

Multiple Description Video Coding Based on Hierarchical B Pictures Using Unequal Redundancy

Wen-Jiin Tsai, *Member, IEEE*, and Hao-Yu You

Abstract—Multiple description video coding (MDC) is one of the approaches for reducing the detrimental effects caused by transmission over error-prone networks. In this paper, a MDC model based on hierarchical B pictures is proposed to optimize the tradeoff between coding efficiency and error resilience. The model produces two descriptors by applying different MDC techniques such as duplication, spatial splitting and temporal splitting on the different frames of video sequences, taking into account unequal importance of frames at different hierarchical levels. Duplication (high redundancy) is for key frames; spatial splitting (medium redundancy) for reference B frames, and temporal splitting (low redundancy) for nonreference B frames. For one descriptor loss, the model applies different estimation methods, but for the two descriptor loss case, the same temporal estimation is employed. As a consequence, better error resilience can be achieved at high coding efficiency. The advantages of the proposed model are demonstrated in error-free and packet loss networks.

Index Terms—Duplication, hierarchical B pictures, multiple description coding (MDC), spatial splitting, temporal splitting.

I. INTRODUCTION

THE DEMAND FOR transmitting video signals over wireless channels or over IP-based networks increases as bandwidth and storage of computer networks grow. Unfortunately, these environments are error-prone. During data transmission, packets may be dropped or damaged, due to channel errors, congestion, and buffer limitation. Moreover, the data may arrive too late to be used in real-time applications. In the case of transmission of compressed video sequences, this loss may be devastating and result in a completely damaged stream at the decoder side. For real-time applications, since retransmission is often not acceptable, error resilience (ER) and error concealment (EC) techniques are required for displaying a pleasant video signal despite the errors and for reducing distortion introduced by error propagation.

Several ER methods have been developed, such as forward error correction [1], intra/intercoding mode selection [2], layered coding [3], and multiple description coding (MDC) [4].

Manuscript received July 15, 2010; revised November 2, 2010; accepted December 19, 2010. Date of publication December 13, 2011; date of current version February 8, 2012. This work was supported in part by grants from the National Science Council of Taiwan, under Contract NSC 99-2221-E-009-139. This paper was recommended by Associate Editor D. S. Turaga.

W.-J. Tsai is with the Department of Computer Science, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: wjtsai@cs.nctu.edu.tw).

H.-Y. You is with the Home and Health Business Unit, Quanta Computer, Inc., Taoyuan 33377, Taiwan.

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.2011.2179450

This paper is concerned with MDC. MDC is a technique that encodes a single video stream into two or more equally important substreams, called *descriptions*, each of which can be decoded independently. Unlike the traditional single description coding (SDC), where the entire video stream (single description) is sent in one channel, in MDC, these multiple descriptions are sent to the destination through different channels, resulting in much less probability of losing the entire video stream (all the descriptions), where the packet losses of all the channels are assumed to be independently and identically distributed. The first MD video coder, called multiple description scalar quantizer [5], was realized in 1993 by Vaishampayan who proposed an index assignment table that maps a quantized coefficient into two indices each could be coded with fewer bits. Due to effectiveness in providing error resilience, a variety of research on different MDC approaches had been proposed afterward. These approaches can be intuitively classified through the stage where it split the signal, such as frequency domain [5], [6], spatial domain [7], [8], and temporal domain [9], [10]. In our previous work [11], a hybrid MDC method has been proposed, which applies MDC first in spatial domain to split motion compensated residual data, and then in frequency domain to split quantized coefficients. The results in [11] show that, by properly utilizing more than one splitting technique, the hybrid MDC method can improve error-resilient performance.

Although a variety of MDC approaches have been proposed, most of them were built upon conventional H.264/AVC coding structure and did not utilize hierarchical B-picture prediction. In a hierarchical B-picture prediction framework, the B frames at the coarser temporal levels can be used as a reference for the B frames at the finer temporal levels, and therefore the coding efficiency can be further improved. Compared with classical H.264/AVC prediction structure IBBP, the improvement can be more than 1 dB as described in [12]. Even though hierarchical-B picture coding has been widely used in scalable extension of H.264/AVC [13] to provide temporal scalability, it is rarely adopted in multiple description coding. In [14], an MDC based on hierarchical B pictures was proposed, where two descriptions are generated by duplicating the original sequence and then coded by hierarchical B structures with staggered key frames in the two descriptions. By using different QPs at different levels, their approach enables each frame to have two different quality fidelities in different descriptions. When two descriptors are received, their approach simply selects the frame with high-fidelity, or uses a linear combination of

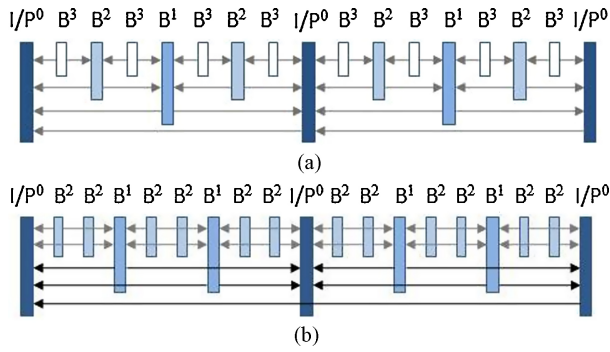


Fig. 1. Hierarchical B-picture prediction structure. (a) Dyadic. (b) Nondyadic.

the high-fidelity and low-fidelity frames to generate a better reconstruction. When only one descriptor is received, the lost frame is recovered by copying from the corresponding frame in the other descriptor. It can be seen that although their MDC approach employs hierarchical B-pictures to improve coding efficiency, it still suffers from high bit-rate redundancy by duplicating the original sequence to two descriptions.

This paper presents a MDC based on hierarchical B-pictures. Our approach employed duplication, spatial splitting, and temporal splitting for the frames at different hierarchical levels to provide unequal redundancy to frames with different fidelity requirements. In [11], we have shown that MDC performance can be improved by taking advantage of combining different MDC techniques. Compared to [11], this paper is significantly different in that a hierarchical B-picture structure is used and unequal redundancy is considered in error protection.

II. PROPOSED MDC METHOD

A. Hierarchical B-Picture Coding

A typical hierarchical prediction framework with four dyadic hierarchy stages is illustrated in Fig. 1(a), where the key frames (which can be I or P frames) are coded in regular intervals. A key frame and all frames that are temporally located between the key frame and the previous key frame form a group of pictures (GOP). The remaining B frames are hierarchically predicted using two reference frames from the nearest neighboring frames of the previous temporal level. In Fig. 1, B^i denotes the B frames at level i . It should be noted that the usage of hierarchical coding structure is not restricted to be the dyadic case. Fig. 1(b) shows the example of a nondyadic hierarchical structure with three levels.

For the optimized encoding, it is better to set smaller QPs for the frames that are referenced by other frames. In the joint scalable video model 11 (JSVM11) [15], QPs of the B frames at level-1 are equal to the QPs of the I/P frames plus 4, and the QPs at level- i increase by 1 from level- $(i-1)$, with $i \geq 2$.

B. Unequal Redundancy Based MDC

In a hierarchical B-picture prediction framework, the frames at lower hierarchical levels can be used as a reference for the frames at higher hierarchical levels. Due to this dependence, the decoding quality of a frame strongly depends on the quality

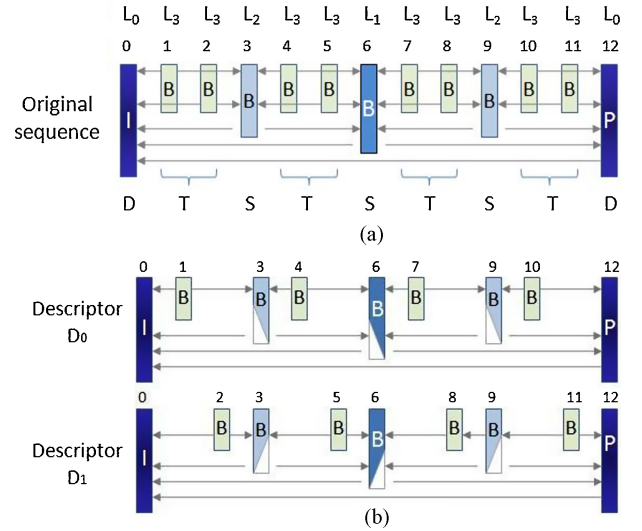


Fig. 2. Proposed MDC based on hierarchical B-picture prediction. (a) Original sequence. (b) Resulting two descriptors.

of the frames at its previous hierarchical level of the same GOP. The frame lost at the lower level will result in more corrupted frames. As an example in Fig. 1(a), the loss of an I or P picture will directly affect seven other frames, while the loss of a level-1, level-2, and level-3 B-pictures will directly affect 4, 2, and 0 other frames, respectively. Based on this observation, the proposed MDC aims at providing unequal redundancy for the hierarchical B pictures, taking into account the unequal importance of the frames at different hierarchical levels.

The proposed MDC model is illustrated in Fig. 2, where a nondyadic hierarchical B-picture structure with four levels is used. We refer to the I/P frames at the lowest hierarchical level as key frames; the B frames at intermediate levels as reference B (RB) frames because they are used as reference; and the B frames at the highest level as nonreference B (NRB) frames because they are not used as reference. As Fig. 2(a) shows, we apply duplication (denoted by D) on key frames for providing the highest error resilience; spatial-splitting (S) on RB frames for modest error resilience; and temporal-splitting (T) on NRB frames for the lowest error resilience. The resulting two descriptions are illustrated in Fig. 2(b), where the rectangles with a missing corner represent incomplete frames (due to spatial splitting). It can be seen that, due to different MDC methods applied, the frames at different hierarchical levels have unequal redundancy to provide robustness again errors. Assuming that description D0 is lost, the lost key-frames (0 and 12) can be easily reconstructed at decoder by using the same frames in description D1. The partially lost level-1 and level-2 frames (3, 6, and 9) can be estimated by using the information of their counterparts in description D1, while the lost level-3 frames (1, 4, 7, and 10), which are not in D1, can only be estimated by using other frames. The estimation methods will be discussed later in the next section.

The encoder architecture of the proposed MDC model is depicted in Fig. 3, where after intraprediction or motion compensation, there are three paths for three different kinds of frames: key frame, RB frame, and NRB frame. Key frames will go through transform, quantization, and entropy coding

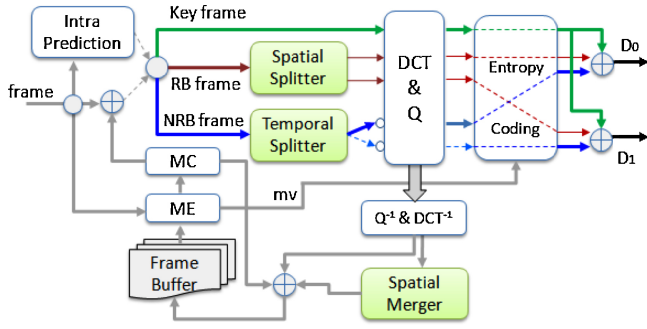


Fig. 3. Encoder architecture of proposed MDC.

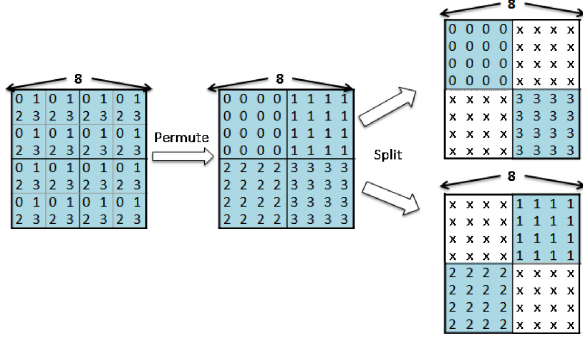


Fig. 4. Spatial splitting of the proposed MDC.

stages before it is duplicated to two descriptions. NRB frames will go to a temporal splitter which assigns the input frames, in turn, to the two output paths such that successive NRB frames will go to different descriptors. RB frames will enter a spatial splitter which splits each input frame into two parts which are then separately transformed, quantized, and entropy encoded before going to their respective descriptors. The spatial splitter performs splitting on an 8×8 block basis in the residual domain. For each 8×8 residual block, it is first *polyphase permuted* inside the block and then is split into two, as shown in Fig. 4. The permuting mechanism is that, for every 2×2 pixels inside the 8×8 residual block, the top-left pixel (labeled 0) is re-arranged to the top-left 4×4 block, the top-right pixel (labeled 1) to the top-right 4×4 block, the bottom-left pixel (labeled 2) to the bottom-left 4×4 block, and the bottom-right pixel (labeled 3) to the bottom-right 4×4 block, as illustrated in the middle of Fig. 4. After polyphase permutation, the 8×8 block is split into two 8×8 blocks, each carries two 4×4 blocks chosen in diagonal and the remaining two 4×4 blocks are given all-zero residuals (labeled as “x” in Fig. 2). Note that there are four 8×8 residual blocks in each macroblock, all of them are permuted and split in the same way. Since these split frames need to be merged to serve as reference frames, a spatial merger is applied after de-quantization (Q^{-1}) and inverse transform (DCT^{-1}) as shown in Fig. 3. The spatial merger first discards the all-zero 4×4 blocks and then adopts *polyphase inverse permuting* (the reversed process of Fig. 4) to reconstruct the original 8×8 blocks.

The decoder architecture of the proposed MDC model is depicted in Fig. 5, where the two descriptors, D_0 and D_1 , are first entropy decoded, dequantized, and inversely transformed

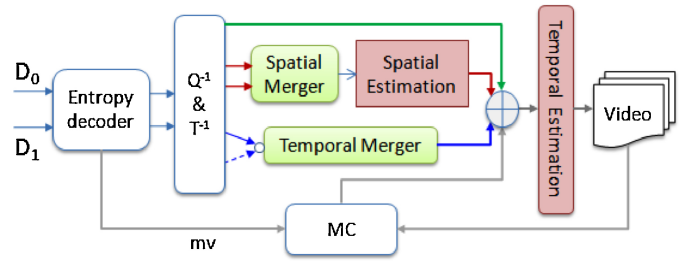


Fig. 5. Decoder architecture of proposed MDC.

TABLE I

SUMMARY OF THE CASES FOR DIFFERENT ESTIMATION METHODS

Estimation Methods	Descriptor Status	
	One-Descriptor Loss	Two-Descriptor Loss
Key frame	D	T
RB frame	S	T
NBR frame	T	T

separately; then a *spatial merger* and a *temporal merger* are applied to RB and NRB frames, respectively. The spatial merger is used to merge two complementary RB frames into a full RB frame. It is performed in the same way as the spatial merger in the encoder side. The temporal merger is used to reconstruct the order of NRB frame for output sequence. If the decoder does not receive the two descriptors intact, then either spatial or temporal estimation will be adopted to reconstruct the lost data.

III. ESTIMATION OF LOST DESCRIPTION

Taking advantage of different MDC methods applied on the frames at different hierarchical levels, different estimation methods are designed for different frames. Table II summarizes the cases for different estimation methods to be applied, where S denotes the spatial method, T the temporal method, and D the duplication method. The columns describe the two loss cases, while the rows describe three types of frames.

A. One-Descriptor Loss

In case of one-descriptor loss, since the lost key-frames can be reconstructed by simply using the duplicated version in the other descriptor, it is marked as **D** in Table I.

As for RB frames, since they are split into the spatial domain, one-descriptor loss only causes partial-frame loss. In this case, spatial method (marked as **S** in Table I) is applied to estimate the lost part. After the received descriptor has been entropy decoded, dequantized, and inversely transformed (see Fig. 5), the spatial merger will apply polyphase inverse permutation on the resulting data and then the residual pixels will be distributed like a checkerboard inside the macroblock as shown in Fig. 6, where each lost residual pixel has four available neighboring pixels. Our spatial estimation uses *bilinear interpolation* to reconstruct the lost residual pixels, as shown in (1) where $f'_{j,i}$ is the reconstructed value of the residual pixel in column i and row j . Since neighboring pixels

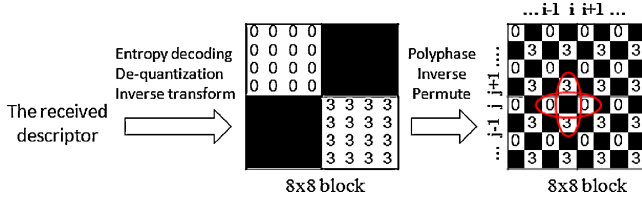


Fig. 6. Spatial concealment by bilinear interpolation.

have high spatial correlation, spatial estimation should be efficient

$$f'_{j,i} = (f_{j+1,i} + f_{j-1,i} + f_{j,i+1} + f_{j,i-1})/4. \quad (1)$$

As for NRB frames, one-description loss will result in whole frame loss because they are split in the temporal domain. In this case, a temporal estimation method (marked as **T** in Table I) is applied to reconstruct the lost frame. Since the temporal method is also adopted for all types of frames in case of two-description loss, we describe it in the next subsection.

B. Two-Descriptor Loss

In case of two-description loss, it will result in whole-frame loss regardless of frame types. For whole-frame loss, each block in the lost frame is recovered based on temporal correlation since all the neighboring blocks are also lost. We refer to the pictures whose pixels are used to predict the missing pixels as the data prediction frame (DF) and the pictures whose block motions are used to predict the motion of the missing blocks as the motion prediction frame (MF). In our method, DF can be different from MF. Besides, the proposed methods adopt a bidirectional motion-compensated signal to recover missing pixels. Thus, we need to select two DFs: a *backward DF* and a *forward DF* (denoted by $\overleftarrow{\text{DF}}$ and $\overrightarrow{\text{DF}}$, respectively); and two MFs: a *backward MF* and a *forward MF* (denoted by $\overleftarrow{\text{MF}}$ and $\overrightarrow{\text{MF}}$, respectively) for a lost picture. Since the data correlation among pictures involved tends to considerably weaken as the temporal distances among these pictures become longer, for a lost picture, it is better to choose the nearest pictures in display order to serve as its DFs. However, to serve as DFs requires that these pictures are decoded earlier than the lost picture. Based on the hierarchical B-picture structure, for a lost picture, we select its reference frames in backward and forward directions as its $\overleftarrow{\text{DF}}$ and $\overrightarrow{\text{DF}}$, respectively.

As for MFs, they are selected differently from DFs. In case of frame loss, even though the frames later than the lost frame (in decoding order) cannot be decoded before the lost frame is recovered, the motion information of these frames is obtainable. Therefore, the MFs need not be located earlier than the lost picture in decoder order. Instead of using the temporal direct mode (TDM) technique which adopts reference pictures as MFs, we choose pictures at higher levels because these pictures are temporally nearer to the lost picture in display order. As an example in Fig. 7(a), if the frame 6 is lost, we will select its reference frames (0 and 12) as its DFs, but frames 3 and 9 as its MFs. In Fig. 7(a), if frame 3 is lost, we will select its reference frames 0 and 6 as DFs, but frames 2 and 4 as MFs. This selection policy is applied to all frames except NRB frames which are at the highest level within the

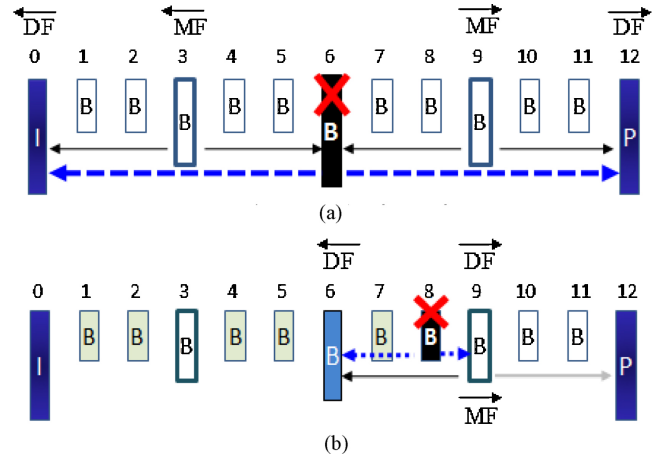


Fig. 7. DF and MF selection for temporal estimation method. (a) DF and MF selection for RB frames. (b) DF and MF selection for NRB frames.

hierarchical structure. For NRB frames, the MFs are selected from their reference frames at the previous level of the lost picture. Fig. 7(b) illustrates the case of NRB frame loss, where frame 8 is the lost frame. In this case, frames 6 and 9 will serve as the DFs, and frame 9 (which is at previous level of frame 8) will serve as the MF. Similarly, if frame 10 is lost, its DFs will be frames 9 and 12, and its MF will be frame 9. Specifically, for the lost picture F_t^l at time instant t with hierarchical level l , we select its $\overleftarrow{\text{MF}}$ and $\overrightarrow{\text{MF}}$ as

$$\overleftarrow{\text{MF}} = \begin{cases} F_{t_{nb}}^{l+1}, & \text{for } l_{\text{base}} \leq l < l_{\text{top}} \\ F_{t_{\text{ref}}}^{l-1}, & \text{if } F_{t_{\text{ref}}}^{l-1} \text{ exists for } l = l_{\text{top}} \end{cases} \quad (2)$$

$$\overrightarrow{\text{MF}} = \begin{cases} F_{t_{nf}}^{l+1}, & \text{for } l_{\text{base}} \leq l < l_{\text{top}} \\ F_{t_{\text{ref}}}^{l-1}, & \text{if } F_{t_{\text{ref}}}^{l-1} \text{ exists for } l = l_{\text{top}} \end{cases} \quad (3)$$

where l_{base} denotes the base level (key-frame level) and l_{top} the top level (NRB-frame level). $F_{t_{nb}}^{l+1}$ and $F_{t_{nf}}^{l+1}$ denote F_t^l 's nearest backward and forward frames at level $l+1$, respectively. $F_{t_{\text{ref}}}^{l-1}$ denotes the F_t^l 's reference frame at level $l-1$.

After determining DFs and MFs, the motion vectors in MFs will be used to estimate the missing motion vectors (pointing to DFs from the lost frame). When the lost frame is a RB frame, since its MFs are located in between DFs and the lost frame [see Fig. 7(a)], the motion vectors are composed if the block in MFs has two motion vectors, or extrapolated if the block has only one motion vector. The motion vector derivation corresponding to Fig. 7(a) is illustrated in Fig. 8(a), where the two motion vectors of b_x in f_3 are composed and the motion vector of b_y is extrapolated, so that the derived motion vectors will point to f_0 from f_6 . The motion vectors pointing to f_{12} from f_6 can also be derived in a similar manner using motion vectors of f_9 . On the other hand, when the lost frame is a NRB frame, since one MF is used for two DFs located on different sides of the lost picture [see Fig. 7(b)], the motion vectors in the MF are interpolated as illustrated in Fig. 8(b) where the motion vector of block b_w is interpolated to obtain two motion vectors respectively pointing to $f_{6\text{and}9}$ from f_8 . Let $\overleftarrow{mv}$ and $\overrightarrow{mv}$ denote the derived motion vectors pointing to $\overleftarrow{\text{DF}}$ and $\overrightarrow{\text{DF}}$ from the lost frame, respectively. For a lost frame, after all the motion vectors in its MFs have been

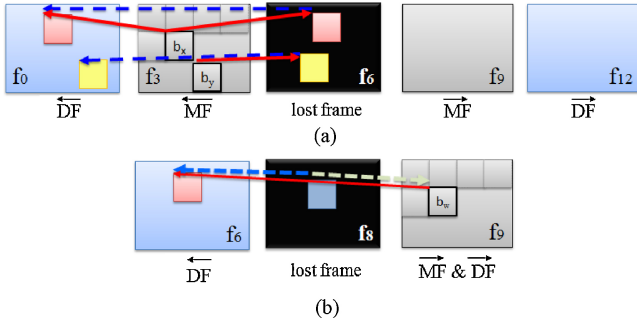


Fig. 8. Temporal estimation using bidirectional predicted signal. (a) Motion composition and extrapolation for a lost RB frame. (b) Motion interpolation for a lost NRB frame.

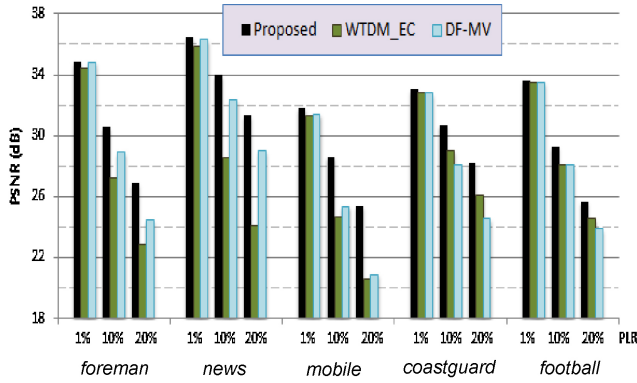


Fig. 9. Performance of temporal estimation methods (QP = 28).

composed, extrapolated, or interpolated, the missing pixels on the lost frame can be classified into four types: the pixels associated with one or more $\vec{m}\vec{v}$, the pixels with one or more $\vec{m}\vec{v}$, the pixels with both $\vec{m}\vec{v}$ and $\vec{m}\vec{v}$, and the pixels without $\vec{m}\vec{v}$ and $\vec{m}\vec{v}$. For a pixel P in the lost picture, we recover it by the predicted signal $\tilde{P}$ obtained as follows:

$$\tilde{P}(x) = \begin{cases} \sum_i \vec{D}\vec{F}(x + \vec{m}\vec{v}_i), & \text{if } P \text{ has } \vec{m}\vec{v} \text{ only} \\ \sum_i \vec{D}\vec{F}(x + \vec{m}\vec{v}_i), & \text{if } P \text{ has } \vec{m}\vec{v} \text{ only} \\ w_0 \sum_i \vec{D}\vec{F}(x + \vec{m}\vec{v}_i) + w_1 \sum_i \vec{D}\vec{F}(x + \vec{m}\vec{v}_i), & \text{if } P \text{ has } \vec{m}\vec{v} \text{ and } \vec{m}\vec{v} \\ w_0 \vec{D}\vec{F}(x) + w_1 \vec{D}\vec{F}(x), & \text{otherwise.} \end{cases} \quad (4)$$

Here, x is a spatial coordinate of P . w_0 and w_1 are the weighting values, which are set in inverse proportion to the temporal distances of $\vec{D}\vec{F}$ and $\vec{D}\vec{F}$, respectively, from the lost picture.

IV. EXPERIMENTAL RESULTS

In this section, the performance of the proposed temporal estimation method is examined first, and then the proposed MDC model is examined by its packet-loss performance, center-decoder performance, and the error propagation effect.

A. Performance of Temporal Estimation Method

To show the performance of the proposed temporal estimation method, experiments were conducted for four CIF

(352 × 288) test sequences: *Foreman*, *Mobile*, *Coastguard*, and *News*, each of which was encoded using a dyadic hierarchical structure with four levels. We compared the proposed method with DF-MV and WTDM_EC [16]. DF-MV is a variation of the proposed temporal estimation method. It estimates the missing motion vectors by adopting motion vectors of DFs (which are the frames used to predict the missing pixels), rather than using motion vectors of the frames in the next hierarchical level. WTDM_EC is a method revised from the TDM of H.264/AVC for the error concealment of whole frame loss in a hierarchical B-picture prediction structure. Both DF-MV and WTDM_EC are implemented by modifying H.264 reference software, JM 16.0 [17]. Three different loss rates (PLRs) are used in our experiments and the results presented in Fig. 9 are the averages of 100 independent simulation runs.

From the results it is observed that the proposed temporal estimation outperformed DF-MV and WTDM_EC for all the sequences under different packet loss rates. The performance gap becomes large as the loss rate increases. Since the motion vectors that DF-MV adopted point to the frames that are far in temporal distance, it is unreliable to have accurate motion vector derivation which is based upon the assumption that object motion is linear. Compared to DF-MV, the proposed method adopted motion vectors that point to relatively nearer frames and thus, more accurate motion vectors can be derived. As for WTDM_EC, it predicts missing motion vector by using the motion vector of co-located block in the selected picture. Such prediction is effective when the two pictures are located closely in the sequences; however, it might not work well for the pictures at lower levels in the hierarchical B-picture structure because these pictures are located far apart in the display order. This can be illustrated by the example in Fig. 10, where the subjective quality comparison of concealed frames at different hierarchical levels in *News* and *Football* sequences is presented. Experiments were conducted independently for each case in Fig. 10, namely, there is no error propagation implemented among them.

From Fig. 10(a) it is observed that visual qualities of three concealed *News* frames at level 3 are almost the same. However, there is an obvious quality improvement by using the proposed method for the concealed frames at levels 2, 1, and 0 (key-level), as shown in Fig. 10(b)–(d). For the frames at levels 2, 1, and 0, there is a noticeable noise on the dancer by using WTDM_EC; and the frames concealed by DF-MV are also obviously blurred. Similar results can also be found for motion intensive sequence such as *Football*, as shown in Fig. 11(a)–(d). Compared with WTDM_EC, the quality improvement by using the proposed method can be up to 2.98 dB for *News* and 3.04 dB for *Football* sequences; and compared with DF-MV, the improvement can be up to 4.86 dB for *News* and 5.57 dB for *Football* sequences.

B. Performance of Proposed MDC

In this section, the proposed MDC model is examined. To see the effects of different MDC techniques adopted in our model, experiments were conducted for three variations of proposed MDC model: *proposed* (S), *proposed* (S+T), and *proposed* (S+T+D). The proposed (S) stands for the

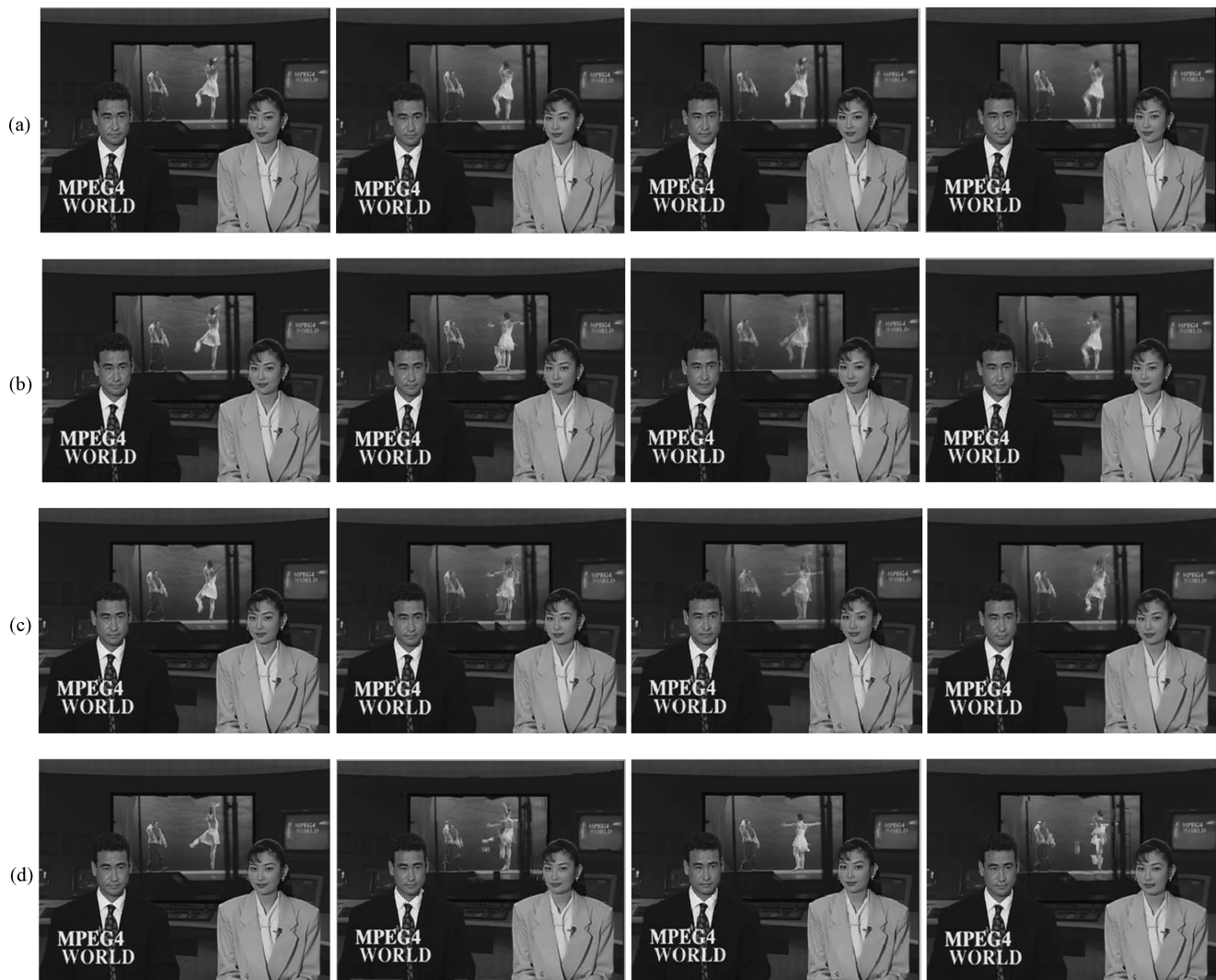


Fig. 10. Subjective quality comparison of the frames at different hierarchical levels of *News* sequence. (a) Subjective quality of the sixth frame (NRB) at level 3. Left to right: correct (36.68 dB), WDTM_EC (34.09 dB), DF-MV (34.29 dB), proposed (34.29 dB). (b) Subjective quality of the seventh frame (RB) at level 2. Left to right: correct (37.53 dB), WDTM_EC (33.88 dB), DF-MV (33.02 dB), proposed (35.51 dB). (c) Subjective quality of the fifth frame (RB) at level 1. Left to right: correct (37.42 dB), WDTM_EC (31.33 dB), DF-MV (29.41 dB), proposed (34.27 dB). (d) Subjective quality of the ninth frame (P) at level 0. Left to right: correct (38.44 dB), WDTM_EC (27.7 dB), DF-MV (29.55 dB), proposed (30.25 dB).

method which adopts spatial splitting only. It applies spatial splitting in the residual domain for all frames, regardless of hierarchical levels. The proposed (S+T) stands for the method which adopts two kinds of splitting: temporal splitting for top-level frames (i.e., NRB frames) and spatial splitting for others. The proposed (S+T+D) stands for the full version of proposed method, which adopts temporal splitting for top-level NRB frames, spatial splitting for RB frames, and duplication for base-level key frames. We compare our three methods with the method of Zhu *et al.* [14], which generates two descriptors by duplicating the original sequence and then coded by hierarchical B structure with staggered key frames in the two descriptions as shown in Fig. 12, where B^i denotes the B frame at level i . This approach is characterized by that each frame at level 0, 1, or 2 of description 1 will be at level 3 of description 2 and vice versa, resulting in two fidelities of each frame in two descriptions. Two variations, *default_QP* and *modified_QP*, in their literature are adopted in our comparison. The *default_QP* follows the QP assignment

rules specified in JSVM11 [15] as described in Section II, while the *modified_QP* modifies the QPs of top-level frames to 51 in order to reduce bitrate redundancy. The results in [14] show that rate-distortion performance of center decoder can be improved remarkably by *modified_QP* in comparison to *default_QP*. In this section, their packet-loss performances are also examined. Table II lists the error concealment methods used by these MDC methods, where D' means the error concealment method in [14], where in case of one-descriptor loss, the lost frame is recovered by the duplicated version in the other description. D' is distinguished from D because the duplicated frame is at the same level in our approach, but at a different level in Zhu *et al.*'s approach. Since Zhu *et al.* did not provide solutions for two-description loss, our temporal estimation method is adopted for fair comparison. The five MDC methods are implemented based on H.264 reference software, JM 16.0 [17].

All the methods encode video sequences using hierarchical B-picture structure of four levels to generate two descriptors.

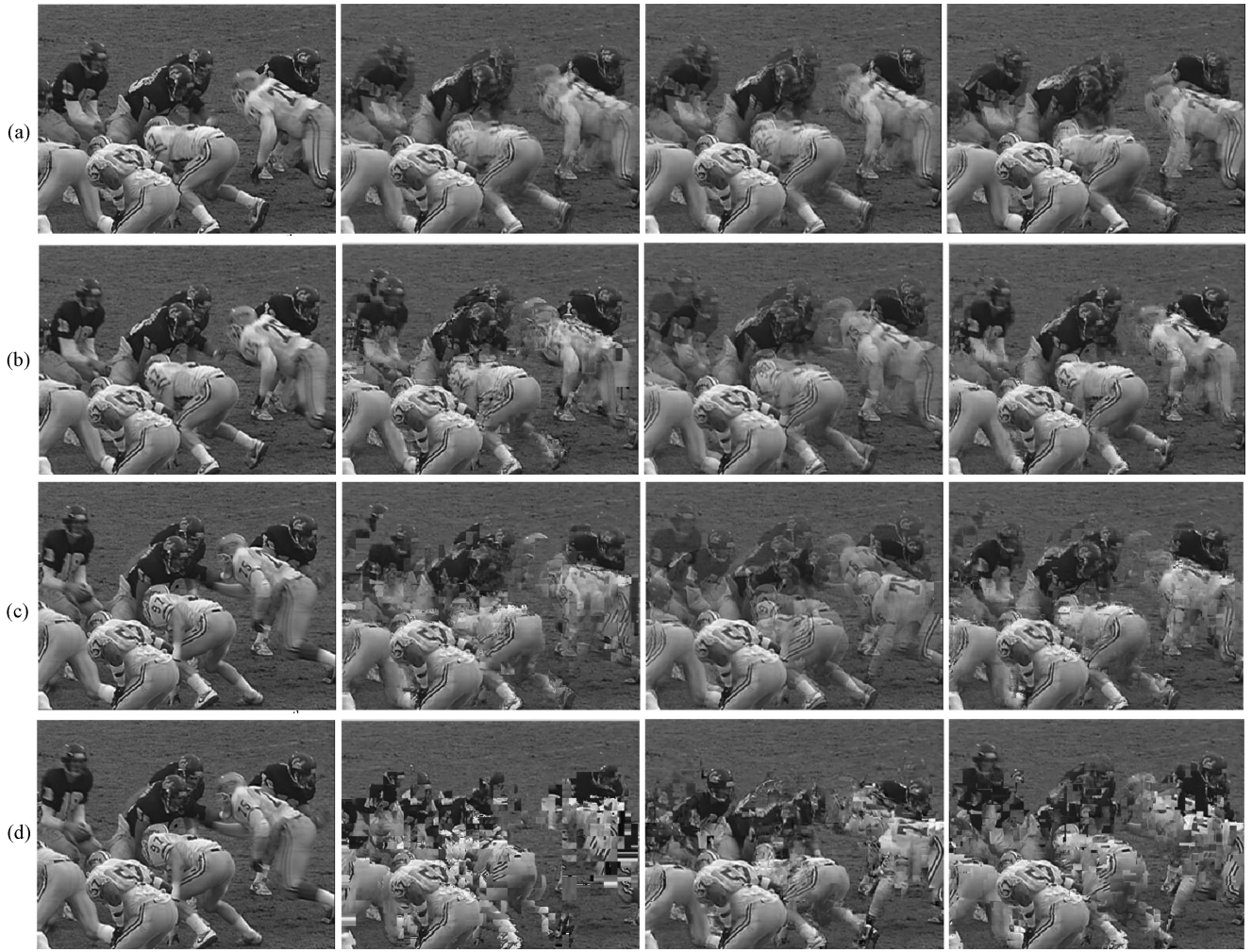


Fig. 11. Subjective quality comparison of the frames at different hierarchical levels of *News* sequence. (a) Subjective quality of the sixth frame (NRB) at level 3. Left to right: correct (32.43 dB), WDTM_EC (20.76 dB), DF-MV (20.83 dB), proposed (20.83 dB). (b) Subjective quality of the seventh frame (RB) at level 2. Left to right: correct: (32.88 dB), WDTM_EC (21.45 dB), DF-MV (18.92 dB), proposed (24.49 dB). (c) Subjective quality of the fifth frame (RB) at level 1. Left to right: correct (32.87 dB), WDTM_EC (18.82 dB), DF-MV (17.54 dB), proposed (21.33 dB). (d) Subjective quality of the ninth frame (P) at level 0. Left to right: correct (35.07 dB), WDTM_EC (15.26 dB), DF-MV (14.44 dB), proposed (17.61 dB).

Frame No.	1	2	3	4	5	6	7	8	9	10
Description 1:	I	B ²	B ²	B ²	B ¹	B ²	B ²	B ²	P	...
Description 2:		I	B ³	B ²	B ³	B ¹	B ³	B ²	B ³	P

Fig. 12. Two descriptions with staggered key frames [14].

The three proposed methods adopt a nondyadic structure which allows temporal splitting on NRB frames as depicted in Fig. 2; while Zhu *et al.*'s two methods adopt a dyadic structure which ensures that each frame has two different fidelities in the two descriptions. To know how the performance might be affected by different hierarchical structures, Fig. 13 shows rate-distortion performance of conventional SDC with these two different structures. As observed in Fig. 13, the two structures perform equally well for low-motion sequences, *Foreman* and *News*; while the dyadic structure outperforms the nondyadic one slightly for high-motion sequences, *Mobile* and *Coastguard*. The summary results of Fig. 13 are included in Table III, where Bjontegarrd bit rate savings (BD-rate) and

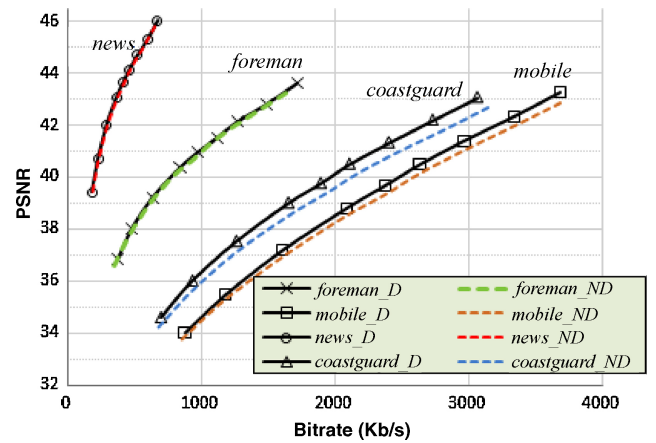


Fig. 13. R-D performance of dyadic (D) and nondyadic (ND) structures.

PSNR gains (BD-PSNR) are calculated using the methodology presented in [18]. The result in Table III shows that the nondyadic structure obtains a relatively worse performance (regardless of bit-rate savings or PSNR gains) than the dyadic

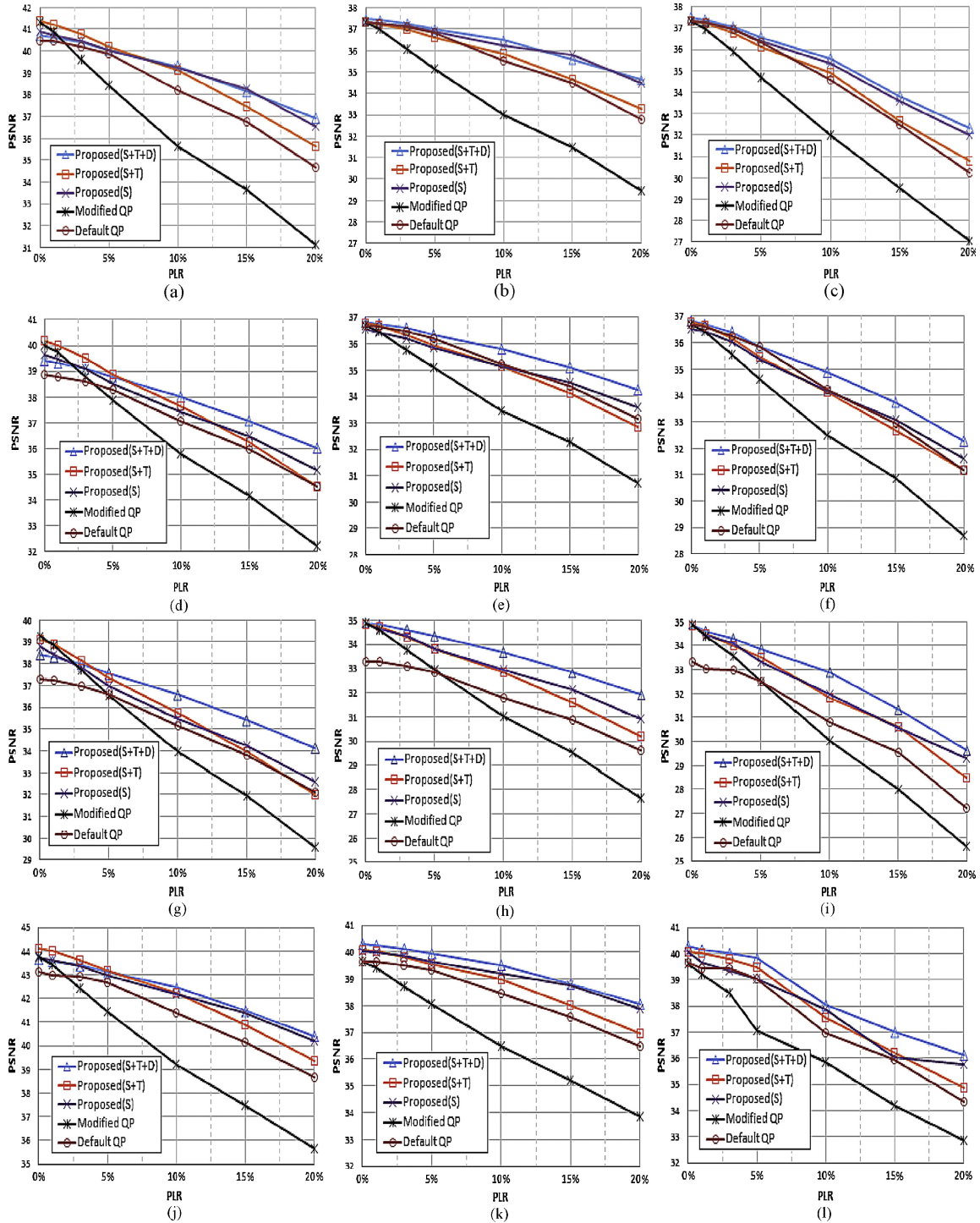


Fig. 14. Performance comparison in packet-loss environments. (a) *Foreman* (CIF, $R_{2D} = 1500$ kb/s, Bernoulli-ch.). (b) *Foreman* (CIF, $R_{2D} = 700$ kb/s, Bernoulli-ch.). (c) *Foreman* (CIF, $R_{2D} = 700$ kb/s, Gilbert-ch.). (d) *Coastguard* (CIF, $R_{2D} = 2800$ kb/s, Bernoulli-ch.). (e) *Coastguard* (CIF, $R_{2D} = 1800$ kb/s, Bernoulli-ch.). (f) *Coastguard* (CIF, $R_{2D} = 1800$ kb/s, Gilbert-ch.). (g) *Mobile* (CIF, $R_{2D} = 2800$ kb/s, Bernoulli-ch.). (h) *Mobile* (CIF, $R_{2D} = 1500$ kb/s, Bernoulli-ch.). (i) *Mobile* (CIF, $R_{2D} = 1500$ kb/s, Gilbert-ch.). (j) *News* (CIF, $R_{2D} = 700$ kb/s, Bernoulli-ch.). (k) *News* (CIF, $R_{2D} = 400$ kb/s, Gilbert-ch.).

structure, indicating that the three proposed methods are based on a slightly worse structure for the following comparison.

1) *Packet Loss Performance*: The five MDC methods were examined in a packet-loss scenario where various packet-loss rates, ranging from 0% to 20%, are adopted. We use one packet for each frame of each descriptor. Two different simulated channels, Bernoulli and Gilbert [19], are used, where the

Bernoulli channel assumes that packets lost are independent of each other; while the Gilbert channel captures the temporal dependence between lost packets, and thus is a widely-used model to simulate burst packet losses. The average burst length adopted in our experiments is 4, which is based on the results in [20], where the loss run lengths fall in between 1 and 2 with very high probability and seldom exceed 5. Fig. 14

shows the results of the two channel models for four CIF sequences, *Foreman*, *Coastguard*, *Mobile*, and *News*. The R_{2D} in Fig. 14 denotes the total bit-rate of two descriptions. For each sequence, two kinds of R_{2D} are used. In each case of Fig. 14, the five methods encode the sequence using the same R_{2D} for fair comparison. The results are the averages of 100 independent runs. It can be seen that, in the case of $PLR = 0\%$, modified_QP and proposed (S+T) have the best performance and default_QP has the worst performance among all methods. This is due to the fact that the default_QP duplicates the entire sequence to two descriptions and therefore suffers from considerable bit-rate redundancy. By providing poorer picture quality at the lowest level, the modified_QP can effectively reduce the bitrate and thus achieve a better performance at $PLR = 0$. We will discuss the error-free performance in the next section. As PLR increases, however, the modified_QP curves drop much more quickly than others for all sequences, showing that the poorer quality at the lowest level will strongly affect error-concealment effectiveness and thus, degrade the performance. Compared with modified_QP, default_QP performed much better as PLR increases. However, the duplication method used in default_QP still cannot avoid quality degradation in recovering lost frames because the same frames in two descriptions are at different hierarchical levels with different fidelities. The degraded error-concealment performance and the high bitrate-redundancy result in the worse performance of default_QP, compared with the three proposed methods.

Among these methods, proposed (S+T+D) has the overall best performance. Although proposed (S) performed slightly better than proposed (S+T+D) for *Foreman* sequence, it performed much worse than proposed (S+T+D) for sequences *Mobile* and *Coastguard*. This is due to the fact that spatial estimation cannot recover lost data well for these sequences when there is packet loss. With temporal splitting on NRB frames, proposed (S+T) reduces bit-rate redundancy and hence, improves the R-D performance at low PLR , but still cannot solve the problem for high PLR . By duplicating key frames, the proposed (S+T+D) can alleviate this problem effectively. When packet loss rate is low ($PLR < 5\%$), the proposed (S+T+D) performed equally to, or slightly worse than, proposed (S+T) and proposed (S). This stems from the fact that the scheme of key-frame duplication adopted in the proposed (S+T+D) cannot take much effect in error concealment when PLR is low. As the PLR increases, however, proposed (S+T+D) outperformed others noticeably. This is due to the fact that the key frames in proposed (S+T+D) can be recovered without quality loss once they are lost. Since key-frames have the maximum number of frames depending on it, the duplication of key-frame can suppress error propagation effectively and improve performance substantially. We will discuss the error propagation issue further in the later section. To summarize, the overall results demonstrate that, by adopting spatial splitting, temporal splitting and duplication for the frames at different levels, the proposed (S+T+D) optimizes the tradeoff between bit-rate redundancy and error-resilient capability and therefore, achieves the best performance among the five MDC methods.

2) *Error Free Performance*: In this section, we compare the performance of the five MDC methods and SDC

TABLE II
SUMMARY OF THE CASES FOR DIFFERENT ESTIMATION METHODS

Estimation Method	Proposed (S)		Proposed (S+T)		Proposed (S+T+D)		Default QP		Modified QP	
	one	both	one	both	one	both	one	both	one	both
Key frame	S	T	S	T	D	T	D'	T	D'	T
RB frame	S	T	S	T	S	T	D'	T	D'	T
NRB frame	S	T	T	T	T	T	D'	T	D'	T

TABLE III
BD-PSNR GAINS AND BD-RATE SAVINGS OF USING
NONDYADIC STRUCTURE

	BD-PSNR (dB)	BD-rate
<i>news</i>	-0.07	1.50%
<i>foreman</i>	-0.05	1.20%
<i>coastguard</i>	-0.37	7.30%
<i>mobile</i>	-0.23	3.80%

in error-free environments. Experiments were conducted for four sequences and the results are presented in Fig. 15. As expected, due to bit-rate redundancy, all the MDC methods have worse rate-distortion performance than SDC. Among MDC methods, default_QP produces noticeably lower PSNR values than others at the same bitrates. Both proposed (S) and proposed (S+T+D) perform slightly worse than modified_QP and proposed (S+T), but the performance gaps among them are insignificant. The results are strongly related to the bit-rate redundancy produced by each MDC method. Let BR denote the bit-rate redundancy of MDC methods. For a MDC method, M_i , its BR is given as

$$BR(M_i) = \frac{R_{2D}(M_i) - R_{SDC}}{R_{SDC}} \times 100\% \quad (5)$$

where R_{2D} is the total bitrates of two descriptions and R_{SDC} is the bit-rate for SDC with a hierarchical B-picture structure. Table IV shows the BR and PSNR produced by the five MDC methods on four CIF sequences with $QP = 28$. It can be seen that default_QP has the best PSNR performance among MDC methods. This is due to the fact that, for each frame in the sequence, default_QP adopts a combination of its high-fidelity and low-fidelity frames from the two descriptors to produce a better reconstruction with PSNR even higher than SDCs. The reason default_QP has higher PSNR than SDC stems from the fact that default_QP generates two descriptors by duplicating the original sequence and encoding each descriptor using hierarchical B-picture structure with key frames in staggered positions (see Fig. 12). Since frames at different hierarchical levels are coded with different QPs (as described in Section II), this approach is characterized by that each frame has two fidelities in two descriptors. When two descriptors are received, rather than simply discarding one whole descriptor, their approach selects each frame from the one with higher fidelity between the two descriptors. For example, frame 2 will be selected from descriptor 2, frame 3 from descriptor 1, frame 4 from descriptor 2, and so on. As a consequence, the

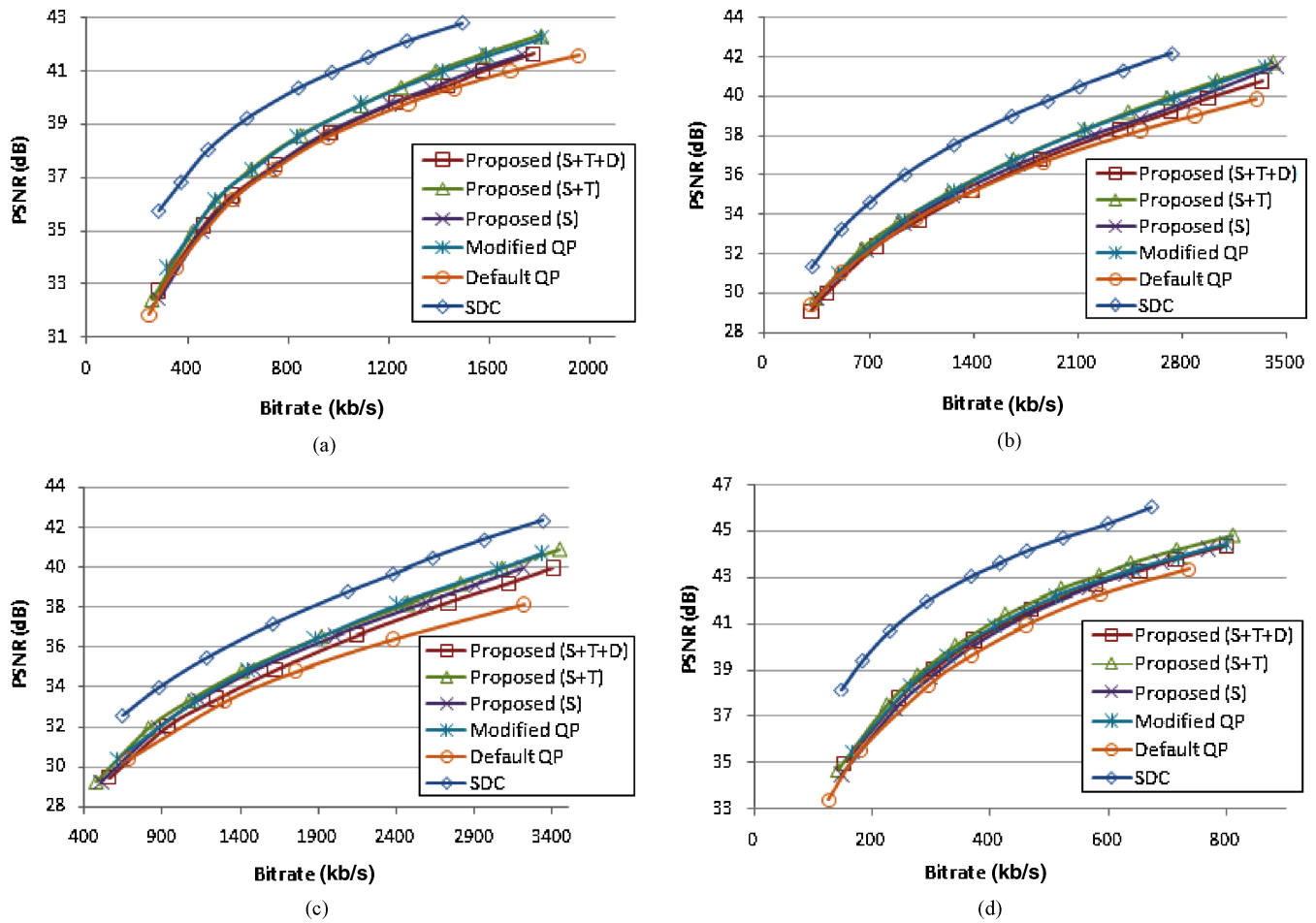


Fig. 15. Rate-distortion performance comparison in error-free environments. (a) *Foreman* sequence (CIF). (b) *Coastguard* sequence (CIF). (c) *Mobile* sequence (CIF). (d) *News* sequence (CIF).

resulting sequence has a higher PSNR than each individual descriptor; namely, it has a higher PSNR than SDC. However, even though default_QP achieves the best PSNR performance, it suffers from substantial BR increase because it duplicates the entire sequence to two descriptions and hence, its bitrate is almost twice the bitrate of SDC. In Fig. 15, the default_QP obtains the worst rate-distortion performance, showing that its gain in PSNR cannot compensate its loss in BR. By modifying the QPs of NRB frames, modified_QP reduces the BR about 20% while keeping the same PSNR as default_QP, as shown in Table IV. This explains why the modified_QP has the best rate-distortion performance in Fig. 15. The improvement of error-free performance is, however, at the cost of reducing the error robustness as shown in Fig. 14, where modified_QP has the worst packet-loss performance. By using splitting techniques, Table IV shows that the three proposed methods can further reduce BRs, but also decrease the PSNRs, resulting in slightly worse or equal rate-distortion performance, compared to modified_QP, as shown in Fig. 15. Besides, the reason that the proposed (S+T+D) cannot outperform proposed (S) or (S+T) in Fig. 15 is due to the fact that the scheme of key-frame duplication in the proposed (S+T+D) cannot take effect when there is no error. Compared to proposed (S) and (S+T), the proposed (S+T+D) suffers from relatively higher bit-rate

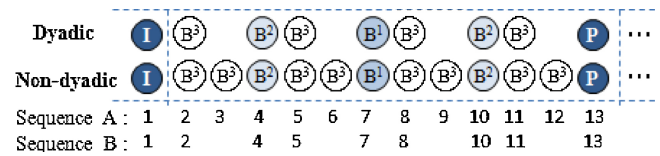


Fig. 16. Frames used in frame-by-frame comparison.

redundancy because it duplicates key-frames. This increase of bit-rate results in slightly worse R-D performance of the proposed (S+T+D) in error-free environments. However, the key-frame duplication did improve the error-resilient capability of proposed (S+T+D). As shown in Fig. 14, the performance gains of proposed (S+T+D), compared with proposed (S) and (S+T), increase obviously as the packet loss rate increases. To summarize, the overall results in Figs. 14 and 15 demonstrate that, with slight degradation in error-free performance, the proposed methods can improve packet-loss performance significantly, especially when the proposed (S+T+D) is employed.

3) *Error Propagation Effects*: This section presents the frame-by-frame comparison of error propagation effects using different MDC methods. The effects of error propagation were examined for a single frame loss occurring at different hierarchical levels of *Mobile* sequence at QP = 28. Since the proposed methods use a nondyadic hierarchy structure and

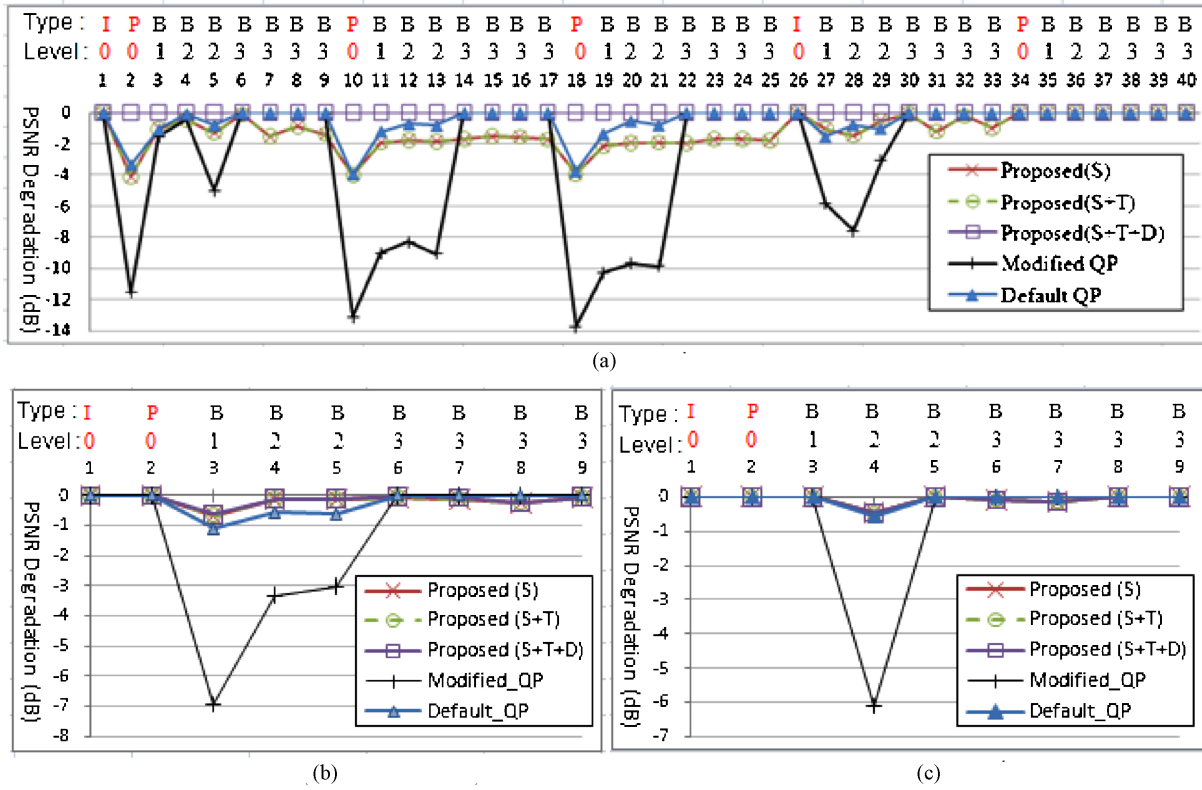


Fig. 17. Frame-by-frame comparison. (a) Frame loss at level 0 (the second frame is lost). (b) Frame loss at level 1 (the third frame is lost). (c) Frame loss at level 2 (the fourth frame is lost).

TABLE IV
ERROR-FREE PERFORMANCE OF MDC METHODS (QP = 28)

Sequences	Foreman		News		Mobile		Coastguard	
	BR	PSNR	BR	PSNR	BR	PSNR	BR	PSNR
SDC	0	35.73	0	38.13	0	32.57	0	33.22
Proposed (S+T)	45%	34.92	51%	37.51	26%	31.95	27%	32.22
Proposed (S+T+D)	60%	35.19	65%	37.78	45%	32.05	47%	32.38
Proposed (S)	57%	34.98	61%	37.36	36%	31.97	32%	32.21
Modified_QP	76%	36.16	79%	38.32	72%	33.31	83%	33.69
Default_QP	100%	36.16	100%	38.32	102%	33.3	98%	33.69

both default_QP and modified_QP use a dyadic structure, the same frame in different MDC methods may be at different levels. To have a fair comparison, some frames in the original sequence are removed for dyadic structure coding so that corresponding GOPs in dyadic and nondyadic structures will start from the same key frames. Let B^i denote the B frame at level i . As illustrated in Fig. 16, four frames at level 3 of the nondyadic structure are removed from the sequence when the dyadic structure is coded and therefore, the level-0, level-1, and level-2 frames in the two structures will be the same. This is applied to each GOP. In Fig. 16, sequence A shows the frame numbers in the original sequence, while sequence B lists the selected frames for frame-by-frame comparison in Fig. 17.

We use one packet for each frame in each descriptor and the error propagation results of a single packet loss at different hierarchical levels are shown in Fig. 17, where we renumber

the selected frames according to decoding order. Fig. 17(a)–(c) shows the results of the frame loss occurring at levels 0, 1, and 2, respectively. In Fig. 17, the y-axis denotes PSNR degradation and the x-axis the frame number (in decoding order). From Fig. 17(a) it is observed that almost all the methods suffer from severe error propagation for the P-frame loss, except the proposed (S+T+D). This is due to that the proposed (S+T+D) duplicates key-frames to two descriptors and therefore, when only one of them is lost, the other one can be used to reconstruct the frame without quality degradation and error propagation. In both proposed (S) and proposed (S+T) methods, key-frames are spatially split into two descriptors and hence, the P-frame loss in one descriptor will cause partial-frame loss which is recovered by using spatial estimation, suffering from quality degradation and error propagation. As for default_QP, although it duplicates the entire sequence to two descriptors, the same frames in the two descriptors are at different levels and thus, the lost key frame can only be recovered by the corresponding low quality frame in the other descriptor. This also results in quality degradation and error propagation. It is worth mentioning that even though the quality degradation of default_QP in Fig. 17(a) is smoother than those of proposed (S) and proposed (S+T), it is at the cost of bit-rate redundancy. That is why default_QP has worse packet-loss performance than proposed (S) and proposed (S+T) as shown in Fig. 14. Compared with default_QP, modified_QP suffers from much severe quality degradation because the top-level frame used to recover the lost key frame has been set to QP = 51 to reduce the bit-rate.

Compared with Fig. 17(a), the results in Fig. 17(b) and (c) show that when the frame loss occurs at level 1 or 2, the error propagation effects are substantially reduced for all the methods and the performance gaps between different methods are also decreased. Since NRB frames will not cause error propagation, the result for level 3 is not presented. To summarize, the results in Fig. 17 show that quality degradation and error propagation in the hierarchical prediction structure are affected by key frames most, and level-1 and level-2 frames the second. By taking into account the importance of frames at different levels, proposed (S+T+D) adopts duplication (high bit-rate redundancy) for key-frames; spatial splitting (low redundancy) for level-1 and level-2 RB frames; and temporal splitting (no redundancy) for NRB frames. As a result, the proposed (S+T+D) optimizes the tradeoff between coding efficiency and the error resilience and achieves the overall best performance.

V. CONCLUSION

A MDC model based on hierarchical B pictures is proposed. The model produces two descriptors by applying different MDC techniques such as duplication, spatial splitting and temporal splitting on frames at different hierarchical levels. Duplication is applied to the frames at base-level which is the most important one in the hierarchical structure; spatial splitting is applied to the frames at intermediate levels; and temporal splitting is applied to the frames at top-level which is the least important one in the structure. By taking account for importance of the frames in the hierarchical structure, the proposed model is able to optimize the tradeoff between coding efficiency and error resilience. In case of data loss, the model takes advantage of different estimation methods in providing different error resilience for the frames with different degrees of importance. Experiments were conducted for five MDC methods: three variations of the proposed model [proposed (S), proposed (S+T), and proposed (S+T+D)] and two methods (default_QP and modified_QP) in [14]. The experimental results show that the proposed (S+T+D) achieves the overall best performance among these five methods.

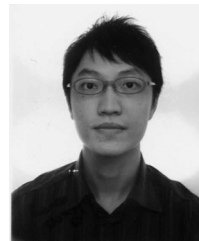
REFERENCES

- [1] A. Nafaa, T. Taleb, and L. Murphy, "Forward error correction strategies for media streaming over wireless networks," *IEEE Commun. Mag.*, vol. 46, no. 1, pp. 72–79, Jan. 2008.
- [2] R. Zhang, S. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 966–976, Jun. 2000.
- [3] C.-M. Fu, W.-L. Hwang, and C.-L. Huang, "Efficient post-compression error-resilient 3D-scalable video transmission for packet erasure channels," in *Proc. IEEE ICASSP*, Mar. 2005, pp. 305–308.
- [4] Y. Wang, A. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.
- [5] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 821–834, May 1993.
- [6] O. Campana and R. Contiero, "An H.264/AVC video coder based on multiple description scalar quantizer," in *Proc. IEEE ACSSC*, Oct.–Nov. 2006, pp. 1049–1053.
- [7] R. Bemardini, M. Durigon, R. Rinaldo, L. Celetto, and A. Vitali, "Polyphase spatial subsampling multiple description coding of video streams with H.264," in *Proc. IEEE ICIP*, Oct. 2004, pp. 3213–3216.
- [8] J. Jia and H. K. Kim, "Polyphase downsampling based multiple description coding applied to H.264 video coding," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E89-A, no. 6, pp. 1601–1606, Jun. 2006.
- [9] J. G. Apostolopoulos, "Error-resilient video compression through the use of multiple states," in *Proc. IEEE ICIP*, vol. 3, Sep. 2000, pp. 352–355.
- [10] S. Gao and H. Gharavi, "Multiple description video coding over multiple path routing networks," in *Proc. ICDT*, Aug. 2006, pp. 42–47.
- [11] C. W. Hsiao and W. J. Tsai, "Hybrid multiple description coding based on H.264," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 1, pp. 76–87, Jan. 2010.
- [12] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MTCF," in *Proc. IEEE ICME*, Jul. 2006, pp. 1929–1932.
- [13] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.
- [14] C. Zhu and M. Liu, "Multiple description video coding based on hierarchical B pictures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 4, pp. 511–521, Apr. 2009.
- [15] J. Reichel, H. Schwarz, and M. Wien, *Joint Scalable Video Model 11 (JSVM 11)*, document JVT-X202, Joint Video Team, Jul. 2007.
- [16] X. Ji, D. Zhao, and W. Gao, "Concealment of whole-picture loss in hierarchical B-picture scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 1, pp. 11–22, Jan. 2009.
- [17] *H.264/AVC Reference Software—JM* [Online]. Available: <http://iphome.hhi.de/suehring/tml>
- [18] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 13th VCEG Meeting, ITU-T SC16/Q6, Austin, TX, Apr. 2001.
- [19] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1256, Sep. 1960.
- [20] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and modeling of the temporal dependence in packet loss," in *Proc. IEEE INFOCOM*, Mar. 1999, pp. 345–352.



Wen-Jiin Tsai (M'07) received the Ph.D. degree in computer science from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1997.

She is currently an Assistant Professor with the Department of Computer Science, NCTU. Before joining NCTU in 2004, she was with Zinwell Corporation, Hsinchu, as a Senior Research and Development Manager for six years. Her current research interests include video coding, video streaming, error-concealment, and error resilience techniques.



Hao-Yu You received the B.S. degree in computer science from National Central University, Zhongli, Taiwan, in 2008, and the M.S. degree from National Chiao Tung University, Hsinchu, Taiwan, in 2010.

He is currently with the Home and Health Business Unit, Quanta Computer, Inc., Taoyuan, Taiwan, as a Researcher. His current research interests include video and audio device development.