Robust Video Coding Based on Hybrid Hierarchical B Pictures

Wen-Jiin Tsai, Yu-Chen Sun, and Po-Jui Chiu

***Abstract*—Compressed video streams transmitted over error-prone environments are usually corrupted by transmission errors. Error-concealment techniques can be used to recover the lost information. In this paper, a hybrid model is proposed to improve the error-concealment performance. The model combines two hierarchical B-picture coding structures such that key-frames, reference B frames, or even nonreference B frames have buddy frames to serve as their data recovery frames when they are lost. With buddy frames, the distance between a lost frame and its recovering frame can be substantially reduced. In addition, an improved estimation method is also proposed to further increase the accuracy of recovering motion. Error-concealment performance can thus be significantly improved with little bit-rate redundancy. We have conducted experiments to compare its performance with other methods, and the results show that the proposed hybrid model outperforms these competed methods. The advantages of the proposed hybrid model are demonstrated in error-free and packet-loss environments.**

***Index Terms*—Error concealment, error-resilient coding, hierarchical B pictures, hybrid model.**

I. Introduction

DURING the stage of transmission through the error-prone environment, packet loss might occur due to signal degradation, oversaturated bandwidth, or routing issues. Moreover, the data may arrive too late to be used in real-time applications. In the case of the transmission of compressed video sequences, this loss may result in a completely damaged stream at the decoder side. 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.

In recent years, several ER methods have been developed. Forward error correction (FEC) [1] is a general approach to allocate information redundancy to combat packet loss. Intra/inter coding mode selection [2] adaptively encodes intra-encoded block to eliminate error propagation. Error-resilient rate-distortion optimization [3] further proposed a coding strategy considering network condition. Flexible macroblock ordering [4], [5] spatially interlaces information to combat burst errors. Multiple description coding (MDC) [6], [7] further generalizes this concept that divides the information into several descriptors. Most of them were built on the conventional H.264/AVC [8] coding structure. Since a hierarchical B-picture structure demonstrates superior compression performance than the conventional one [9], this paper is concerned with the hierarchical B-picture structure. 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 [9]. Even though hierarchical-B picture coding has been widely used in scalable extension of H.264/AVC (SVC) [10] to provide temporal scalability, error-concealment algorithms adopted on it are still very simple. Zhu and Liu [11] proposed an MDC based robust video coding method based on the hierarchical B-picture structure. It generates two descriptors by duplicating the original sequence and then encoding them by using hierarchical B structure with staggered key frames in the two descriptors. By using different QPs at different levels, their approach enables each frame to have two different quality fidelities in different descriptors for error resilience.

In the case of packet loss, EC techniques can be used to recover the lost information. There are many existing EC algorithms, such as spatial interpolation [12], frequency domain interpolation [13], [14], and temporal compensation based on interframe correlation [15]. Among them, temporal error concealment is the most widely used approach, especially to combat the whole-frame loss problem, when hierarchical B-picture coding is used. The simplest temporal EC method is frame copy [16], in which each damaged macroblock is directly replaced by the colocated one in the temporally previous picture. Although it seems to be simple and fast, it suffers from large distortion in the case of fast motion in the erroneous block area. Thus, some methods based on motion compensation have been proposed, which replace the lost block with the one from the previous frame that is shifted to compensate for the estimated motion. To eliminate the complexity of motion estimation in these methods, an approach based on temporal direct mode [17] has been adopted in H.264/AVC (SVC) [10]. It derives the motion vector for each block in the lost B-picture according to the motion vector of the colocated block in the temporally subsequent reference

Manuscript received November 9, 2012; revised May 2, 2013 and July 5, 2013; accepted October 14, 2013. Date of publication November 19, 2013; date of current version May 2, 2014. This paper was recommended by Associate Editor J. Zhang.

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

Y.-C. Sun is with the Multimedia Technology Development Division, Corporate Technology Office, MediaTek, Inc., Hsinchu 30078, 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.2013.2291280

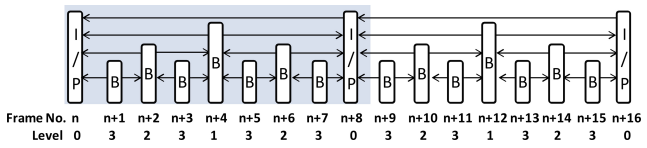

Fig. 1. Hierarchical B-picture prediction structure (n is a multiple of 8).

picture. This method has low computational complexity due to no motion estimation. However, its EC efficiency is usually unsatisfactory. Yan and Gharavi [18] proposed a method that estimates the missing motion of a block by using the motion of its neighboring blocks. For the hierarchical B-picture structure, Ji *et al.* [19] proposed a method based on enhanced temporal direct mode [20], in which the motion vectors for each block are allowed to be derived from the temporally not only subsequent reference picture, as in H.264/AVC, but also the previous reference picture. Thus, the approach in [19] derived motion of each block in the lost picture from the motion vector of the colocated block in the temporal subsequent or previous reference picture. In addition, they also proposed that the motion of the damaged block can be derived from the motion vectors of the colocated blocks in the temporally neighboring left and/or right B-pictures at the next higher temporal level. Their experimental results show that motion prediction in this way can improve the EC performance.

However, in the aforementioned temporal EC approaches, only the prediction of missing motion is discussed. The estimation of missing pixel values is seldom addressed. A widely used technique is to estimate the missing pixels by using pixels on the reference pictures of the lost picture. However, in the hierarchical B-picture coding structure, the distances between a lost picture and its reference pictures are often far apart in display order, especially when the loss occurs at lower (coarse) hierarchical levels. Estimating missing pixels in this way usually result in bad EC performance. Furthermore, when the reference pictures are far apart from the lost pictures, the accuracy of the prediction for missing motions is also degraded because large distance motion may not remain linear as it is usually assumed in motion interpolation or extrapolation.

In this paper, an error resilient coding based on hierarchical B pictures is proposed. In our approach, a new hierarchical coding structure that combines two conventional hierarchical coding structures is employed to reduce the distance between a lost picture and its recovering pictures. In addition, based on the new structure, an improved estimation method is proposed to further increase the accuracy of recovering motion. Experiments demonstrate that encoding hierarchical B pictures in this way can improve error-concealment performance significantly. It is worth mentioning that, for error concealment, the proposed method is the first one that utilizes buddy frames to reduce the recovery distance, and thus increase the confidence of the frames used for recovery. The proposed method could cooperate with most state-of-the-art studies.

II. Background

A typical hierarchical prediction framework with four dyadic hierarchy levels is illustrated in Fig. 1, where the key

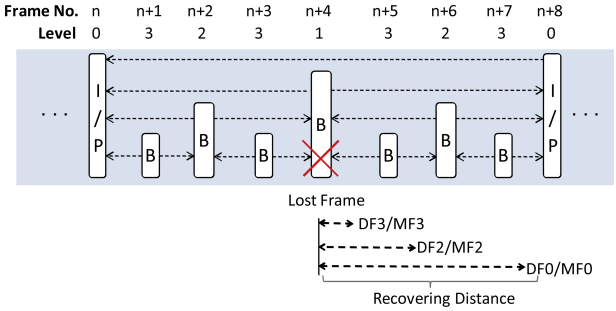

Fig. 2. Experimental setting for different combinations of motion frames (DF0, DF2, and DF3) and data frames (MF0, MF2, and MF3). (n is a multiple of 8.)

TABLE I
EXPERIMENTAL RESULT FOR ALL COMBINATIONS OF MOTION FRAMES AND DATA FRAMES

Concealment performance (db)			
Data Frame \ Motion Frame	MF0	MF2	MF3
DF0	20.4[13]	23.2[14]	21.9
DF2 (not valid)	23.7	26.6	27.2
DF3 (not valid)	26.2	29.5	29.8

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, as shown in Fig. 1. For 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) [21], QPs of the B frames at level-1 equal to the QPs of the I/P frames plus 4, and the QPs increase by 1 from one hierarchical level to the next level.

We refer to the I/P frames at the lowest hierarchical level as key frames; the B frames at intermediate levels as reference B frames (RB frames) because they are used as reference; and the B frames at the highest level as nonreference B frames (NRB frames) because they are not used as reference. Hierarchical B-frame structure has the characteristic that the frames at different levels have different reference distances (which means the temporal distance between a frame and its reference frame). Among the three types of frames, key frames have the longest reference distance, RB frames the medium, and NRB frames the shortest.

III. Motivation

Temporal error concealment in the hierarchical B-picture structure includes lost-motion and lost-pixel recovery. The lost information is estimated by referring other valid frames. The lost-motion would be estimated by interpolating, extrapolating, or compositing the motion vectors of the blocks in the motion prediction frame (MF). Then, the lost-pixel could be recovered from pixels of the data prediction frame (DF) according to estimated motions. DF can be different from MF. To have

TABLE II
MINIMAL PIXEL RECOVERING DISTANCES FOR LOST FRAMES AT DIFFERENT HIERARCHICAL LEVELS

Hierarchical level	Lost frame	Recovering frame	Recovering distance
Level 0	n+8	n+0	8
Level 1	n+4	n+0, n+8	4
Level 2	n+2	n+0, n+4	2
	n+6	n+4, n+8	2
Level 3	n+1	n+0, n+2	1
	n+3	n+2, n+4	1
	n+5	n+4, n+6	1
	n+7	n+6, n+8	1

better error concealment, it is important to select appropriate MFs and DFs. Because the correlation of the lost frame and the referred frame increases when the frame distance decreases, frames within smaller distance could provide more reliable recovering information and thus better concealment performance. To explore the relation between error concealment performance and recovery distance, experiments were conducted for *Foreman* sequence (CIF), where assume that a four-level hierarchical B-frame structure is adopted and frame-loss occurs on every level-1 frame. To recover these lost frames, temporal concealment is applied with various selections of DFs and MFs. As illustrated in Fig. 2, to recover frame $n+4$, both MFs and DFs were chosen from frames $n+5$, $n+6$, or $n+8$, denoted by MF3/DF3, MF2/DF2, and MF0/DF0, respectively, because they are on levels 3, 2, and 0, respectively. We conducted experiments for all the possible combinations of DFi and MFi, where $i=0$, 2, 3, and the results are shown in Table I. Note that some combinations in Table I may not be realistic because the DF frames are unavailable (not yet been decoded) when performing error concealment, e.g., DF2 and DF3 are not available when level-1 frames are under error concealment. We still simulate these cases for illustrating the rationale behind the proposed method. To simulate these cases, we first assume the lost level-1 frame is not actually lost, so it can be correctly decoded and used for decoding both DF2 and DF3. Then, the decoded DF2 and DF3 are used to recover the lost level-1 frame. In other words, this experiment was conducted to ignore the error propagation effect and to focus on evaluating how the concealment performance is affected by various recovering distances. Obviously, this is not a realistic system but only for explaining the importance of recovering distance. In Table I, the cell (MF0, DF0) means to use level-0 frames (i.e., frame $n+8$) as both DF and MF for recovery. Note that frame $n+8$ is the reference frame of the lost frame in this example; choosing MF and DF in this way is known as temporal direct mode (TDM) [17] of H.264/AVC.

Instead of using reference frames as both MFs and DFs, WTDM [19] chooses MFs from the frames on the next higher level of the lost frame to reduce motion recovering distance. Note that since motion vector decoding of a frame does not depend on other frames in H.264/AVC, it is possible to obtain motion vectors of an MF frame located at a higher level than the lost frame, even though all the pixels of this MF frame have not been decoded. In our example, the corresponding performance is the case shown in cell (MF2, DF0). Compared with the one in (MF0, DF0), the performance is improved because motion recovering distance becomes shorter. From this result we might expect that choosing MF3 as the motion prediction frame should produce the best error-concealment quality because MF3 has the shortest motion recovering distance. However, it is not as expected when DF0 is adopted as the data prediction frame, as can be seen in Table I where (MF3, DF0) performs worse than (MF2, DF0). The reason might be that even though MF3 is located close to the lost frame, it is far away from DF0. Therefore, the motion vectors in MF3 need to be greatly extrapolated to reach DF0, resulting in the decrease in motion accuracy and hence the degradation in error-concealment quality. The result implies that MF and DF should not be determined independently.

Although many methods have been proposed to select proper MFs to reduce motion recovering distance, how to reduce data recovering distance is seldom discussed. Most studies use pixels on the reference frames to recover missing pixels. Selecting DFs in this way may result in long data recovering distance. Take level-1 frame loss as an example, reference frames of the frame $n+4$ in Fig. 1 are frames n and $n+8$, both of them are four frames away from frame $n+4$ in display order; namely, the data recovering distance will be 4 if frame $n+4$ is lost. Table II shows data recovering distances for frame loss in different hierarchical levels, respectively, assuming that their reference frames are used for recovery. It can be seen that data recovering distances are large, especially for the cases of frame loss in lower hierarchical levels. However, long data recovering distance may result in severe quality degradation, as can be seen in Table I where the performance with DF0 is always the worst, while that with DF3 is always the best, if the same MFs are adopted. This implies that if data recovering distance can be reduced, it is very promising that error-concealment performance can be improved. However, with a hierarchical coding structure, it is hard to take advantage of those frames with recovering distances shorter than reference frames because these frames have not yet been decoded when the lost frame is under recovery. To solve this problem, we propose a variation of hierarchical B structure to reduce data recovering distance.

In summary, both motion and data recovering distances influence error-concealment performance significantly. In this paper, an approach based on the hierarchical B-picture structure is proposed, which is aimed at jointly determining MFs and DFs to reduce both motion and data recovering distances.

IV. PROPOSED METHOD

Here, a variation of hierarchical B structure is proposed for better error-concealment performance. As mentioned above, key frames have the longest reference distance, resulting in the worst error-concealment performance when they are lost. To improve the performance, a hybrid model called H_{N+1} is proposed, which combines an N-level with a one-level hierarchical B-picture structure. As an example in Fig. 3(a) where $N=4$, by combining a four-level and a one-level

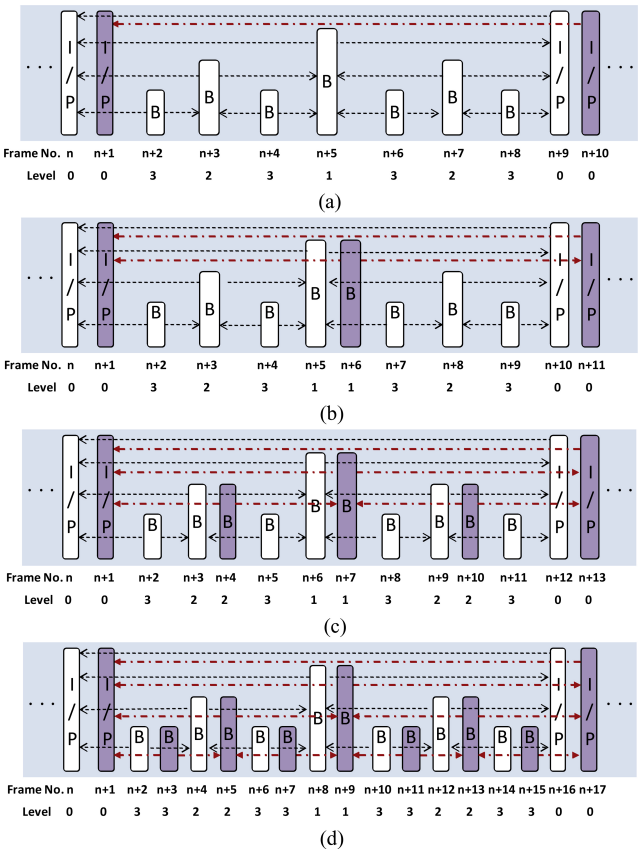

Fig. 3. Proposed hybrid model based on hierarchical B structure. (a) H_{4+1} model (n is a multiple of 9). (b) H_{4+2} model (n is a multiple of 10). (c) H_{4+3} model (n is a multiple of 12). (d) H_{4+4} model (n is a multiple of 16).

hierarchical B structures, each key frame in the resulting sequence has a neighboring frame located at the same level. Rather than encoding frame n + 10 as a level-2 RB frame in the conventional hierarchical B structure shown in Fig. 1, the proposed model will encode frame n + 10 as a key frame (I/P frame). We call a key frame and its neighboring key frame the buddy frames which are a pair of frames used to recover each other when there is a loss. In Fig. 3(a), frame n + 9 and frame n + 10 are buddy frames. If frame n + 9 is lost, instead of using its reference frame (frame n) for missing pixel recovery, its buddy frame n + 10 is used. Compared with WTDM [19] described in the previous section, the proposed H_{4+1} reduces the recovering distance of key frames from eight to one frame. By employing buddy frames in this way, error-concealment performance of key-frame loss can be improved significantly.

In addition to key frames, RB frames also suffer from the problem of long data recovering distance. The proposed buddy frames can also be applied to RB frames. The hybrid model H_{N+2} that is a variation of H_{N+1} is proposed for this. It combines an N-level hierarchical B structure with a 2-level hierarchical B structure as an example in Fig. 3(b), where N = 4. By combining a four-level and a two-level hierarchical B structures, not only each key frame but also each level-1 RB-frame such as frame n + 5 in the resulting sequence has buddy frames located at the same level. In Fig. 3(b), if RB-frame n + 5 is lost, instead of using its reference frames for missing pixel recovery, its buddy frame [frame n + 6 in Fig. 3(b)] will be used. Compared with WTDM [19] where reference frames are used for recovery, H_{4+2} reduces the data recovering distance of RB-frame n + 5 from four frames (the distance between frame n + 4 and its reference frames shown in Fig. 1) to one frame only [the distance between frame n + 5 and its buddy frame in Fig. 3(b)].

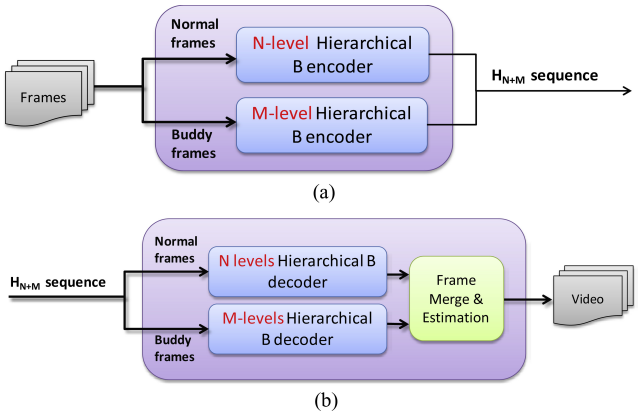

Fig. 4. Architecture of the proposed hybrid model H_{N+M}. (a) Encoder architecture. (b) Decoder architecture.

Similarly, the proposed buddy frames can also be applied to level-2 RB-frames and level-3 NRB-frames to reduce their data recovering distances. Two variations of hybrid model, H_{N+3} and H_{N+4}, are shown in Fig. 3(c) and (d), respectively. The H_{N+3} model in Fig. 3(c) combines a four-level and a three-level hierarchical B structures, while the H_{N+4} model in Fig. 3(d) combines two four-level hierarchical B structures. As observed in these figures, H_{4+3} model reduces the recovering distance of level-2 RB-frame [e.g., frame n + 3 in Fig. 3(c)] from two to one frame and H_{4+4} model keeps the recovering distance of level-3 NRB frame [e.g., frame n + 2 in Fig. 3(d)] as one frame.

The proposed various hybrid models can be generalized as a H_{N+M} model which means that the resulting sequence is the combination of an N-level hierarchical B-picture structure and an M-level one. The encoder architecture of the H_{N+M} model is depicted in Fig. 4(a). As the figure shows, the frames in the sequence are split into two groups: *G0* and *G1* first, and then each group will go through a standard hierarchical B picture encoder to perform motion estimation, transform, quantization, and entropy coding. The G0 frames are encoded as an N-level hierarchical structure and the G1 frames as an M-level structure, resulting in a H_{N+M} sequence.

Different hybrid models are made up by different G0 and G1 frames. For example, in H_{4+1} model, the G1 frames consist of frames 1, 10, 19, 28, . . . , etc., while in the H_{4+2} model, they are frames 1, 6, 11, 16, . . . , etc. For N = 4, the G1 frames of the four variations are summarized as follows, where *m* is an integer

$$H_{4+1}: \textit{frames}\ 1, 10, 19, 28, \ldots, 9m+1$$
$$H_{4+2}: \textit{frames}\ 1, 6, 11, 16, \ldots, 5m+1$$
$$H_{4+3}: \textit{frames}\ 1, 4, 7, 10, \ldots, 3m+1$$
$$H_{4+4}: \textit{frames}\ 1, 3, 5, 7, \ldots, 2m+1.$$

The decoder architecture of the proposed hybrid model H_{N+M} is depicted in Fig. 4(b), where the received frames

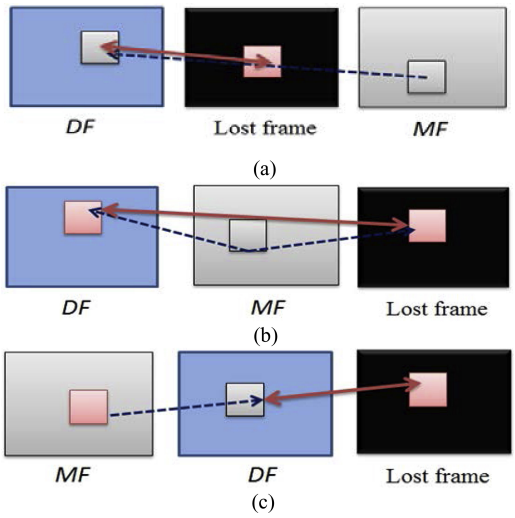

Fig. 5. Motion interpolation, composition, and extrapolation. (a) Motion vector interpolation. (b) Motion vector composition. (c) Motion vector extrapolation.

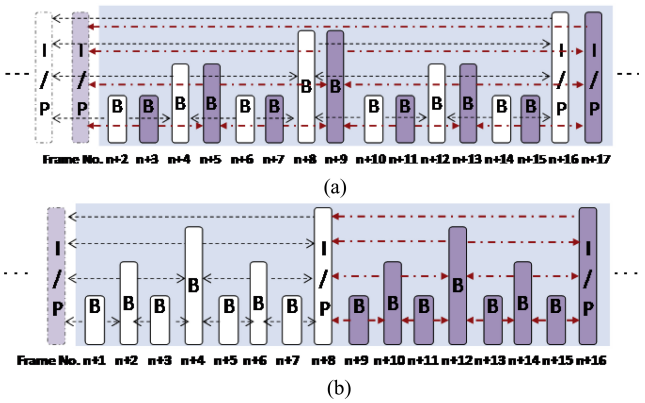

Fig. 6. Coding structures of hybrid model H_{4+4} and the original model. (a) Hybrid model H_{4+4} (n is a multiple of 16). (b) Original model (n is a multiple of 8).

are first split into two groups, G0 and G1. Then, each group will go through a standard hierarchical B decoder for entropy decoded, de-quantized, and inversely transformed. G0 frames are decoded with an N-level hierarchical structure; while G1 frames are decoded with an M-level one. Finally, the frame-merge and estimation procedure is used to reconstruct the order of frames for generating output sequence. If the decoder does not receive the two structures intact, the estimation procedure will be used to estimate the lost data. The estimation method is detailed in the next section.

V. Estimation of Lost Pictures

In the proposed method, we assume that each frame is divided into slices in a raster scan order. In the case of packet loss, it will result in successive macroblock loss, regardless of frame types and levels. Each lost block is recovered based on temporal correlation since the neighboring blocks are also lost. Namely, both data and motion prediction frames (i.e., DF and MF) must be determined for error concealment.

To serve as DFs requires that these pictures are decoded earlier than the lost picture. Therefore, for the hierarchical B structure, almost all the error-concealment methods choose reference frames of the lost frame to serve as the DFs. The DF can be in the backward direction, forward direction, or both. Since 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 pictures near in the display order to serve as its DFs. Therefore, in the proposed hybrid model, we choose the buddy frame of the lost frame to serve as DF because it is usually located near in temporal distance. However, not every frame has buddy frame. For example, in H_{4+1} model, only level-0 frames have buddy frames; while in H_{4+2} model, both level-0 and level-1 frames have buddy frames. If the lost frame has no buddy frame or the buddy frame is also lost, we simply use its reference frames to serve as DFs. That is, for the lost frame F_t^l with hierarchical level l at time instant t, we select its DF as

$$\mathrm{DF}(F_t^l) = \begin{cases} F^l_{t_{\mathrm{buddy}}}, \text{ if has buddy frame} \\ F^k_{t_{\mathrm{ref}}}, \text{ otherwise} \end{cases}$$

where k can be l, $l-1$, or $l-2$, depending on what level the reference frame of the lost frame is. As an example, for the H_{4+2} model in Fig. 3(b), if frame n + 8 is lost, its DFs are frame n + 5 and frame n + 10 because it has no buddy frame. But if frame n + 5 is lost, the DF will be its buddy frame n + 6.

As for MFs, since we can obtain motion information of a frame even though it has not been decoded, the MFs can be the frames later than the lost frame (in decoding order). As discussed in Section III, how to choose MF depends on not only motion recovering distance but also pixel recovering distance. Therefore, instead of using reference pictures at lower levels, if the lost frame has a buddy frame, we choose the nearest pictures at higher levels to serve as MFs because these pictures are temporally closer to the lost picture in display order. Otherwise, we choose the pictures at next higher level to serve as MFs [19] to prevent motion interpolation/extrapolation. As an example, if frame n + 5 in Fig. 3(b) is lost, we will select frames n + 4 and n + 7 (rather than its reference frames n and n + 10) as its MFs. But if frame n + 5 in Fig. 3(a) is lost, since it has no buddy frame, we will select frames n + 3 and n + 7 as its MFs. This selection policy is applied to all frames except NRB frames that are at the highest level within the hierarchical structure. For NRB frames, the MFs are selected from the reference frame at the next lower level or the buddy frame at the same level. As an example, if NRB frame n + 5 in Fig. 3(c) is lost, its reference frames n + 3 and n + 6 at lower levels are chosen as its MFs because it has no buddy frame. But if NRB frame n + 6 in Fig. 3(d) is lost, its buddy frame (frame n + 7) at the same level will be chosen.

Once both DFs and MFs of the lost picture have been determined, for every block in MF, its motion vector(s) are composed, extrapolated, or interpolated so that the motion vectors pointing to the lost frame from DFs can be obtained. Such motion vectors are called recovery motion vectors (RMV). If DF and MF are on different sides of the lost frame along

TABLE III
PERFORMANCE COMPARISON BETWEEN HYBRID MODEL H_{4+4} AND THE ORIGINAL MODEL. BOTH MODELS ENCODE *Mobile* SEQUENCE (CIF) AT 1900 KB/S

Concealment performance (db)		
Model / Loss rate	H_{4+4} hybrid model	Original model
0%	34.7	36.9
5%	32.1	31.9
10%	29.7	28.1
15%	27.7	25.2
20%	25.7	23.0

temporal dimension, the MV pointing to DF from MF are interpolated to obtain the RMV as illustrated in Fig. 5(a), where the RMV is denoted using a solid arrow. If DF and MF are on the same side of the lost frame, the MV pointing to DF from MF are either extrapolated or composed to get RMV, as illustrated in Fig. 5(b) and (c). In the case of motion vector composition, we would like to derive the motion vector pointing to DF from the lost frame by using two reference vectors, MF to DF and MF to the lost frame. Therefore, we composite these two reference vectors to get the RMV. Once all the RMVs have been derived, if a location on the lost picture is pointed by more than one RMV, its pixel value is replaced by the average of these pointing pixels on the DFs. If a location on the lost picture is not pointed by any RMV, the motion vector of the colocated pixel on the DF will be scaled and serve as RMV [24].

VI. EXPERIMENTAL RESULTS

A. Effects of Hybrid Structures

To see the effects of the proposed hybrid model, the experiment was first conducted for comparing the proposed H_{4+4} with a standard hierarchical B-frame structure with four levels. It is interesting to observe that both structures contain two key frames, two level-1 frames, four level-2 frames, and eight level-3 frames for every successive 16 frames, as seen in Fig. 6(a) and (b). With the same number of frames for each frame type, standard structure encodes the 16 frames as two successive GOPs, while H_{4+4} encodes them as two independent GOPs with interleaved positions in display order. Table III shows the resulting performance of the two structures. It can be seen that H_{4+4} performs worse than the standard structure for the error-free case as expected because temporal prediction distance in H_{4+4} is much farther than that in the standard one. However, in the case of packet loss, H_{4+4} shows superior performance. The result shows that, with the proposed hybrid structure, H_{4+4} did improve error resilience significantly. Fig. 7 shows the visual comparison of concealment results between a conventional model and the H_{4+4} hybrid model.

B. Effects of Hybrid Structure Variations

There are four variations of hybrid models: H_{4+1}, H_{4+2}, H_{4+3}, H_{4+4}. This section examines how they affect error resilient capability. Since how often a key frame is encoded as an

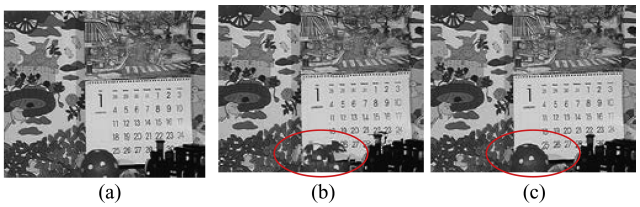

Fig. 7. Visual performance comparison between hybrid model H_{4+4} and the conventional model. Both models encode *Mobile* sequence (CIF) at 1900 kb/s. The major differences are highlighted by red circles. (a) Original frame. (b) Conventional model. (c) H_{4+4} hybrid model.

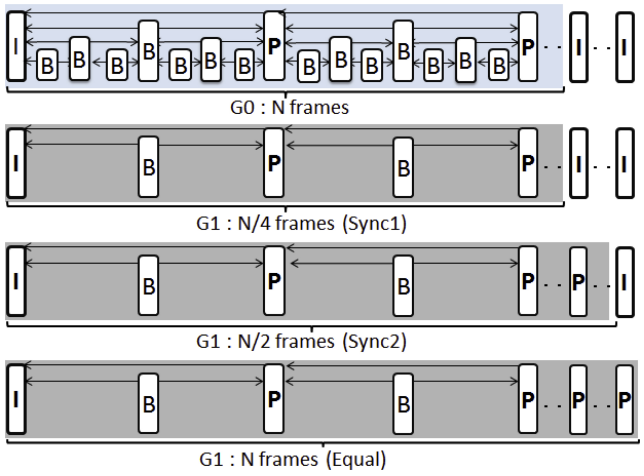

Fig. 8. Difference between Sync1, Sync2, and Equal (using H_{4+2}).

I-frame instead of a P-frame also affects the performance of the overall sequence, we adopt the same I-frame period, 32, for G0 frames in the four hybrid models. As for G1 frames, three different I-frame period settings are used for comparison. The first setting, called *Equal*, is to use the same I-frame period (i.e., 32) for G1 frames in the four models. Using H_{4+2} as an example, adopting Equal setting means that GOP size of G1 is also 32. Since G0 is a four-level hierarchy and G1 is two-level hierarchy in H_{4+2}, the same GOP size of 32 means that the distance between successive I-frames in G1 is much larger than that in G0. The second setting, *Sync1*, is to synchronize the positions of I-frames in G0 frames and G1 frames. With Sync1, the I-frame periods of G1 frames are 4 in H_{4+1}, 8 in H_{4+2}, 16 in H_{4+3}, and 32 in H_{4+4}. Using H_{4+2} as an example where the GOP size of G0 is 32, its G0 will have four frames at key frame level, four B-frames at level 1, eight B-frames at level 2, and 16 B-frames at level 3 for one GOP. Using Sync1 for this H_{4+2} model means that its GOP size of G1 will be 8 because it consists of four key frames and four B-frames at level 1, just like G0. The third setting, *Sync2*, simply doubles the Sync1 I-frame periods for G1 frames, and keeps I-frame period as 32 for G0 frames. Using H_{4+2} as an example, one GOP of G1 will consist of 16 frames, including eight key-level frames and eight B-frames at level 1. It is obvious that, among three settings, Equal setting has the best coding efficiency and Sync1 the worst. However, Sync1 has the shortest (the best) error propagation length and Equal setting has the longest (the worst) one. Fig. 8 illustrates the difference of Sync1, Sync2, and Equal settings.

TABLE IV
PACKET-LOSS PERFORMANCE COMPARISON

Sequence	Loss rate	Sync1				Sync2				Equal			
		H_{4+1}	H_{4+2}	H_{4+3}	H_{4+4}	H_{4+1}	H_{4+2}	H_{4+3}	H_{4+4}	H_{4+1}	H_{4+2}	H_{4+3}	H_{4+4}
Foreman @800kbps	5%	35.31	35.61	**35.77**	35.26	35.10	35.36	35.23	34.95	34.83	35.14	35.23	35.26
	20%	28.83	**29.58**	28.92	28.46	28.17	28.92	28.24	27.61	27.54	28.35	28.24	28.46
Mobile @1900kbps	5%	32.29	32.40	**32.47**	32.10	28.17	28.92	28.24	27.61	31.47	31.76	31.70	32.10
	20%	25.86	**26.18**	26.17	25.66	25.09	25.51	25.20	24.68	24.02	24.90	25.20	25.66
News @400kbps	5%	38.72	38.60	**38.94**	38.83	38.56	38.35	38.67	38.68	38.32	38.23	38.67	38.83
	20%	34.08	33.83	34.21	**34.23**	33.54	33.21	33.63	33.58	32.82	32.99	33.63	34.23
Soccer @1300kbps	5%	33.79	34.03	**34.45**	33.43	33.58	33.79	33.88	33.05	33.16	33.60	33.88	33.43
	20%	25.76	26.16	**26.52**	25.41	25.30	25.70	25.53	24.44	24.60	25.34	25.53	25.41

Table IV shows the PSNR as a function of packet-loss rate (PLR) for 12 combinations of the four hybrid models with three I-frame period settings under four CIF sequences: *Foreman*, *Mobile*, *News*, and *Soccer*. All combinations encode the same sequence using the same bit-rate for fair comparison and the results presented are the averages of 100 independent runs. It is observed that, among the three I-frame period settings, Sync1 has the best performance. Among the 12 combinations, H_{4+3} with Sync1 achieves the overall best performance for all the sequences. Why H_{4+3} performed the best among the four variations of hybrid models can be explained as follows. For the four hybrid models, buddy frames are allocated at different levels to decrease recovering distances. As can be seen in Table II, allocating buddy frames at levels 0, 1, 2 can reduce recovering distances, respectively, from 8, 4, and 2 to 1. However, using buddy frames at level 4 cannot reduce the distance because the distance in the conventional structure is already 1. Thus, H_{4+3} can fully take advantage of buddy frames at levels 0, 1, 2 to reduce recovering distances. H_{4+4} cannot perform better than H_{4+3} because the buddy frames at level 3 cannot reduce recovering distance to improve error concealment. Besides, allocating more frames at high levels in H_{4+4} means that more frames will be encoded with large QP because the quantization setting that we adopted is based on the suggestion in [21]. As a result, H_{4+3} have better overall performance than H_{4+4}. As for various GOP settings of buddy frames (Sync1, Sync2, and Equal), when GOP size increases, the coding efficiency of buddy frames would increase because more frames are intercoded. However, large GOP size would increase error propagation lengths, and thus degrade the performance. The experimental result shows that Sync1 which has the smallest GOP size obtains the best performance, indicating that, compared with coding efficiency, error robustness has more impacts on overall performance.

C. Packet-Loss Performance

Since H_{4+3} with Sync1 outperforms all the other hybrid models, it was adopted for the comparison with other methods in packet-loss scenarios. We adopt four CIF and four 720p sequences to conduct the experiments. The Bernoulli channel is adopted, which assumes that each packet is lost randomly and independently. Each frame was encoded into three slices in raster scan order. Namely, a frame might be partially lost in the experiments. In the case of partial lost, the successfully received part could be normally decoded and the lost part would be recovered by the proposed method. Each video is encoded with multiple QPs to produce R-D (bitrate-PSNR) curves. To evaluate the performance at a specific target bitrate, we simply adopt a linear model that chooses two RD points most close to the target rate and then use them to linearly interpolate the PSNR value under the specified target rate. We compare H_{4+3} with Ji *et al.*'s method [19] and Zhu *et al.*'s method [11]. Ji *et al.*'s method called WTDM is a method based upon TDM of H.264/AVC for error concealment in a hierarchical B-picture prediction structure. The I-frame period is 32. Since the proposed method divides a video sequence into two independent coding units (G0 frames and G1 frames), it can be considered one of unbalanced MDC video approaches. Zhu *et al.*'s method is an MDC approach based on the hierarchical B-picture prediction structure. It duplicates each test sequence into two and then encodes by hierarchical B structure with staggered key frames in the two sequences. For example, if one sequence is encoded with the structure shown in Fig. 1 where frames n, n+8, n+16, ... are key frames, then the other one will have frames n+1, n+9, n+17, ... encoded as key frames. This approach is characterized by that each frame at levels 0, 1, or 2 of one sequence will be at level 3 of the other sequence and vice versa, resulting in two fidelities of each frame. Two variations, *defaultQP* and *modifiedQP*, in their literature are adopted in our comparison. The defaultQP follows the QP assignment rules specified in JSVM11 [21], while modifiedQP modifies the QPs of top-level frames to 51 to reduce bit-rates redundancy. The results in [11] show that RD performance of center decoder can be improved remarkably by modifiedQP, compared to defaultQP. All these methods are implemented based on H.264 reference software, JM 16.0 [22].

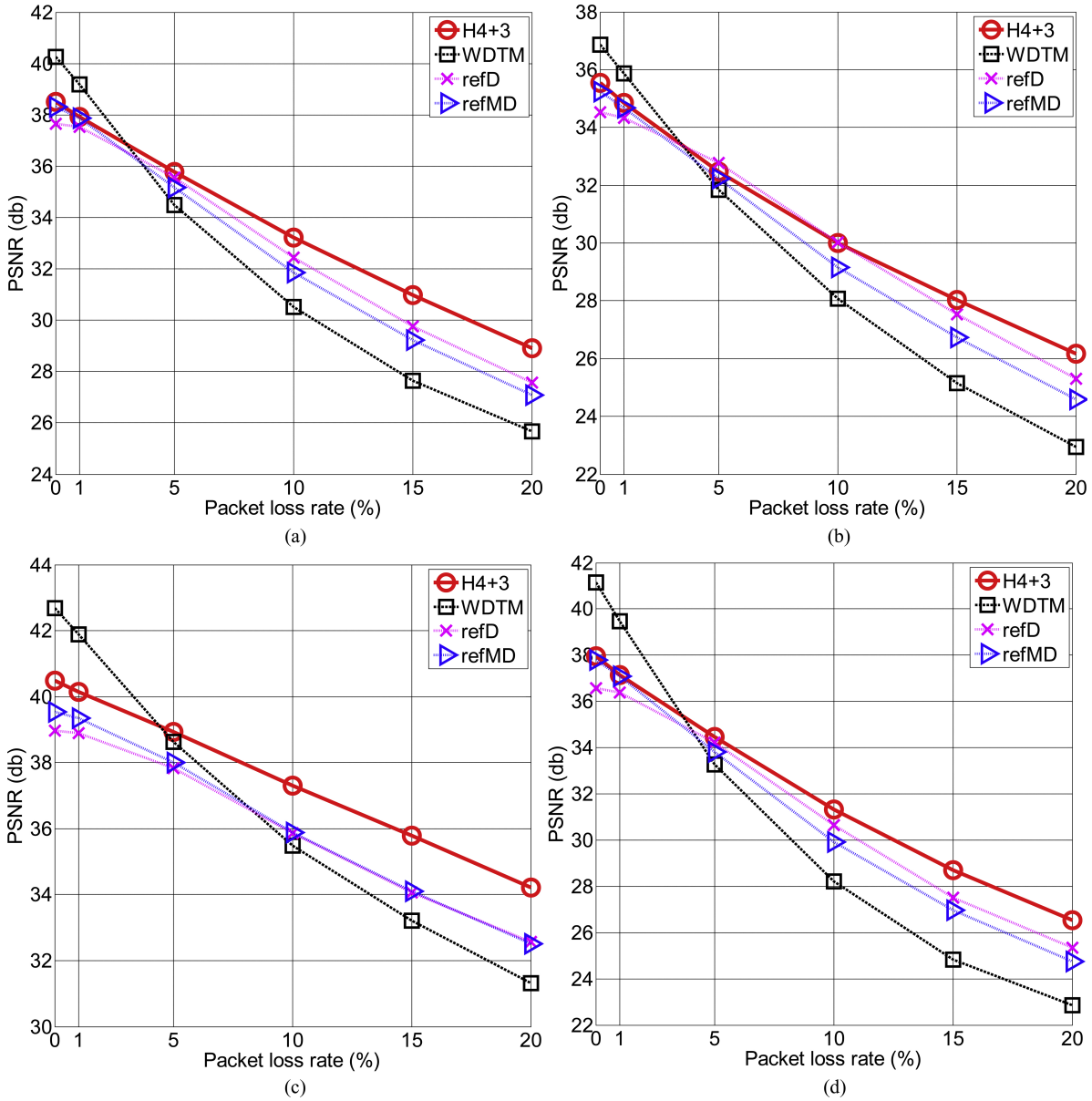

Fig. 9. Packet-loss performance of various methods using CIF sequences. (a) *Foreman* (CIF at 800 kb/s). (b) *Mobile* (CIF at 1900 kb/s). (c) *News* (CIF at 400 kb/s). (d) *Soccer* (CIF at 1300 kb/s).

Fig. 9 shows the result for four different methods with four CIF test sequences. In Fig. 9, the four methods encode the same sequence using the same bit-rate for fair comparison and the results are the averages of 100 independent runs. It can be seen that, as PLR increases, WTDM curves drop much more quickly than others, showing its poor error resilience. By duplicating the entire sequence, defaultQP (refD) and modifiedQP (refMD) achieve better error robustness than WTDM. Compared with defaultQP, the modifiedQP method shows better performance at low PLR because of its reduced bit-rate at top-level frames (NRB frames). However, such a reduction in bit-rate strongly affects its error-concealment effectiveness, and hence, degrades its performance dramatically at high PLR. Among all methods, the proposed H_{4+3} performed the best because it modifies hierarchical B coding structure by encoding more key frames and RB frames as buddy frames, resulting in reduced recovery distance and better error-concealment effect, especially at high PLRs. To summarize, the overall results demonstrate that, by combining two hierarchical B-picture structures, the proposed hybrid model offers a better tradeoff between bit-rate redundancy and error-resilient capability, and thus, achieves the best performance among the four methods.

Packet-loss performance of the four methods using 720p sequences is shown in Fig. 10. It is observed that the proposed method shows significant performance improvement against others for these high-definition sequences, and the performance gaps are even larger when compared with CIF sequences in Fig. 9. The result shows the proposed method can be a potential approach for next-generation video delivering applications.

Since the proposed method improves error resilience by utilizing bitrate redundancy, it may perform worse than

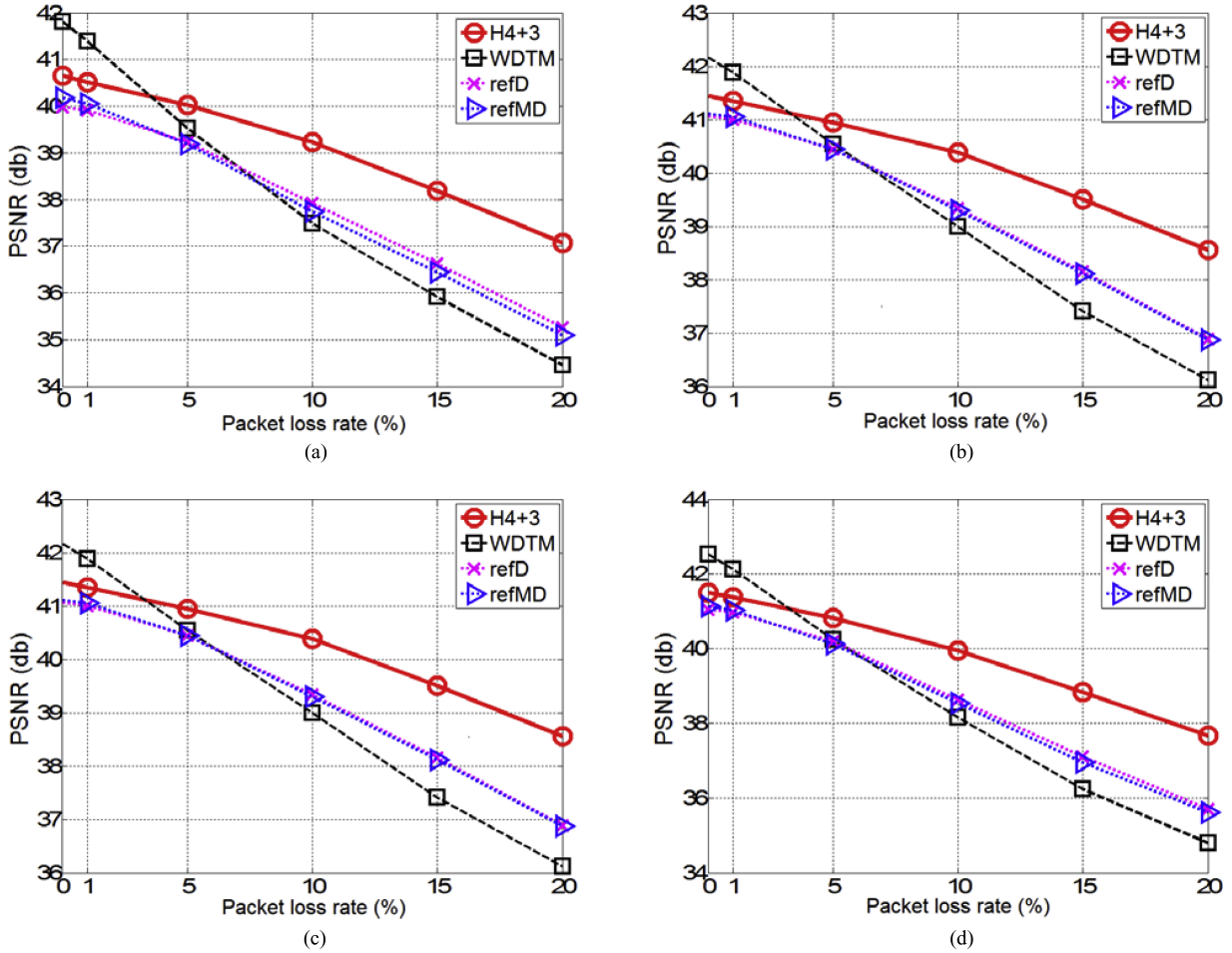

Fig. 10. Packet-loss performance of various methods using 720p sequences. (a) *FourPeople* (720p at 3600 kb/s). (b) *Johnny* (720p at 2400 kb/s). (c) *KristenAndSara* (720p at 3000 kb/s). (d) *Vidyo1* (720p at 2600 kb/s).

TABLE V
STATISTICAL INFORMATION FOR 100 INDEPENDENT RUNS

PSNR Statistical Information(dB)											
Sequence	Loss Rate	Avg.	Max.	Min.	Std.	Sequence	Loss Rate	Avg.	Max.	Min.	Std.
Foreman (CIF)	1%	37.55	36.30	38.10	0.44	FourPeople (720p)	1%	40.69	39.84	40.84	0.18
	10%	33.01	30.22	35.56	1.10		10%	39.38	37.53	40.35	0.59
	20%	28.80	24.61	32.32	1.74		20%	37.17	32.87	39.55	1.45
Mobile (CIF)	1%	34.47	32.99	35.13	0.48	Johnny (720p)	1%	41.24	40.97	41.33	0.09
	10%	29.80	26.90	32.23	0.90		10%	40.30	38.60	40.87	0.42
	20%	26.08	22.62	29.20	1.35		20%	38.50	35.59	40.43	1.11
News (CIF)	1%	40.03	39.20	40.36	0.28	KristenAnd Sara (720p)	1%	41.29	40.59	41.42	0.14
	10%	37.24	33.90	38.83	0.87		10%	39.88	37.87	40.70	0.58
	20%	34.17	28.99	37.07	1.59		20%	37.61	34.09	39.93	1.45
Soccer (CIF)	1%	36.94	35.42	37.71	0.51	Vidyo1 (720p)	1%	41.66	41.04	41.77	0.12
	10%	31.24	28.47	33.61	1.10		10%	40.55	39.12	41.33	0.49
	20%	26.53	21.24	29.62	1.68		20%	38.40	33.65	40.70	1.39

conventional structures at error-free or low pack-loss rates when the error robustness improved by the proposed method cannot compensate for the bitrate overhead. That is, there is a tradeoff between error robustness and coding efficiency. Fortunately, we found that for all the test sequences we have adopted, the proposed method performed worse than H.264/AVC when the PLR is less than 3∼5%. This implies that we can adopt a mechanism that chooses to use the proposed hybrid model to enhance error robustness only when PLR > 5% and use a conventional hierarchical B-picture structure to have better coding efficiency if PLR <= 5%. We have conducted such a mechanism and the R-D performance of the combined method is shown in Fig. 11. As the result shows, with a simple adaptive mechanism on the top of the proposed hybrid model, it can provide more robust video quality for all network conditions.

To evaluate the stability of the performance, Table V lists the statistical information of 100 independent runs for each loss case. It is observed that the standard deviation of PSNR values over 100 runs is below 1.1 dB when the PLR is smaller than 10% and it is about 1.5 dB when the loss rate increases to 20%. The result shows that the performance of the proposed method is robust under varying network conditions and stable for 100 independent runs.

D. Error-Free Performance

This section examines the error-free performance of all the methods and the results are presented in Fig. 12. It is observed that the eight curves in Fig. 12 can be divided into three groups: WTDM and JM have the best rate-distortion performance, defaultQP and modifiedQP have

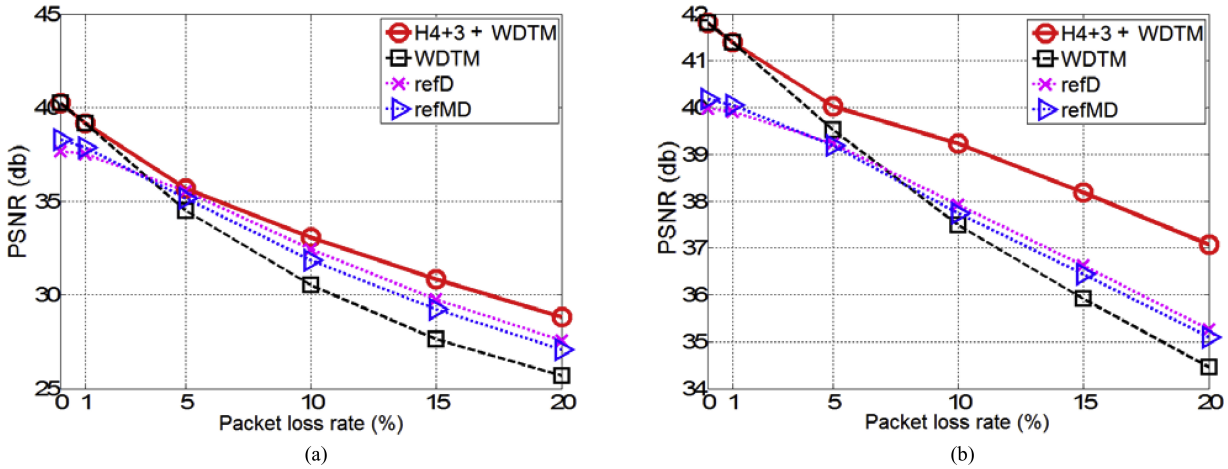

Fig. 11. Packet-loss performance of the combination of the proposed hybrid model and the conventional hierarchical B-picture structure. (a) *Foreman* (CIF at 800 kb/s). (b) *FourPeople* (720p at 3600 kb/s).

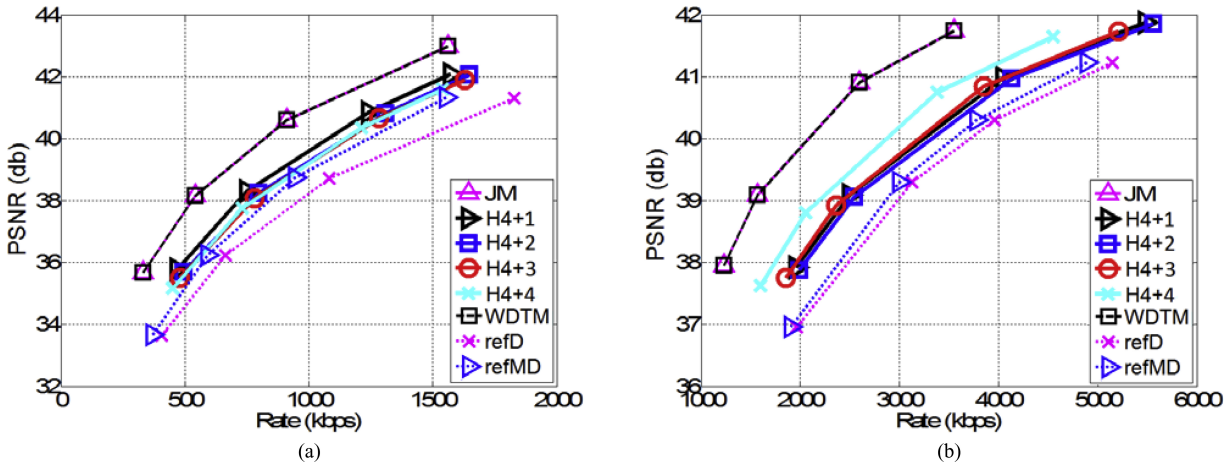

Fig. 12. Rate-distortion performance comparison in error-free environment. (a) *Foreman* (CIF). (b) *FourPeople* (720p).

TABLE VI
BIT-RATE REDUNDANCY COMPARISON

Bit-rate redundancy (%)							
Sequence	WTDM	RefD	RefMD	H4+1	H4+2	H4+3	H4+4
Foreman(CIF)	0	76.9	53.9	32.3	41.9	43.7	44.0
Mobile(CIF)	0	60.2	26.6	17.5	19.8	20.1	20.1
News(CIF)	0	91.1	73.4	46.1	50.4	46.7	44.8
Soccer(CIF)	0	90.6	54.1	29.2	42.3	46.4	52.9
FourPeople(720p)	0	87.8	79.8	56.2	60.0	54	38.5
Johnny(720p)	0	89.9	87.8	60.5	63.6	54.7	38.4
KristenAndSara(720p)	0	90.3	84.7	62.9	69.0	57.3	49.2
Vidyo1(720p)	0	88.3	82.2	61.0	63.5	57.4	43.1

the worst performance, and the four hybrid models have the performance between them. Table VI shows the bitrate redundancy produced by each method. It is defined as the Bjontegaard delta bitrate between JM and each method, which is calculated by the method in [23]. In Fig. 12, WTDM performs the same to JM16.0 because it focused its error-concealment approach on missing motion recovery and did not modify the hierarchical B picture coding structure. Thus, it did not produce any bit-rate redundancy that may reduce the R-D performance in Fig. 12. Both defaultQP and modifiedQP have large bit-rate redundancy that degrades their performance in Fig. 12. As shown in Table VI, defaultQP produces redundancy about 60%∼90%. Compared with defaultQP, while modifiedQP reduces the redundancy about 10% by modifying the QPs of NRB frames, the RD performance improvement as shown in Fig. 12 is quite limited. Compared with modifiedQP, the proposed hybrid models have much lower redundancy as shown in Table VI and much better RD performance than defaultQP and modifiedQP as shown in Fig. 12.

VII. CONCLUSION

A hybrid model based on hierarchical B pictures is proposed, which improves error-concealment effects by combining two hierarchical B-picture coding structures. For a four-level hierarchical structure, there are four variations of the proposed hybrid model. They are H_{4+1}, H_{4+2}, H_{4+3}, and H_{4+4}. In the H_{4+1} model, each base-level key frame has a buddy frame that is used to serve as the data recovery frame when it is lost. In H_{4+2} and H_{4+3}, not only key-frames, but also

RB-frames have buddy frames. In H_{4+4}, all the frames, including NRB-frames, have buddy frames. With buddy frames, data recovery distance can be reduced and the error-concealment performance can be substantially improved. Experiments have been conducted for eight methods: four variations of the proposed model (H_{4+1}, H_{4+2}, H_{4+3}, and H_{4+4}), WTDM [19], two methods (defaultQP and modifiedQP) in [11], and JM16.0. The experimental results show that the proposed H_{4+3} has the overall best performance among them.

In the proposed method, the pixels that are not pointed by any RMV might belong to static/uncovered background pixels. It is possible to utilize some advanced methods, e.g., the McFIS method proposed in [25], to recover these pixels. However, the result of our current experiments is not good. The reason might be due to that the accurate motion from loss frame to McFIS is hard to estimate. This would be one of potential future works.

Besides, current rate control algorithms cannot be directly applied to the proposed hybrid model. A method that determines the best bit-allocation between G0 and G1 frames is required. Once this method has been designed, the target bitrates for G0 and G1 frames can be determined, respectively. Then, G0 and G1 frames can be encoded independently for their respective target bitrates, using currently available rate control algorithms. Designing an optimal bit-allocation algorithm is challenge and could be one of potential future works.

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] S. Wan and E. Izquierdo, "Rate-distortion optimized motion-compensated prediction for packet loss resilient video coding," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1327–1338, May 2007.

[4] P. Lambert, W. De Neve, Y. Dhondt, R. Van de Walle, "Flexible macroblock ordering in H.264/AVC," *J. Visual Commun. Image Representation*, vol. 17, no. 2, pp. 358–375, Apr. 2006.

[5] J.-Y. Shih and W.-J. Tsai, "A new unequal error protection scheme based on FMO," *Multimedia Tools Appl.*, vol. 47, no. 3, pp. 461–476, Aug. 2009.

[6] 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.

[7] W. J. Tsai and J. Y. Chen "Joint temporal and spatial error concealment for multiple description video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1822–1833, Dec. 2010.

[8] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[9] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MTCF," in *Proc. IEEE ICME'06*, pp. 1929–1932.

[10] 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.

[11] 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.

[12] W. Zhu, Y. Wang, and Q.-F. Zhu, "Second-order derivative-based smoothness measure for error concealment in DCT-based codecs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 10, pp. 713–718, Oct. 1998.

[13] S. Cen and P. C. Cosman, "Decision trees for error concealment in video decoding," *IEEE Trans. Multimedia*, vol. 5, no. 1, pp. 1–7, Mar. 2003.

[14] M. Ancis, D. D. Giusto, and C. Perra, "Error concealment in the transformed domain for DCT-coded picture transmission over noisy channels," *Eur. Trans. Telecomm.*, vol. 12, no. 3, pp. 197–204, 2001.

[15] S. C. Hsia, S. C. Cheng, and S. W. Chou, "Efficient adaptive error concealment technique for video decoding system," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 860–868, Oct. 2005.

[16] J. W. Suh and Y. S. Hu, "Error concealment based on directional interpolation," *IEEE Trans. Consumer Electron.*, vol. 43, no. 3, pp. 295–302, Aug. 1997.

[17] M. Flierl and B. Girod, "Generalized B pictures and the draft H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 587–597, Jul. 2003.

[18] B. Yan and H. Gharavi, "A hybrid frame concealment algorithm for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 98–107, Jan. 2010.

[19] X. Ji, D. Zhao, and W. Gao, "Concealment of whole-picture loss in hierarchical B-picture scalable video coding," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 11–22, Jan. 2009.

[20] J. Zheng, X. Ji, G. Ni, W. Gao, and F. Wu, "Extended direct mode for hierarchical B picture coding," in *Proc. IEEE ICIP*, vol. 2. Sep. 2005, pp. 265–268.

[21] J. Reichel, H. Schwarz, and M. Wien, *Joint Scalable Video Model 11 (JSVM 11)*, document JVT-X202, Joint Video Team, Jul. 2007.

[22] *H.264/AVC Ref. Software—JM* [Online]. Available: http://iphome.hhi.de/suehring/tml/

[23] G. Bjontegaard, *Improvement of the BD-PSNR Model*, VCEG document VCEG-AI11, ITU-T SG16/Q6, 35th VCEG Meeting, 2008.

[24] S. K. Bandyopadhyay, Z. Wu, P. Pandit, and J. M. Boyce, *Frame Loss Error Concealment for H.264/AVC*, JVT-P072, 73rd MPEG Meeting and 16th JVT Meeting, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Jul. 2005.

[25] M. Paul, W. Lin, C. T. Lau, and B.-S. Lee, "McFIS in hierarchical bipredictve pictures-based video coding for referencing the stable area in a scene" in *Proc. IEEE ICIP*, Sep. 2011, pp. 3521–3224.

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

She is currently an Associate 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 research interests include video coding, video streaming, error concealment, and error resilience techniques.

Yu-Chen Sun received the Ph.D. degree in computer science from National Chiao-Tung University, Hsinchu, Taiwan, in 2013.

From February 2013 to June 2013, he was an Intern with Microsoft Research Asia, Beijing, China. Since September 2013, he has been with MediaTek, Inc., Hsinchu, where he is currently a Senior Engineer with the Multimedia Technology Development Division, Corporate Technology Office. His research interests include image/video processing, compression, and communications.

Po-Jui Chiu received the B.S. degree in computer science from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2012. He is currently pursuing the Ph.D. degree at the Department of Computer Science, NCTU.

He joined the Video Information Processing Laboratory, led by Prof. W.-J. Tsai, in 2011 and passed the Ph.D. qualifying examination in 2012. His research interests include robust video coding, error resilience, and concealment techniques in video coding.