors of B frames and P frames are different, the expected distortions associated with these errors are similar when they are quantized using the same QP. As long as the quantization steps are sufficiently fine, we can assume a uniform distribution of the transformed coefficients within each quantization level. The relation between the quantization errors and the quantization step can be described as

$$D = 1/12 * (\Delta)^2 \tag{36}$$

where Δ denotes the quantization step. Notably, in H.264/AVC, QP_P and QP_B are the QPs that control the quantization steps; therefore, there is a fixed relation between Δ and QP in H.264/AVC, and they can easily be converted into one another. Therefore, r^2 can be expressed as $(\mathrm{QP}_P/\mathrm{QP}_B)^2$. Here, we can see that r will be very small if $\mathrm{QP}_B \gg \mathrm{QP}_P$, in which case, a good approximation is $r \approx 0$. Finally, QP_B will be fixed in the range $[\mathrm{QP}_P, 51]$.

III. Experimental Results and Analysis

In this section, experiments are conducted using the following stereoscopic video sequences with three different resolutions, which are *Ballroom* (640 × 480), *Rena* (640 × 480), *Race* (640 × 480), *Exit* (640 × 480), *Soccer* (720 × 480), and *Poznan_street* (1920 × 1088). Among these sequences, *Rena*, *Exit*, and *Poznan_street* are relatively smooth, whereas the other three contain more motion. The proposed algorithm is implemented using JM18.5 H.264/AVC reference software. The GOP size is chosen to be 45, and the GOP structure for our scheme is IBP. To model the packet loss channel, packet loss patterns are selected from the error patterns for Internet experiments, as specified in Q15-I-16r1 [25]. The results are obtained using the complete loss patterns, which contain 10 000 binary characters.

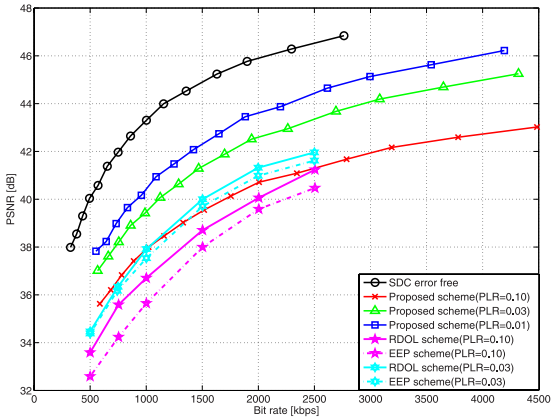

Fig. 4. Results for *Rena* at various PLRs.

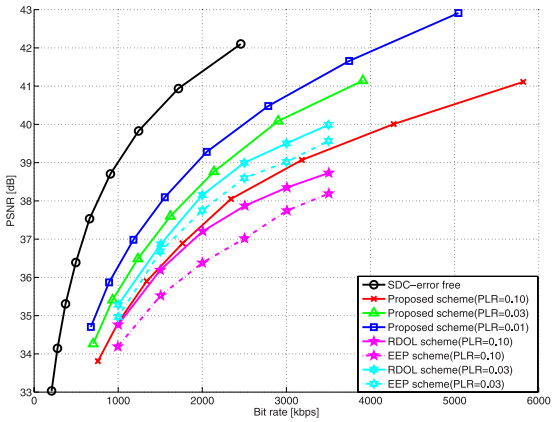

Fig. 5. Results for *Soccer* at various PLRs.

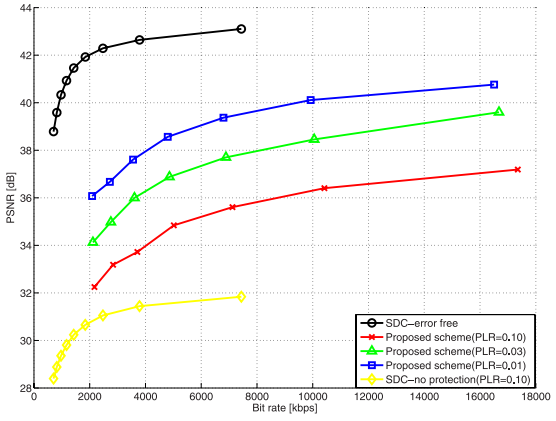

Fig. 6. Results for *Poznan_street* at various PLRs.

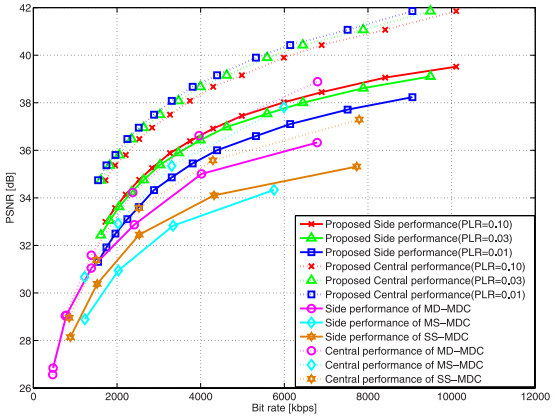

Fig. 7. Side/central performance results for *Ballroom*.

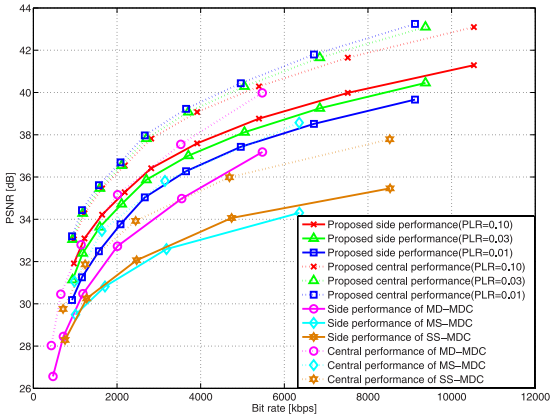

Fig. 8. Side/central performance results for *Race*.

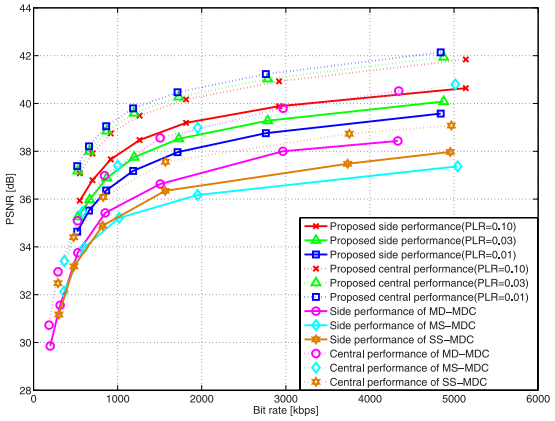

Fig. 9. Side/central performance results for *Exit*.

Figs. 4–6 provide the results for various PLRs. For comparison, the results of equal error protection and the RD optimized layer (RDOL) schemes from [4] are also presented for *Rena* and *Soccer*. In [4], systematic Luby transform code and Reed-Solomon codes are utilized to protect the stereoscopic video data according to RD curve modeling for each layer. Furthermore, the results of SDC for the error-free case are also provided. It can be seen that up to 1.3-dB gain can be achieved for *Rena* sequence, whereas the gains are lower for *Soccer* sequence. This difference arises because that the proposed scheme is more efficient when the sequence is smooth. For *Poznan_street*, only a comparison with the case of SDC is provided, as no comparable results are presented in [4]. In general, FEC schemes offer higher RD performance than MDC schemes, whereas MDC schemes have the advantage of low complexity and are suitable for the case of burst packet loss. Nevertheless, it is observed that even in the case of packet loss, the proposed MDC scheme is superior to those with which it is compared. In addition, if some burst packet loss occurs in one description, the proposed MDC scheme can still function, whereas the RDOL scheme presented in [4] cannot work well.

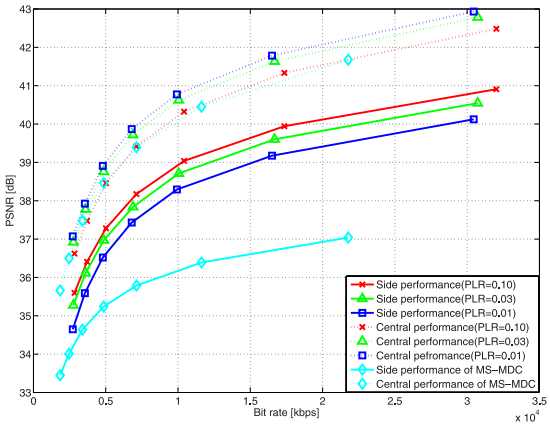

Fig. 10. Side/central performance results for *Poznan_street*.

Figs. 7–10 provide the results of side/central performance with various configurations. It can be seen that the side performance is still acceptable when one description is lost. For comparison, the results from [12] are also provided, in which MD is used to form an MDC scheme; this scheme will be referred to as MD-MDC. Moreover, the results of the SS-MDC and MS-MDC schemes [11] are also presented. The SS-MDC scheme scales one view in one description and the other view in the other description, whereas the MS-MDC algorithm exploits temporal subsampling to construct two descriptions. These three types of schemes are classical methods of constructing MDC descriptions.

With similar or higher central performance, the proposed scheme outperforms the other schemes by more than 2 dB in terms of side performance. In Fig. 10, only a comparison with the MS-MDC scheme [11] is provided for *Poznan_street* sequence. These MS-MDC results were simulated using our implementation. The central performances of MS-MDC and the proposed scheme were tuned to be similar to facilitate comparison between the side performances. In addition, the redundancy of the proposed scheme can be tuned to achieve different levels of tradeoff between side and central performance depending on the practical situation. Note that the average distortion of the two views, instead of the reference view alone, is used here for the performance comparison. However, because of the asymmetric encoding of the reference and auxiliary views, the better quality will be the dominant one in our case. Hence, this comparison is unfavorable to our case. Even so, the proposed scheme still outperforms the other schemes. All results testify the effectiveness of the proposed scheme.

In Table I, the luminance Bjøntegaard delta (BD) rate savings achieved with fusion scheme described in Section II-B are presented. The gain is large when the PLR is high, which can be attributed to the high probability of fusion in such cases. Note that the complexity introduced by simply averaging the two corresponding frames is quite small, but the BD rate savings are nevertheless approximately 2%–17%. In the extreme case, if QP_P and QP_B are equal, BD rate savings of more than 30% can be achieved; this corresponds to the case with the highest redundancy and the maximum protection for our MDC scheme.

TABLE I
BD RATE SAVINGS WITH THE FUSION SCHEME

Sequence	*PLR*(%)	Y BD-rate (%)
Rena	1	-2.91
	3	-6.4
	10	-7.65
Ballroom	1	-2.79
	3	-4.09
	10	-5.98
Race	1	-4.52
	3	-5.68
	10	-6.47
Soccer	1	-2.17
	3	-4.42
	10	-5.58
Exit	1	-2.76
	3	-6.53
	10	-7.94
Poznan_street	1	-12.6
	3	-14.71
	10	-17.2

IV. CONCLUSION

In this paper, an MDC scheme with stagger frame order is proposed for stereoscopic video communication. Because of the introduction of one dumb frame delay, the B frames are inserted into different positions in the two descriptions. By tuning the bitrate allocation for the B frames using a closed-form equation, different types of protection can be achieved depending on the network status, video contents, position in the GOP, and the layer of views. The proposed scheme also employs a fusion scheme to allow B frames to be used to enhance the performance. Compared with classical schemes, the proposed scheme can achieve gains of up to 1.3 and 2 dB in the case of packet loss and in side/central performance, respectively. Moreover, the fusion scheme yields BD rate savings of approximately 2%–17% compared with the case without fusion. In conclusion, the proposed MDC scheme is a promising approach for the transmission of stereoscopic 3-D video in error-prone channels.

ACKNOWLEDGMENT

The authors would like to thank Prof. C. Cai for providing us with his experimental results, Dr. C. Yao and American Journal of Experts for their assistance in modifying this paper, and the anonymous reviewers for multiple suggestions that resulted in a significantly improved manuscript.

REFERENCES

[1] A. Vetro, A. M. Tourapis, K. Müller, and T. Chen, "3D-TV content storage and transmission," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 384–394, Jun. 2011.

[2] A. S. Tan, A. Aksay, C. Bilen, G. B. Akar, and E. Arikan, "Error resilient layered stereoscopic video streaming," in *Proc. 3DTV Conf.*, May 2007, pp. 1–4.

[3] A. S. Tan, A. Aksay, C. Bilen, G. B. Akar, and E. Arikan, "Rate-distortion optimized layered stereoscopic video streaming with raptor codes," in *Proc. Packet Video*, Nov. 2007, pp. 98–104.

[4] A. S. Tan, A. Aksay, G. B. Akar, and E. Arikan, "Rate-distortion optimization for stereoscopic video streaming with unequal error protection," *J. Adv. Signal Process.*, vol. 2009, no. 7, pp. 1–14, 2009.

[5] J. Xiao, T. Tillo, C. Lin, Y. Zhang, and Y. Zhao, "A real-time error resilient video streaming scheme exploiting the late-and early-arrival packets," *IEEE Trans. Broadcast.*, vol. 59, no. 3, pp. 432–444, Sep. 2013.

[6] J. Xiao, T. Tillo, and Y. Zhao, "Real-time video streaming using randomized expanding Reed–Solomon code," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1825–1836, Nov. 2013.
[7] A. Vosoughi, V. Testoni, P. Cosman, and L. Milstein, "Joint source-channel coding of 3D video using multiview coding," in *Proc. IEEE ICASSP*, May 2013, pp. 2050–2054.
[8] N. Ramzan, A. Amira, and C. Grecos, "Efficient transmission of multiview video over unreliable channels," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 1885–1889.
[9] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.
[10] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.
[11] A. Norkin, A. Aksay, C. Bilen, G. B. Akar, A. Gotchev, and J. Astola, "Schemes for multiple description coding of stereoscopic video," in *Proc. Int. Conf. Multimedia Content Represent., Classification Secur.*, 2006, pp. 730–737.
[12] X. Wang and C. Cai, "Mode duplication based multiview multiple description video coding," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2013, p. 527.
[13] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Nov. 2013, pp. 1–6.
[14] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625–2636, Jun. 2014.
[15] H. Karim, A. Sali, S. Worrall, A. Sadka, and A. Kondoz, "Multiple description video coding for stereoscopic 3D," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2048–2056, Nov. 2009.
[16] J. Gutierrez, P. Perez, F. Jaureguizar, J. Cabrera, and N. Garcia, "Subjective evaluation of transmission errors in IPTV and 3DTV," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Nov. 2011, pp. 1–4.
[17] G. Saygili, C. G. Gurler, and A. M. Tekalp, "Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 593–601, Jun. 2011.
[18] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4021–4024.
[19] G. Saygili, C. G. Gürler, and A. M. Tekalp, "Quality assessment of asymmetric stereo video coding," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4009–4012.
[20] L. Pinto and P. Assuncao, "Asymmetric 3D video coding using regions of perceptual relevance," in *Proc. Int. Conf. 3D Imag. (IC3D)*, Dec. 2012, pp. 1–6.
[21] A. Abdelazim, S. J. Mein, M. R. Varley, and D. Ait-Boudaoud, "Fast prediction algorithm for multiview video coding," *Opt. Eng.*, vol. 52, no. 3, p. 037401, 2013.
[22] L. Shen, Z. Liu, S. Liu, Z. Zhang, and P. An, "Selective disparity estimation and variable size motion estimation based on motion homogeneity for multi-view coding," *IEEE Trans. Broadcast.*, vol. 55, no. 4, pp. 761–766, Nov. 2009.
[23] T. Tillo, M. Grangetto, and G. Olmo, "Redundant slice optimal allocation for H.264 multiple description coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 59–70, Jan. 2008.
[24] C. Lin, Y. Zhao, and C. Zhu, "Two-Stage diversity-based multiple description image coding," *IEEE Signal Process. Lett.*, vol. 15, pp. 837–840, 2008.
[25] Y.-K. Wang, S. Wenger, and M. M. Hannuksela, *Common Conditions for SVC Error Resilience Testing*, JVT document P206, Aug. 2005.

Chunyu Lin was born in Liaoning, China. He received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2011.

He was a Visiting Researcher with the Information and Communication Theory Group, Delft University of Technology, Delft, The Netherlands, from 2009 to 2010. From 2011 to 2012, he was a Post-Doctoral Researcher with the Multimedia Laboratory, Ghent University, Ghent, Belgium. His research interests include image/video compression and robust transmission, 2-D-to-3-D conversion, and 3-D video coding.

Yao Zhao (M'06–SM'12) received the B.S. degree from the Department of Radio Engineering, Fuzhou University, Fuzhou, China, in 1989; the M.E. degree from the Department of Radio Engineering, Southeast University, Nanjing, China, in 1992; and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He became an Associate Professor with BJTU in 1998 and a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science at BJTU. He leads several national research projects in the 973 Program, 863 Program, and National Science Foundation of China. His research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

He serves on the Editorial Boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE SIGNAL PROCESSING LETTERS, and *Circuits, System, and Signal Processing* (Springer), and an Area Editor of *Signal Processing: Image Communication* (Elsevier). He was named as a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013.

Tammam Tillo (M'05–SM'12) received the Dipl.-Ing. degree in electrical engineering from Damascus University, Damascus, Syria, in 1994 and the Ph.D. degree in electronics and communication engineering from Politecnico di Torino, Turin, Italy, in 2005.

He was a Visiting Researcher with École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2004 and was a Post-Doctoral Researcher with the Image Processing Laboratory, Politecnico di Torino, from 2005 to 2008. For a few months, he was an Invited Research Professor with Digital Media Laboratory, Sungkyunkwan University, Seoul, Korea, before joining Xi'an Jiaotong–Liverpool University (XJTLU), Suzhou, China, in 2008. He was promoted to Full Professor in 2012. From 2010 to 2013, he was the Head of the Department of Electrical and Electronic Engineering at XJTLU, where he was the Acting Head of the Department of Computer Science and Software Engineering from 2012 to 2013. His research interests include robust transmission of multimedia data, image and video compression, and hyperspectral image compression.

Jimin Xiao received the B.S. and M.E. degrees in telecommunication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2007, respectively, and the dual Ph.D. degree in electrical engineering and electronics from University of Liverpool, Liverpool, U.K., in 2013.

In 2013, he served as a Visiting Researcher at Nanyang Technological University, Singapore. From 2013 to 2014, he was a Senior Researcher with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and an External Researcher with the Nokia Research Center, Tampere. He is currently a Lecturer with Xi'an Jiaotong Liverpool University, Su Zhou, China. His research interests include video streaming, image and video compression, and multiview video coding.