國立交通大學

資訊科學與工程研究所

碩 士 論 文

建構於階層式 B 幀的混合式多重描述編碼

Hybrid Multiple Description Coding Based on Hierarchical B Pictures

研 究 生：游顥榆

指導教授：蔡文錦 教授

中 華 民 國 九 十 九 年 八 月

建構於階層式 B 幀的混合式多重描述編碼

Hybrid Multiple Description Coding Based on Hierarchical B Pictures

研 究 生：游顥榆　　　　Student：Hao-Yu You

指導教授：蔡文錦　　　　Advisor：Wen-Jiin Tsai

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis
Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

建構於階層式B幀的混合式多重描述編碼

學生：游顥榆　　　　指導教授：蔡文錦 教授

國立交通大學

資訊科學與工程研究所

摘　要

當視訊資料透過易發生傳輸錯誤的網路傳輸時，多重描述視訊編碼就是一種用於降低錯誤影響的技術。在多重描述編碼上有著許多的方法，都勢必考量到編碼效能與錯誤恢復，並在其中取得一個最好的平衡點。而在此篇論文中，我們提出了一種架構於階層式B幀混合式的多重描述編碼技術。

在混合式多重描述編碼架構模型上，我們利用了不同的多重描述編碼方法切割在不同架構下的幀，如重複法、空間域切割和時間域切割，考慮到不同階層下之不同重要性的幀，將其視訊來源編碼分割產生兩個描述子，當有描述子發生錯誤或遺失時，混合式多重描述編碼以不同的預估方式提供不同程度重要性的幀不平等的錯誤恢復。結果顯示，可以使得視訊在有效率的編碼效能下，達到較好的錯誤恢復。在理想無出錯的網路，或在網路隨機封包遺失的情況下，都可以顯示混合式多重描述編碼的優勢。

關鍵字：多重描述編碼、階層式B幀、空間分割、時域分割、重複法

Hybrid Multiple Description Coding Based on Hierarchical B Pictures

Student: Hao-Yu You Advisor: Dr. Wen-Jiin Tsai

College of Computer Science

National Chiao Tung University

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 the unequal importance of the frames at different hierarchical levels. In case of data loss, the model takes advantages of different estimation methods in providing unequal error resilience for the frames with different degrees of importance. As a consequence, better error resilience can be achieved at high coding efficiency. The advantages of the proposed MDC model are demonstrated in error-free and packet loss networks.

Keywords: Multiple description coding (MDC), hierarchical B pictures, spatial splitting, temporal splitting, duplication.

Table of Contents

List of Figures

List2 of Table

Chapter 1 Introduction

Introduction

1.1 Preface

Through the growing of the communication technology, video streaming has recently become a popular field. There had been more and more application services about video streaming being developed and provided, such as , IPTV, peer-to-peer live video and video phone; the scale of these services also becomes larger. Transmitting video streams smoothly to effectively combat network errors is an important subject.

H.264/AVC is one of the most newly introduced video coding standard developed by Joint Video Team founded by ITU-T and ISO/IEC, which has a better video quality and compression efficiency than existing standards, such as MPEG2 and H.263. When transmitting the H.264/AVC encoded bit-stream, as the coding efficiency is higher, the bits of the encoded stream carry more information of the video source, and the bit-stream would be more vulnerable to transmission errors. As a result, there had been a lot of error resilience tools proposed to combat transmitting error; **Table 1-1**from [1] by A. Vetro, J. Xin and H. Sun summarizes recently proposed error resilience tool. These tools are classified into four different groups according to field of categories and their benefits are listed separately. Localization is a technique that can restrain the error to propagate in a limit range; data partitioning separates the encoded bit-stream into different parts, each has unequal importance so that one can protect each part with different levels of security; redundant coding protects the bit-stream with additional data bits, that is when error occurs, the correctly received parts can be used to recover the lost parts; concealment-driven aims to predict lost part of data with the aid of correlation on either spatial or temporal domain. H.264/AVC had incorporated almost all tools in the four categories from Table 1-1: 1) adaptive intra refresh; 2) reference picture selection; 3) multiple reference pictures; 4) data partition of MV, header and texture; 5) Redundant slice; 6) Flexible macroblock order.

Category	Benefit	Tools
Localization	Reduces error propagation	–Resynchronization marker –Adaptive Intra Refresh –Reference picture selection –Multiple reference pictures
Data partitioning	Enables unequal error protection and transport prioritization	–Frequency coefficients –Motion, header, texture
Redundant coding	Enables robust decoding	–Reversible variable-length coding –Multiple-description coding –Redundant slice
Concealment-driven	Enables improved error concealment	–Concealment motion vectors –Flexible macroblock order

Table 1-1 Benefits of error resilience tools according to category. From[1].

Low-bandwidth handheld devices have become more popular and backbone capacities of the Internet has increased, thus for a video streaming service, the client bandwidth varies in a wide range, from hundreds of kilo-bytes to tens of mega-bytes. Clients on hand-held devices such as cell phone, smart phone or PDA, usually have lower bandwidth, while in desktop, higher bandwidth is common. As a result, a service that is adaptive to the varying bandwidth of heterogeneous networks would become more appealing.

Real-time is another important characteristic in video streaming services. A system that utilizes retransmission or feedback channel may result in an unacceptable delay; since retransmitting lost packets would add at least one round-trip time delay, thus the packet would expired its display timeline. In the streaming on P2P network, the receiving of data stream may come from different source peers through different paths, and the path may failed if one peer along the path failed, thus the receiver could constantly losing part of data from some peers. As the failure of peer is not predictable, the part of data which will get lost during transmission does not know a priori. In this circumstance, using unequal error protection would not be effective. If receivers can make use of whatever they received and utilize the appropriate error concealment and/or resilience tools, the system will have a better performance.

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 (FEC) [1], intra/inter coding mode selection [2], layered coding [3], and multiple description coding (MDC) [4]. This paper is concerned with MDC.

1.2 Multiple Description Coding

Multiple description coding is a technique that encodes a single video stream into two or more equally important sub-streams, called descriptions, each of which can be decoded independently. Different from 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 (MDSQ)[4], has been 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.

Most MDC approaches focus on how to generate two descriptors so that each descriptor would have good decoding quality and the overall two channel bit-rate would be minimized. Figure 1-1 shows the conventional MDC system architecture. The encoder encodes the source into two individual descriptors and then sends through two channels. The decoder has multiple decoder states: side decoder and center decoder; when receiving only one descriptor, the side decoder will be responsible to decode the one descriptor bit-stream; if both descriptors were received the center decoder will produce the best quality output.

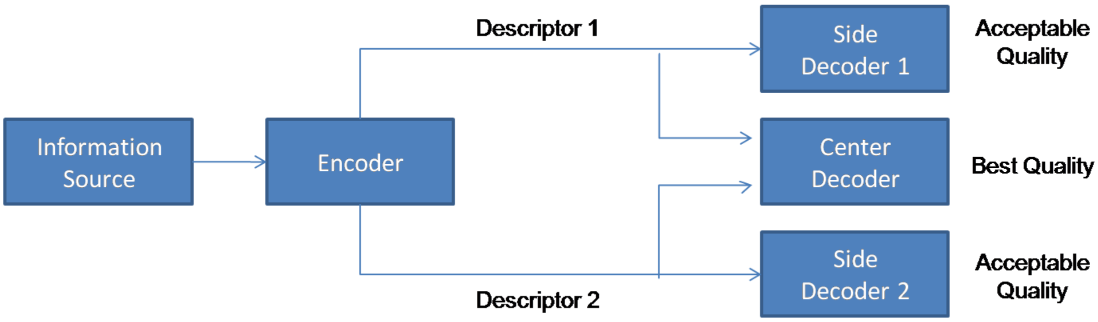


Figure 1-1 Conventional MDC System Architecture

Layered coding, such as scalable video coding(SVC), is a technique that encodes the bit-stream into base layer and enhancement layers; base layer has lower bit-rate and a basic acceptable quality of video, and enhancement layers are used to refine the video quality. If the network traffic is congested, the receiver can receive only base layer; if the bandwidth is sufficient for the receiver to obtain more data, the enhancement layers will be used to further refine the decoding quality. The more enhancement layers are received, the better the decoding quality can be obtained.

SVC has similar features with MDC, but they are different in the view of data importance: SVC treats base layer more important, while the descriptors are equally important in MDC. The different importance of base layer and enhancement layers are due to the fact that enhancement layers cannot be reconstructed without the base layer. In other words, if the base-layer data packets are corrupted, then the corresponding enhancement layers' data packets will be useless. Contrary to SVC, each descriptor of MDC has equal importance, bit-rate and quality.

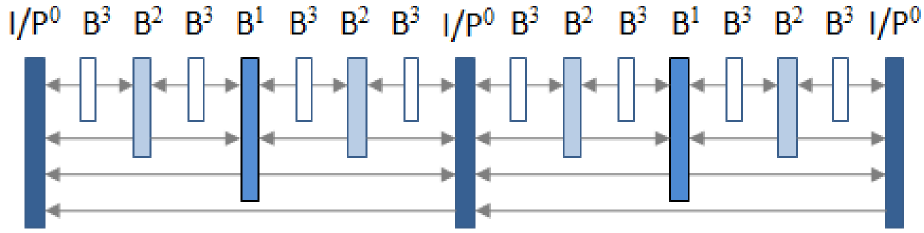
Chapter 2

Related Work

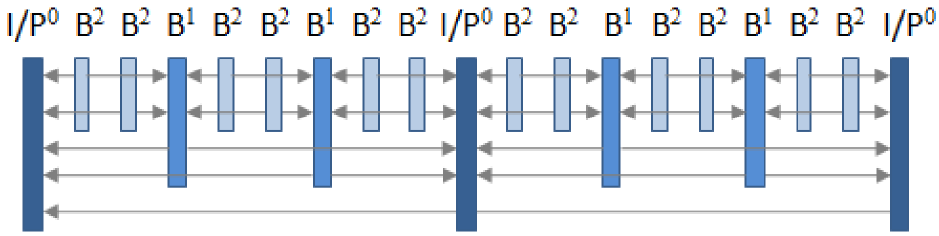
In a hierarchical B-picture prediction framework, the B frames at the coarser temporal levels can to be used as reference for the B frames at the finer temporal levels and therefore, the coding efficiency can be further improved. Compared with classical H.264/AVC prediction structure IBBP, the improvement can be more than 1dB as described in [5]. Even though hierarchical-B picture coding has been widely used in scalable extension of H.264/AVC (SVC) [6] to provide temporal scalability, it is rarely adopted in multiple description coding.

A typical hierarchical prediction framework with 4 dyadic hierarchy stages is illustrated in Figure 2-1(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 picture (GOP). The remaining B frames are hierarchically predicted using two reference frames from the nearest neighboring frames of the previous temporal level. In Figure 2-1, 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. Figure 2-1(b) shows the example of a non-dyadic hierarchical structure with 3 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) [7], 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.



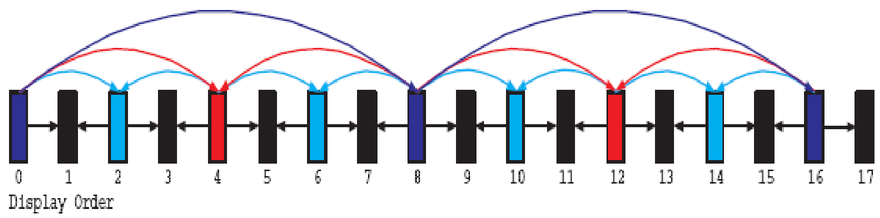
(a) Dyadic



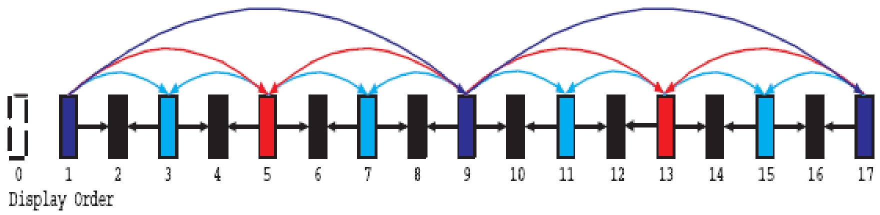
(b) Non-dyadic

Figure 2-1 Hierarchical B Picture Prediction Structure

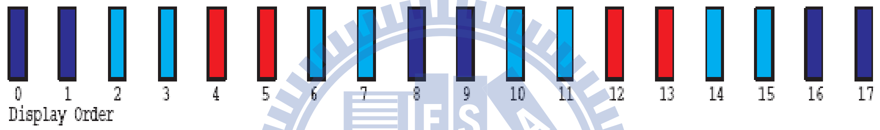
In [8], an MDC based on hierarchical B pictures was proposed, where two descriptions are generated by duplicating the original sequence and then coded by hierarchical B structure with staggered key frames in the two descriptions, as shown in Figure 2-2. By using different QPs at different levels, their approach enables each frame to have two different fidelities in different descriptions. When two descriptors are received, their approach simply selects the frame with high-fidelity, or uses a linear combination of 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.



Hierarchical coding structure of Description 1



Hierarchical coding structure of Description 2



Central Decoder

Figure 2-2 Hierarchical coding structures of the MDVC. From[8]

Chapter 3 MOTIVATION

Motivation

Hierarchical B-frame structure has the characteristic that the each frame has different levels with different importance as mentioned in chapter 2, we refer to the I/P frames at the lowest hierarchical level as key frames, which has the longest reference distance and smaller QP that make the key frame have lots of residual; 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 non-reference B frames (NRB frames) because they are not used as reference, the NRB frames have lots of motion vectors and the biggest QP that make residual are less than other reference frame.

Due to effectiveness in providing error resilience, a variety of researches on different MDC approaches had been proposed afterwards. These approaches can be intuitively classified through the stage where it split the signal, such as, frequency domain [9, 10], spatial domain [11, 12], and temporal domain [13, 14]. In [15], a novel 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 [15] show that, by properly utilizing more than one splitting technique, the novel MDC method can improve error-resilient performance.

According to the two considerations mentioned above: 1) different frame levels structure; 2) hybrid segmentation method in different signal domain, we would like to propose a novel MDC model with hierarchical B-frame architecture, and combine the spatial domain and temporal domain with different levels to make the proposed model more adaptive to the clients from heterogeneous networks.

Chapter 4 HYBRID MODEL

Hybrid Model

In this chapter, a hybrid MDC model is proposed. It is based upon hierarchical B-frame structure and combines different segmentation methods in different domains to generate two descriptors. Based on hierarchical B-frame structure, the error propagation effect in different levels is presented first; then, different segmentation methods adopted in hybrid MDC model are presented, estimation methods are described last.

4.1 Error Propagation Effect

In hierarchical B-picture prediction framework, the frames at lower hierarchical levels can be used as reference for the frames at higher hierarchical levels. Due to this dependency, the decoding quality of a frame strongly depends on the quality of the frames at its previous hierarchical level of the same GOP. The lower level at which a frame is lost, the more frames that will be corrupted. Thus, the error propagation in hierarchical B picture is much more serious than in P frame architecture. For example, in a dyadic hierarchical B picture with four levels, the error in the key frame (level 0) will directly affect seven frames. The error in level-1 and level-2 reference B frames will directly affect four and two frames respectively. For example in Figure 4-1, the error in frame 9 will directly affect frames 5, 7, 8, 10, 11, 13 and 17. The error in frame 5 will directly affect frames 3, 4, 6 and 7. Only the non-reference B frames (level 3) will not directly affect other frames in hierarchical B-picture structure.

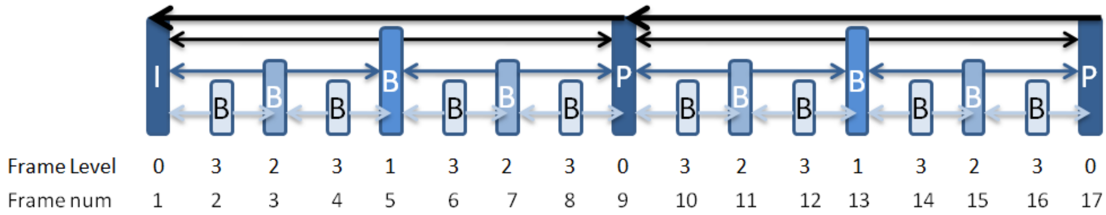
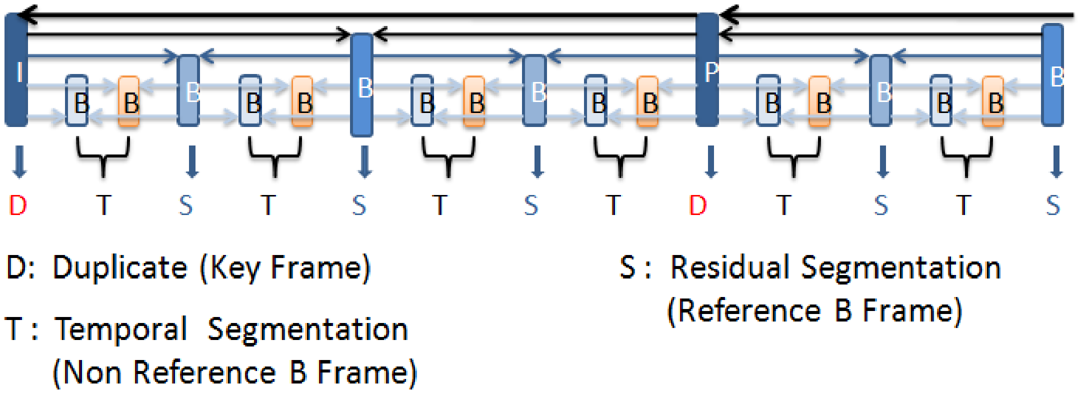


Figure 4-1 Error Propagation of Hierarchical B-picture Structure

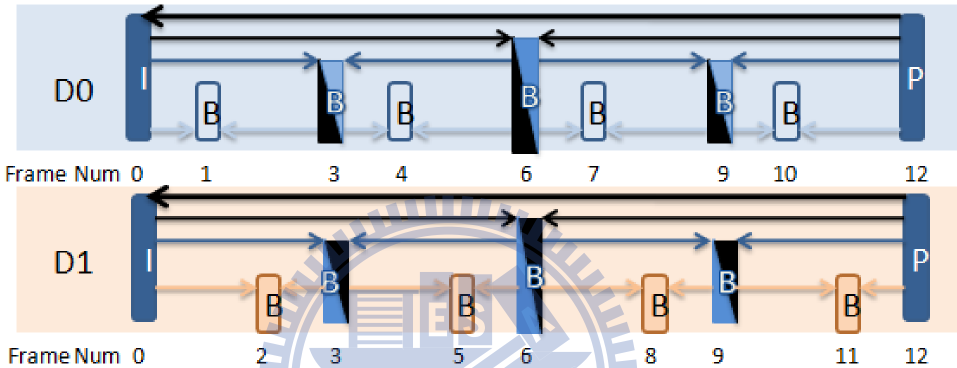
Based on the different error propagation effects of the frames at different hierarchical levels, the proposed hybrid MDC model aims at providing unequal error protections for different frames in hierarchical B-picture structure.

4.2 Hybrid Encoder

The hybrid MDC model uses a four-level non-dyadic hierarchical B picture structure, where there are eight non-reference frames in one GOP. Based on this non-dyadic structure, we apply duplication (denoted by D) for key frames the lowest level which require the highest error resilience; apply spatial splitting (denoted by S) for reference B frames at internal levels; and apply temporal splitting (denoted by T) for non-reference B frames at the highest level which require the lowest error resilience. The proposed hybrid MDC is illustrated in Figure 4-2(a), and the resulting two descriptors are also presented in Figure 4-2(b). Based on level principle, hybrid model has specific segmentation methods in each level. It take good performance in both bit-rate and reconstruction quality.



(a) Original sequence



(b) The resulting two descriptors

Figure 4-2 Proposed MDC based on hierarchical B-picture prediction

The encoder architecture of the proposed MDC model is depicted in Figure 4-3, where after intra-prediction or motion compensation, there the 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 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 go to their respective descriptors.

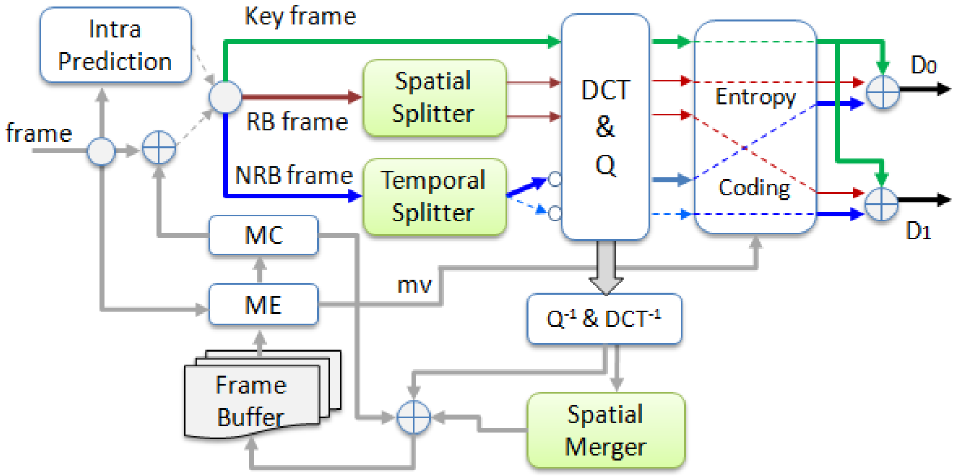


Figure 4-3 Encoder architecture of proposed MDC

4.2.1 Duplicate Key Frame

The duplicate method is designed to duplicate key frames, which are the most important frames in each GOP because they are referenced by lots of B frames. In hybrid model, the key frames duplicated to the every descriptor would cost more bits, but they can avoid the error propagation problem and improve the side decoder reconstruction quality. As show in Figure 4-4, two descriptors have the same I & P frames.

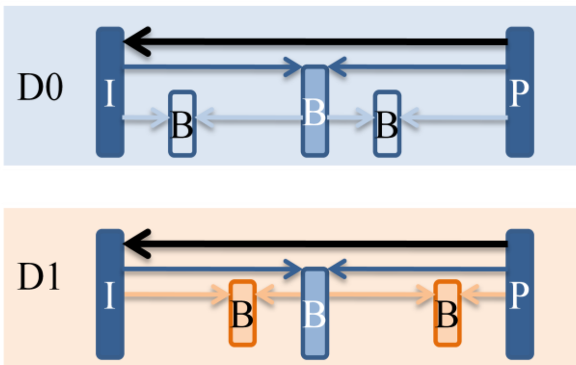


Figure 4-4 Duplicate key frame of the proposed MDC

4.2.2 Residual Segmentation

After motion estimation and compensation in the reference B frames, the spatial splitting, is performed on an 8x8-block basis using polyphase permuting and splitting in the residual domain. The residual data in each 8x8 block will be polyphase permuted inside the block as shown in Figure 4-5. The residual pixels in the 8x8 block are all labeled with a number from: 0, 1, 2 and 3, where for every 2x2 pixels, 0 is labeled on top-left, 1 is on top-right, 2 is on bottom-right, 3 is on bottom-right. The polyphase permuting then rearranges the top-left pixel to the top-left 4x4 block, top-right pixel to the top-right 4x4 block, and so on. After permuting, the pixels labeled with the same number are grouped into the same 4x4 block, the four 8x8 blocks in each macroblocks are all permuted in the same way.

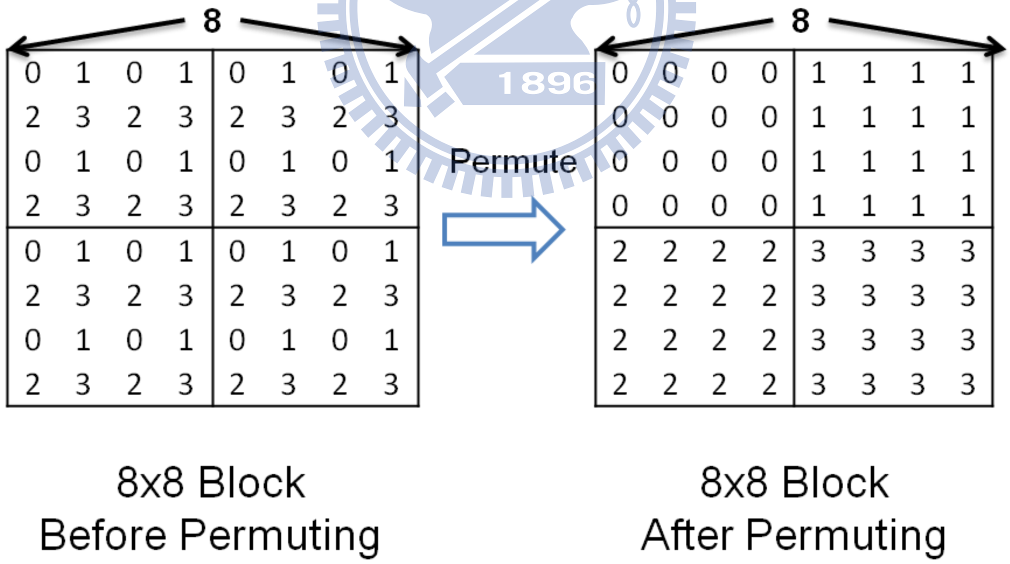


Figure 4-5 Polyphase Permuting of a 8x8 Block

The splitting process as shown in Figure 4-6 is performed to split each 8x8 block into two 8x8 blocks, called residual 0 (R0) and residual 1 (R1), each carries two 4x4 blocks chosen in diagonal: top-left and bottom-right 4x4 residual blocks belong to one

8x8 block, while top-right and bottom-left blocks belong to the other 8x8 block. For each 8x8 block, the remaining two 4x4 blocks with pixels all labeled with'x'in the Figure are given residual pixels all set to zero. The encoder has no need to encode the coefficient of these two all-zero 4x4 blocks.

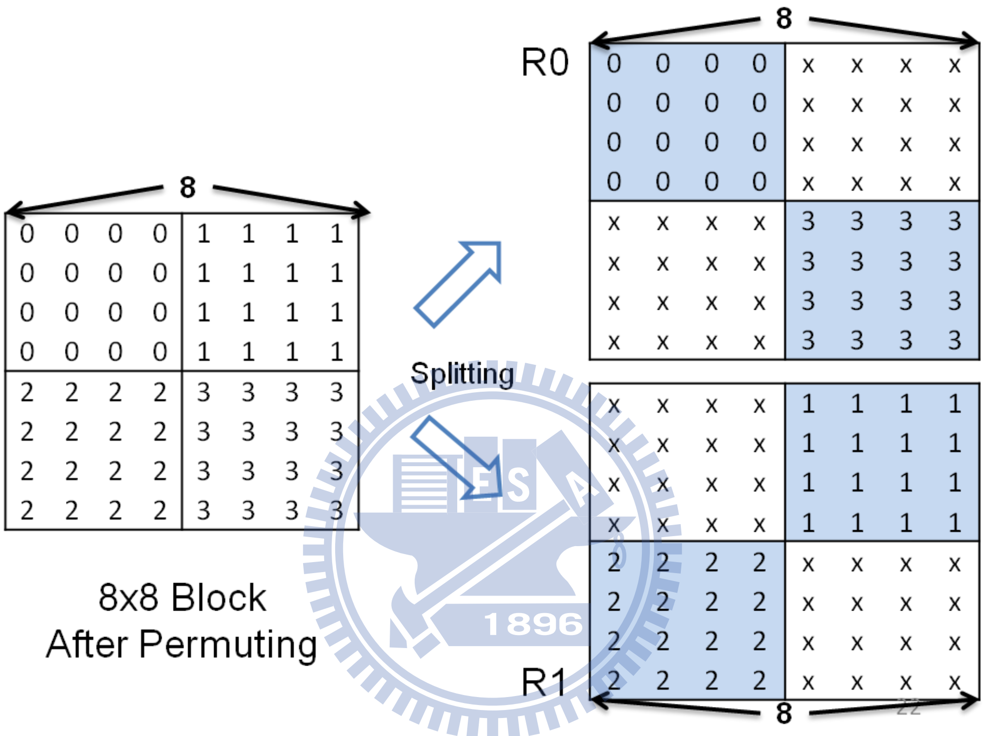


Figure 4-6 Splitting of a 8x8 Block

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 Figure 4-3. The Spatial Merger first discards the all-zero 4x4 blocks and then adopts *Polyphase Inverse Permuting* (the reversed process of Figure 4-6) to reconstruct the original 8x8 blocks.

4.2.3 Temporal Segmentation

Temporal segmentation method split the non-reference frames into two descriptors. As show in Figure 4-7, frames 1, 4, 7 and 10 are assigned to descriptor D0; while frames 2, 5, 8 and 11 are assigned to descriptor D1. Since temporal segmentation the frame rate of each descriptor to 2/3 of original frame rate, it results in minimum bit-rate redundancy, compared to duplication and spatial segmentation. However, with temporal segmentation, one descriptor loss will cause whole frame loss for some frames, and therefore requires efficient error concealment method to prevent performance degradation. The details of the error concealment method will be discussed in section 4.4.3.

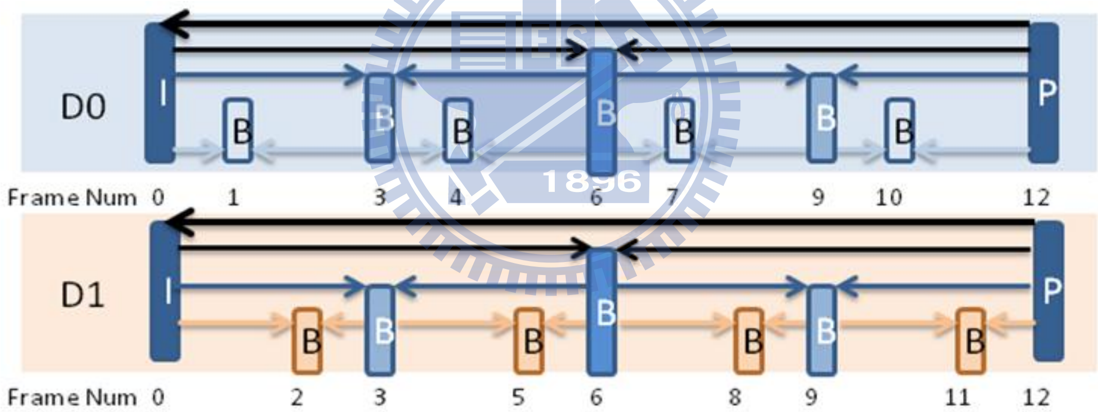


Figure 4-7 Temporal segmentation of the proposed MDC

4.3 Hybrid Decoder

The decoder architecture of the proposed MDC model is depicted in Figure 4-8, where the two descriptors, D_0 and D_1, are first entropy decoded, de-quantized, and inversely transformed 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.

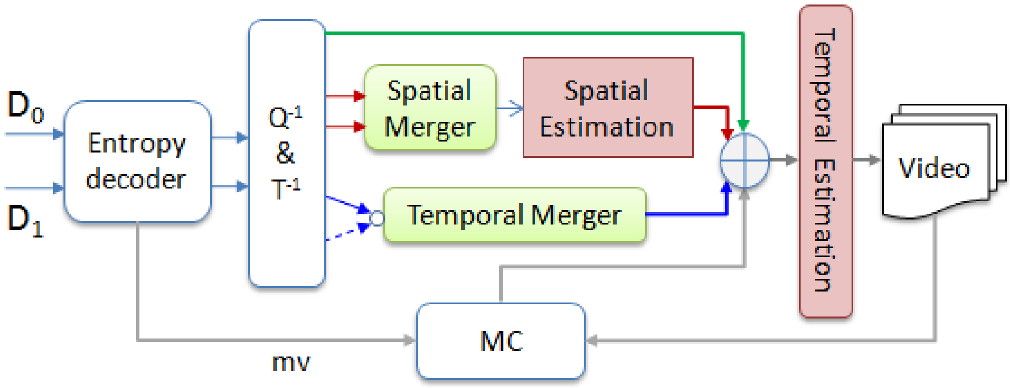


Figure 4-8 Decoder architecture of proposed MDC

4.4 Estimation of Lost Description

Taking advantages of different MDC methods applied on the frames at different hierarchical levels, different estimation methods are designed for different frames. Table 4-1 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.

Estimation Method	Hybrid(S+T+D)	
Descriptor loss / Frame level	One	Both
Key frame	D	T
Reference B frame	S	T
Non-reference B frame	T	T

Table 4-1 Summary of the cases for different estimation methods

4.4.1 Duplicate Method

In center decoder, the key frames loss will cause a series error propagation, the duplicate method can avoid this problem; based on the duplicate method in encoder side, the decoder will receive the same key frames in each descriptor; when one description loss, the lost key-frames can be reconstructed by simply using the duplicated version in the other descriptor.

4.4.2 Spatial Estimation Method

Spatial estimation method explores the spatial correlation between residual pixels to estimate the lost part of reference B frame. For example, assuming that R0 and R1 are two descriptions split from the reference B frames, and R1 is lost during transmission, to reconstruct the missing R1, the decoder will apply spatial estimation method when R0 is correctly received. After the polyphase inverse permutation of R0, the residual pixels are distributed like a checkerboard within a macroblock as shown in Figure 4-9, where for each lost residual pixel, four neighboring pixels are available.

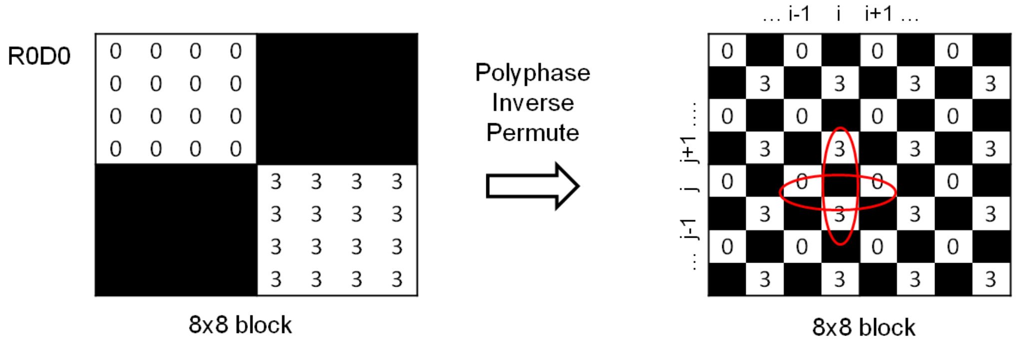


Figure 4-9 Spatial Concealment by Bilinear Interpolation

$$\widetilde{f_{j,i}} = \left(f_{j+1,i} + f_{j-1,i} + f_{j,i+1} + f_{j,i-1}\right)/4 \quad (4.1)$$

The spatial estimation method uses bilinear interpolation to reconstruct the lost residual pixels, as shown in Equation (4.1) where $\tilde{f}_{j,i}$ is the reconstructed value of the residual pixel in column i and row j. Since neighboring residual pixels have high spatial correlation, spatial estimation will be efficient.

4.4.3 Temporal Estimation Method

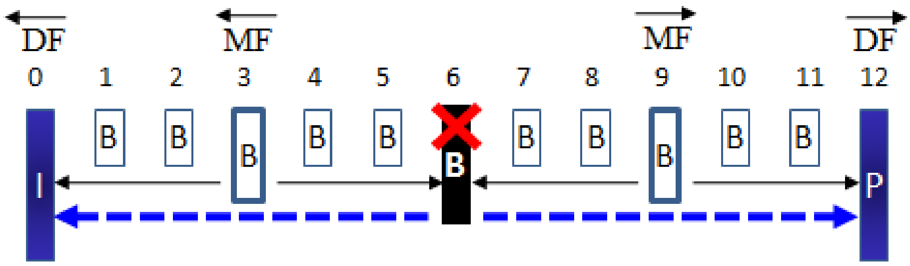
For whole-frame loss, each block in the lost frame is recovered based on temporal correlation since all the neighboring blocks are also lost. We refer to the pictures whose pixels are used to predict the missing pixels as the data prediction frame (DF) and the pictures whose block motions are used to 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 bi-directional motion-compensated signal to recover missing pixels. Thus, we need to select two DFs: a backward DF and a forward DF (denoted by $\overleftarrow{\mathrm{DF}}$ and $\overrightarrow{\mathrm{DF}}$, respectively); and two MFs: a backward MF and a forward MF (denoted by $\overleftarrow{\mathrm{MF}}$ and $\overrightarrow{\mathrm{MF}}$, 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 pictures nearest to it 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 prediction structure, for a lost picture, we select its reference frames in backward and forward directions as its $\overleftarrow{\mathrm{DF}}$ and $\overrightarrow{\mathrm{DF}}$, 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 to be located earlier than the lost picture in decoder order. Instead of using temporal direct mode (TDM) technique which adopts reference pictures as MFs, we choose pictures at higher levels because these pictures are temporally closer to the lost picture in display order. As an example in Figure 4-10(a), if the frame 6 is lost, we will select its reference frames (0 and 12) as its DFs, but select frames 3 and 9 as its MFs. Similarly, if frame 3 is lost, we will select frames 0 and 6 as its DFs, but frames 2 and 4 as its MFs. This selection policy is applied to all frames except NRB frames which are at the highest level within the hierarchical structure. For NRB frames, the MFs are selected from their reference frames at the previous level of the lost picture. Figure 4-10(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 $\overleftarrow{MF}$ and $\overrightarrow{MF}$ as :

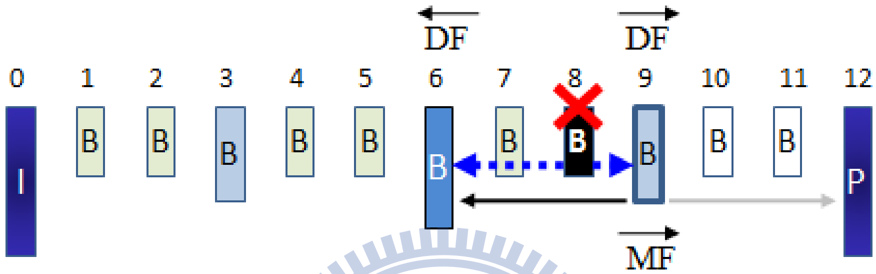
$$\overleftarrow{MF} = \begin{cases} F_{t_{nb}}^{l+1} & \text{for } l_{base} \leq l < l_{top} \\ F_{t_{ref}}^{l-1} & \text{if exists for } l = l_{top} \end{cases} \tag{4.2}$$

$$\overrightarrow{MF} = \begin{cases} F_{t_{nf}}^{l+1} & \text{for } l_{base} \leq l < l_{top} \\ F_{t_{ref}}^{l-1} & \text{if exists for } l = l_{top} \end{cases} \tag{4.3}$$

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. $F_{t_{ref}}^{l-1}$ denotes the F_t^l's reference frame at level l-1.



(a) DF and MF selection for RB frames



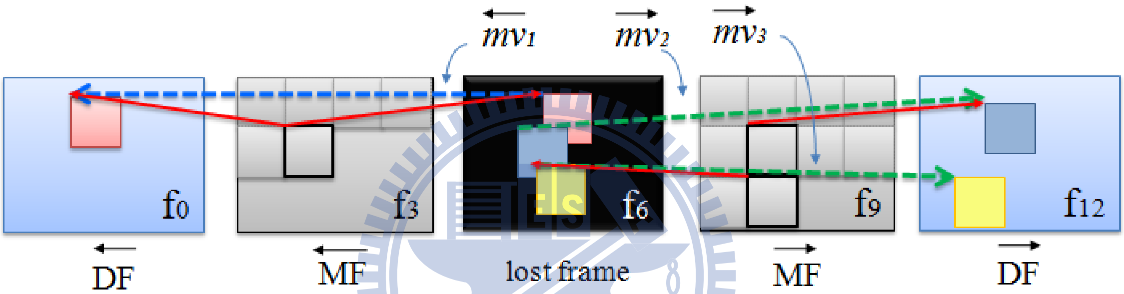
(b) DF and MF selection for NRB frames

Figure 4-10 DF and MF selection for temporal estimation method

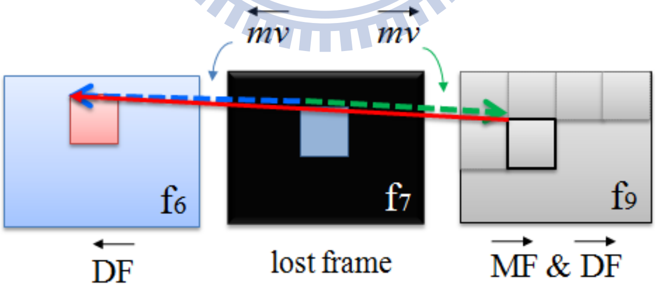
After the determine of DFs and MFs, for every block in $\overleftarrow{MF}$ and $\overrightarrow{MF}$, its motion vectors(s) are composed, extrapolated, or interpolated so that the motion vectors pointing to $\overleftarrow{DF}$ and $\overrightarrow{DF}$ from the lost frame can be obtained (called $\overleftarrow{mv}$ and $\overrightarrow{mv}$, respectively).. For RB frames, since its MFs are located in between DFs and the lost frame (see Figure 4-10(a)), the motion vectors in MFs are composed (e.g., $\overleftarrow{mv}_1$ in Figure 4.11(a)) if the block has two motion vectors, or extrapolated (e.g., $\overrightarrow{mv}_2$ and $\overrightarrow{mv}_3$) if the block has only one motion vector. For NRB frames, since one MF is used for two DFs located on different sides of the lost picture (see Figure 4-10(b)), the motion vectors in the MF are interpolated as shown in Figure 4-11(b). In this way, the pixels in the lost picture can be classified into four types: the pixels with one or more $\overleftarrow{mv}$, the pixels with one or more $\overrightarrow{mv}$, the pixels with both $\overleftarrow{mv}$ and $\overrightarrow{mv}$, and the pixels without $\overleftarrow{mv}$ and $\overrightarrow{mv}$. For a pixel P in the lost picture, we recover it by the predicted signal $\tilde{P}$ obtained as follows

$$\tilde{P}(x) = \begin{cases} \sum_i \overleftarrow{DF}(x + \overleftarrow{mv}_i) & ;\ \text{if P has } \overleftarrow{mv} \text{ only} \\ \sum_i \overrightarrow{DF}(x + \overrightarrow{mv}_i) & ;\ \text{if P has } \overrightarrow{mv} \text{ only} \\ w_0 \sum_i \overleftarrow{DF}(x + \overleftarrow{mv}_i) + w_1 \sum_i \overrightarrow{DF}(x + \overrightarrow{mv}_i) & ;\ \text{if P has } \overleftarrow{mv} \text{ and } \overrightarrow{mv} \\ w_0 \overleftarrow{DF}(x)) + w_1 \overrightarrow{DF}(x)) & ;\ \text{otherwise} \end{cases} \quad (4.4)$$

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



(a) Motion composition and extrapolation for a lost RB frame



(b) Motion interpolation for a lost NRB frame

Figure 4-11 Temporal estimation using bi-directional predicted signal

Chapter 5 EXPERIMENTAL results

Experimental Results

In this chapter, the performance of the proposed temporal estimation method is examined first, and then the Hybrid MDC model is examined by its packet-loss performance, rate-distortion performance, center-decoder performance, and the error propagation effect. The experimental results of the five models: Hybrid(S), Hybrid(S+T), Hybrid(S+T+D), Modified QP[8] and Default QP[8]; the four test sequences: mobile, news, foreman and coastguard with CIF (352x288) resolution are used for performance evaluation. These models are implemented in H.264/AVC reference software, JM 16[16]; the intra period is 48 frames and the hierarchical level is 4.

5.1 Performance of Temporal Estimation Method

To show the performance of the proposed temporal estimation method, experiments were encoded using a dyadic hierarchical structure with 4 levels. We compared the proposed method with WTDM_EC[17] which is a method based upon temporal direct mode (TDM) of H.264/AVC for the error concealment of whole frame loss in hierarchical B-picture prediction structure. Three different loss rates (PLRs) are used in our experiments and the results presented in Figure 5-1 are the averages of

100 independent simulation runs.

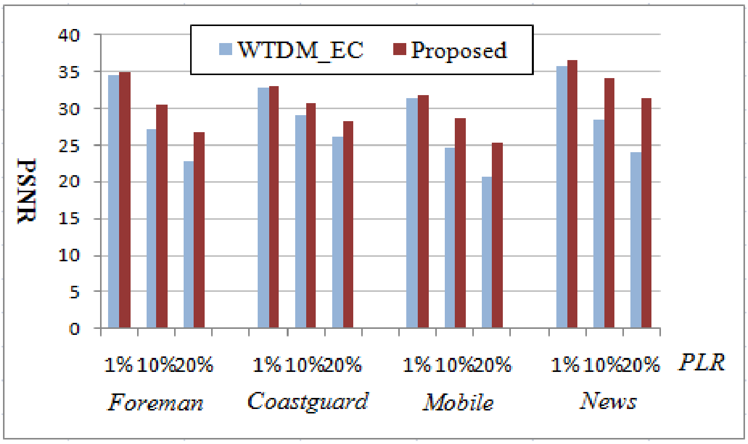


Figure 5-1 Performance of temporal estimation methods. (QP=28)

From the results it is observed that the proposed temporal estimation outperformed the WTDM_EC method for all the sequences under different packet loss rates. The performance gaps become large as the loss rates increase. The reason is due to that, for every block in the lost picture, WTDM_EC predicts its 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 Figure 5.2, where the subjective quality comparison of concealed frames at different hierarchical levels in News is presented. Experiments were conducted independently for each case in Figure 5.2, namely, there is no error propagation implemented among them. As shown in Figure 5.2 (a), for the frame at level 3, the visual quality of the concealed frame by WTDM_EC and that by proposed method are almost the same. The PSNR difference between them is only 0.2dB. In Fig.10(b), the quality difference of the two concealed frames at level 2 becomes larger.

There is a noticeable noise on the dancer of the frame by WTDM_EC and the PSNR difference of using two methods is about 1.7dB. In Figure 5.2(c) and (d), the concealed frames by WTDM_EC are obviously blurred. This is due to inadequate motion vectors obtained from co-located blocks in a selected frame which is far from the lost frame if the lost frame is at low hierarchical levels. By using the proposed method, the quality improvement can be up to 2.9 dB at level 1 and 2.5dB at level 0.

(a) Subjective quality of the 2nd frame at level 3: WDTM_EC (left): 34.09dB, Proposed (right): 34.29dB, Correct: 36.68dB.

(b) Subjective quality of the 4th frame at level 2: WDTM_EC (left): 33.88dB, Proposed (right): 35.51dB, Correct: 37.53dB.

(c) Subjective quality of the 7th frame at level 1: WDTM_EC (left): 31.33dB, Proposed (right): 34.27dB, Correct: 37.42dB.

(d) Subjective quality of the 13th frame at level 0: WDTM_EC (left): 27.7dB, Proposed (right): 30.25dB, Correct: 38.44dB.

Figure 5-2 Subjective quality comparison of the frames at different hierarchical levels of News sequence

5.2 Performance of Proposed MDC

In This s 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: hybrid (S), hybrid (S+T), and hybrid (S+T+D). The hybrid (S) stands for the method which adopts spatial splitting only. It applies spatial splitting in the residual domain for all frames, regardless of the hierarchical levels. The hybrid (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 hybrid (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 Zhu et al.'s method [8] 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 Figure 5.3, 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[18] as described in Chapter 2, while the modified QP modifies the QPs of top-level frames to 51 in order to reduce bitrates redundancy. The results in [8] 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 examined. Table 5-1 lists the error concealment methods used by the five MDC methods, where D' means the error concealment method presented in [8], where in the case of one-descriptor loss, a lost frame is recovered by the duplicated version in the other description. D' is distinguished from D because the duplicated frame is at a different level and thus, with different quality fidelity. Since Zhu et al. did not provide solutions for the cases of 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.

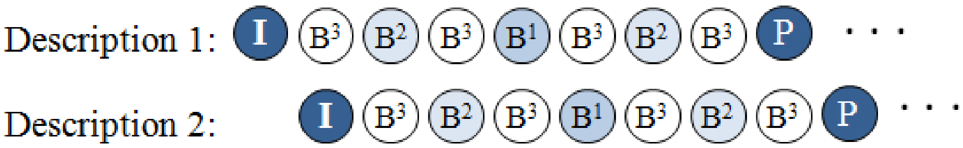


Figure 5-3 Two descriptions with staggered key frames [8]

Estimation Method	Hybrid(S+T)		Hybrid(S+T+D)		Hybrid(S)		Modified QP[8]		Default QP[8]	
Frame level \ Descriptor loss	One	Both	One	Both	One	Both	One	Both	One	Both
Key frame	S	T_e	D	T_e	S	T_e	T_e	T_e	D	T_e
Reference B frame	S	T_e	S	T_e	S	T_e	T_e	T_e	D	T_e
Non-reference B frame	T_i	T_i	T_i	T_i	S	T_i	T_i	T_i	D	T_i

Table 5-1 Summary of the cases for different estimation methods

All the methods encode video sequences with hierarchical B-picture structure of four levels to generate two descriptors. The three proposed methods adopt a non-dyadic structure which allows temporal splitting on NRB frames; 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, Figure 5.4 shows rate-distortion performance of single description coding (SDC) with these two different structures. As observed in Figure 5.4, the two structures perform equally well for low-motion sequences, foreman and news; while the dyadic structure outperforms the non- dyadic one slightly for high-motion sequences, mobile and coastguard. Namely, the three proposed methods are based upon a structure with slightly worse coding performance in comparison.

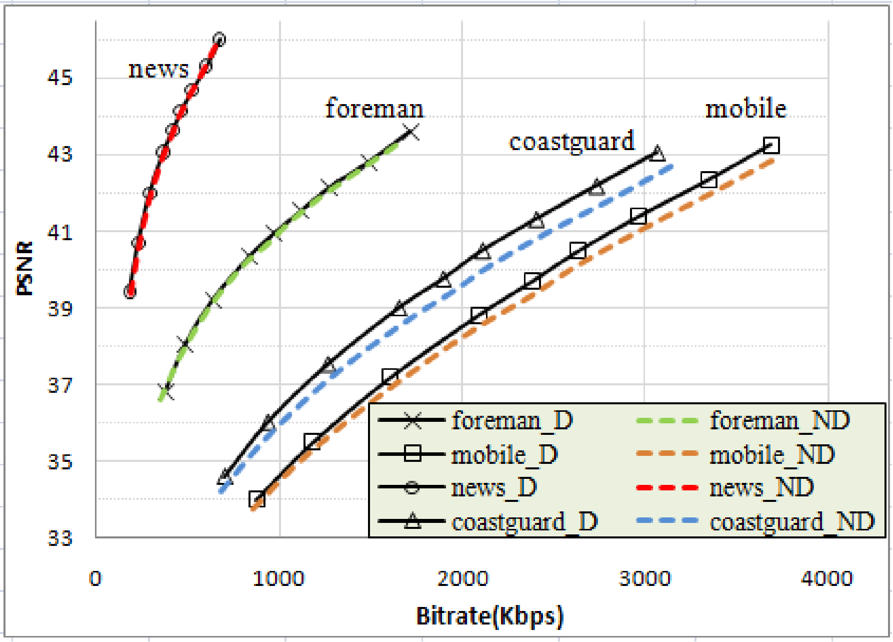
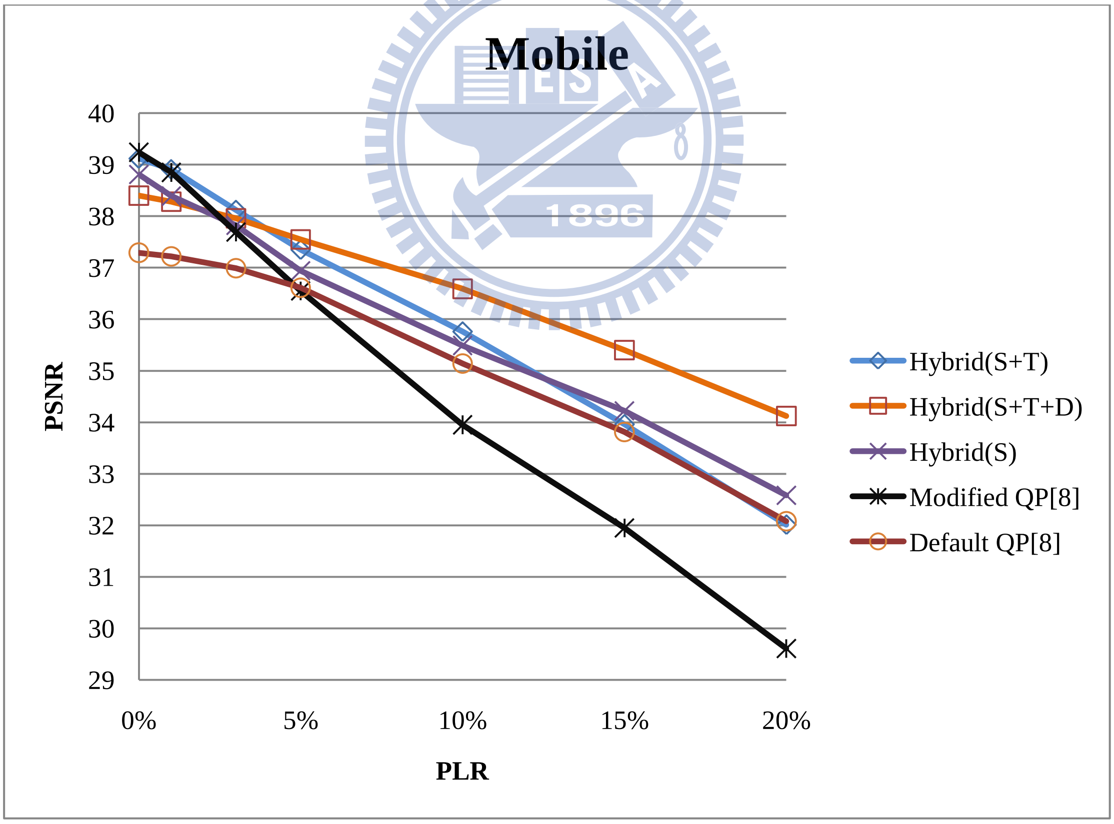


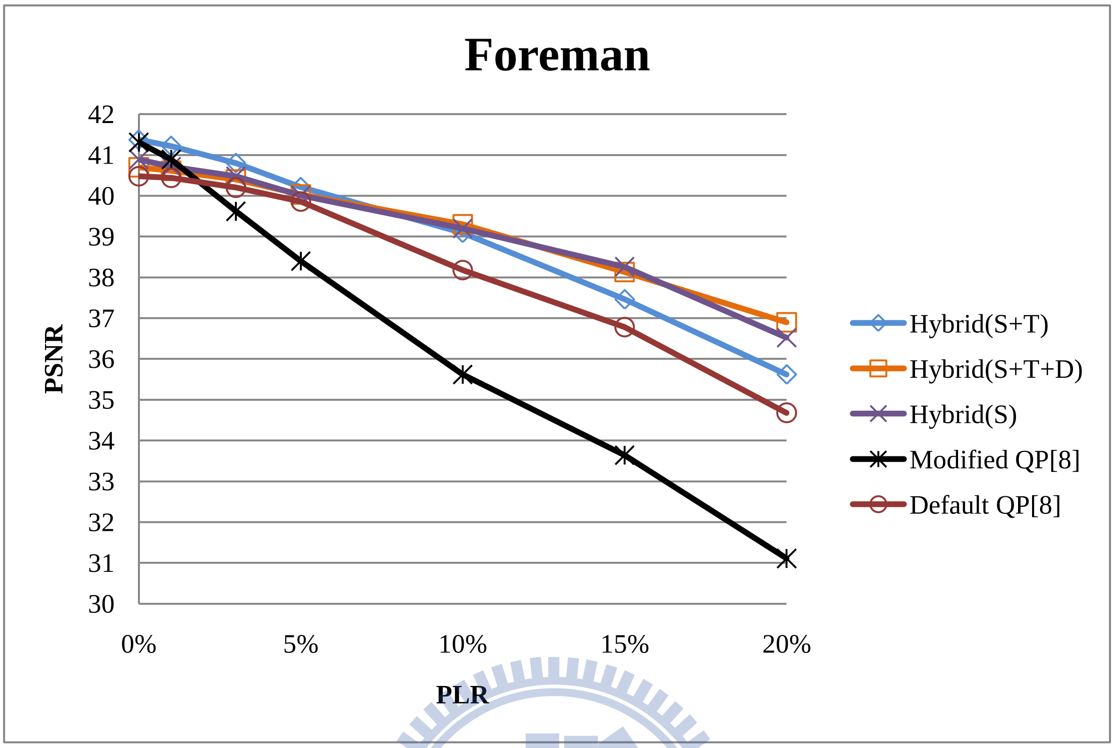
Figure 5-4 R-D performance of dyadic (D) and non-dyadic (ND) structures

5.2.1 Packet Loss Performance

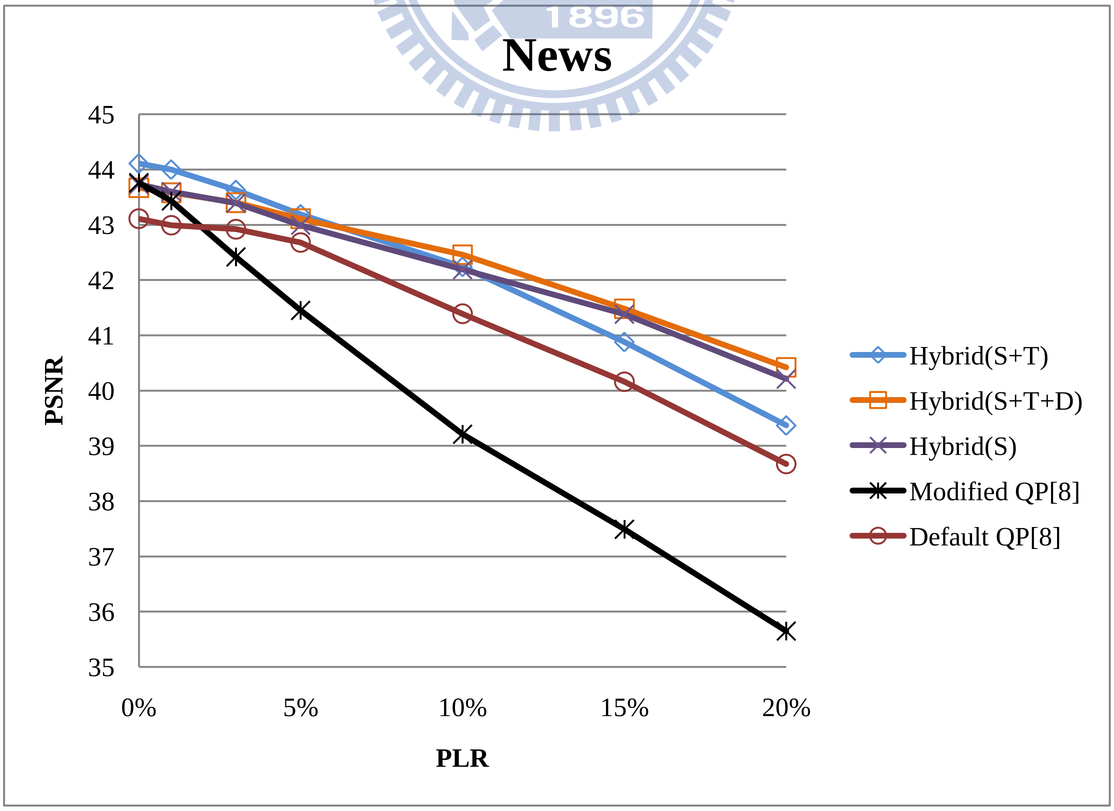
The five MDC methods were examined in a packet-loss scenario where various packet-loss rates, ranging from 0% to 20%, are adopted. We use one packet for each frame of each descriptor. Each packet is lost randomly and independently. Figure 5.5 shows the PSNR as a function of packet-loss rate (PLR) for four CIF test sequences, Foreman, Coastguard, Mobile, and News. The results are the averages of 100 independent runs. It can be seen that, in the case of PLR=0%, both modified QP and hybrid (S+T) have the best performance and default QP has the worst performance among all methods. This is due to 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 bitrates and thus achieve a better performance at PLR=0. We will further discuss the error-free performance in 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 degrade the performance. Compared with modified QP, default QP performed much better as PLR increases. However, the duplication mechanism 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 as well as the high bitrates-redundancy result in the worse performance of default QP, compared with the three proposed methods.



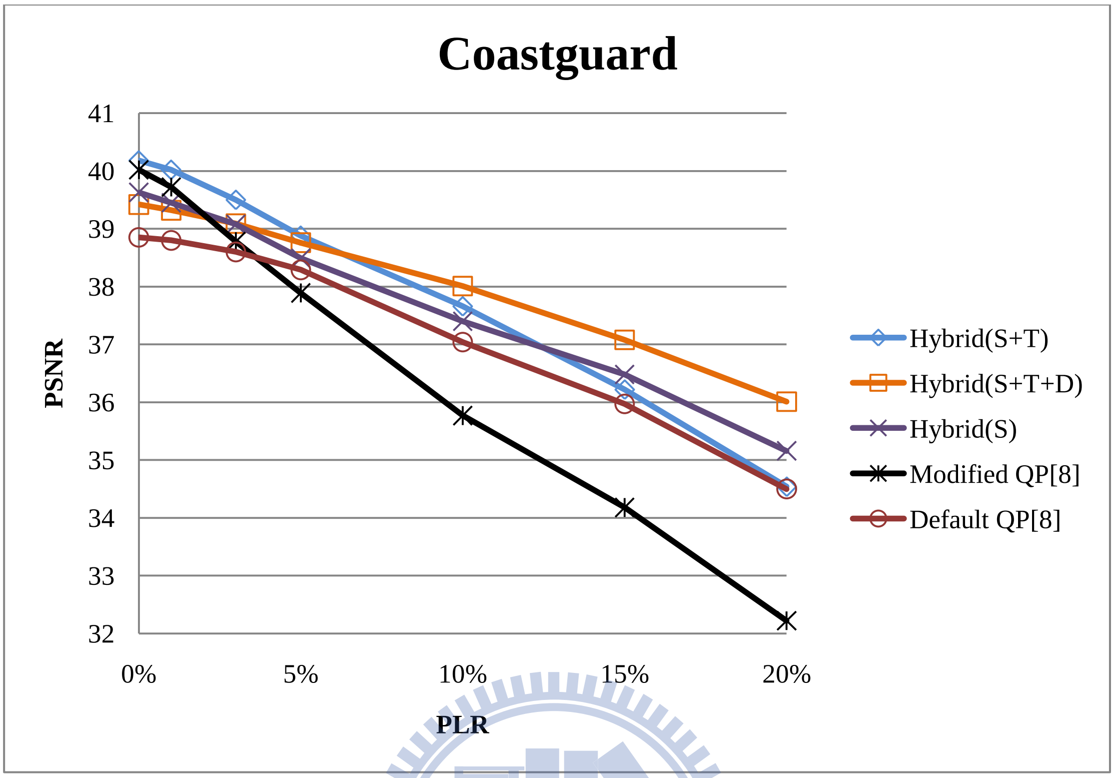
(a) Mobile sequence (CIF@2800kbps)



(b) Foreman sequence (CIF@1500kbps)



(c) News sequence (CIF@700kbps)



(d) Coastguard sequence (CIF@2800kbps)

Figure 5-5 Performance comparison in packet-loss environments

Among these methods, hybrid (S+T+D) has the overall best performance. Although hybrid (S) performed slightly better than hybrid (S+T+D) for sequences foreman and news, it performed much worse than hybrid (S+T+D) for sequences mobile and coastguard. This is due to that spatial estimation cannot recover lost data well for these sequences when there is packet loss. With temporal splitting on NRB frames, hybrid (S+T) is able to reduce bit-rate redundancy and hence, improve the performance at low PLRs, but still cannot solve the problem. By duplicating key frames, the hybrid (S+T+D) can alleviate this problem effectively because key frames 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 is able to suppress error propagation effectively and improve the performance substantially. We will discuss the error propagation issue further in later section. To summarize, the overall results demonstrate that, by adopting spatial splitting, temporal splitting and duplication for the frames at different levels, the hybrid (S+T+D) optimizes the trade-off between bit-rate redundancy and error-resilient capability and therefore, achieves the best performance among the five MDC methods.

5.2.2 Error Free Performance

In this section, we compare the performance of the five MDC methods and single description coding (SDC) in error-free environments. Experiments were conducted for four sequences and the results are presented in Figure 5.5. 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 hybrid (S) and hybrid (S+T+D) perform slightly worse than modified QP and hybrid (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

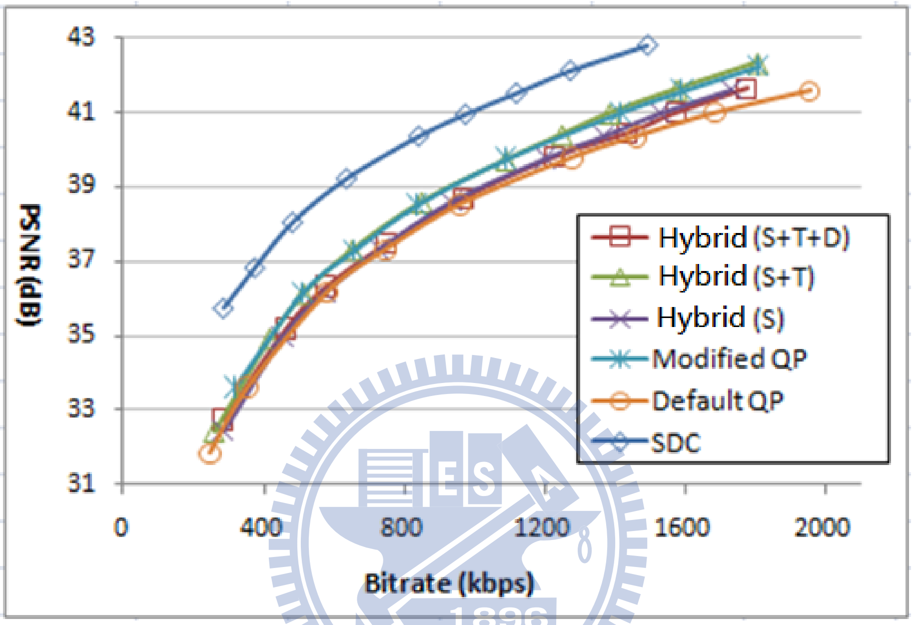
$$\mathrm{BR}(M_i) = \frac{\mathrm{R}_{2D}(M_i) - \mathrm{R}_{SDC}}{\mathrm{R}_{SDC}} \times 100\%$$

where R_{2D} is the total bitrates of two descriptions and R_{SDC} is the bit-rate for SDC with hierarchical B-picture structure. Table III show 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.

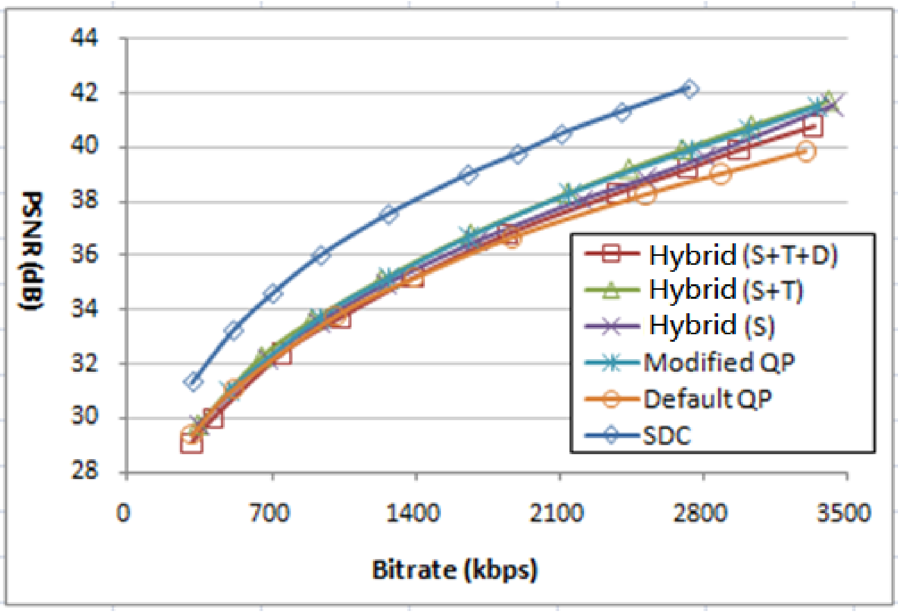
QP 28	Foreman		News		Mobile		Coastguard	
SDC PSNR	35.73 dB		38.13 dB		32.57 dB		33.22 dB	
	R	PSNR	R	PSNR	R	PSNR	R	PSNR
Hybrid(S+T)	1.45	34.92 dB	1.51	37.51 dB	1.26	31.95 dB	1.27	32.22 dB
Hybrid(S+T+D)	1.6	35.19 dB	1.65	37.78 dB	1.45	32.05 dB	1.47	32.38 dB
Hybrid(S)	1.57	34.98 dB	1.61	37.36 dB	1.36	31.97dB	1.32	32.21 dB
Modified QP[8]	1.76	36.16 dB	1.79	38.32 dB	1.72	33.31 dB	1.83	33.69 dB
Default QP[8]	2	36.16 dB	2	38.32 dB	2.02	33.30 dB	1.98	33.69 dB

Table 5-2 Error-free Performance of MDC methods (QP=28)

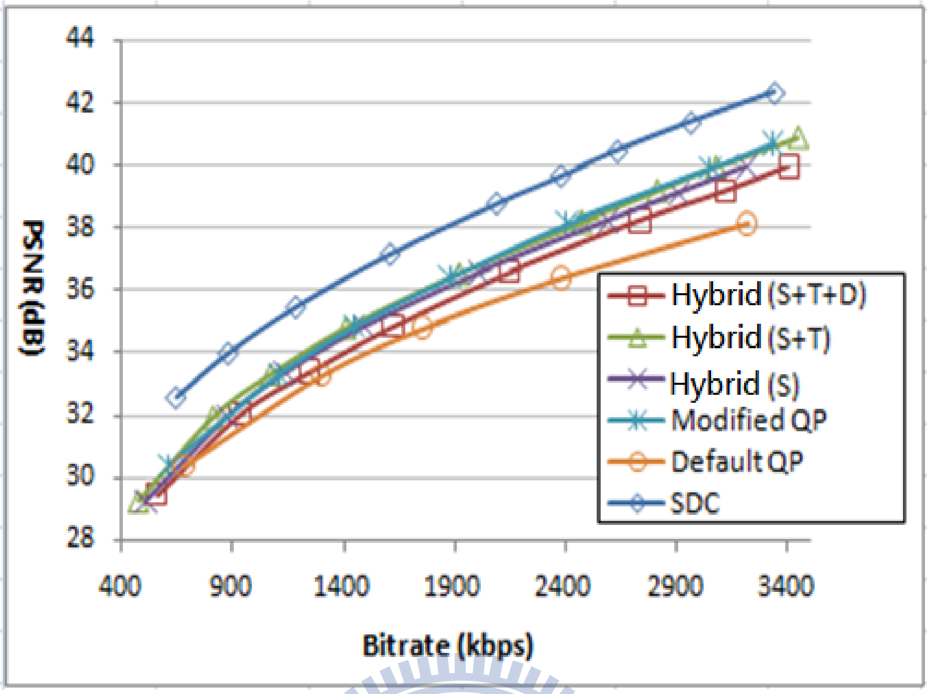
This is due to that, for each frame in the sequence, default QP adopts a combination of its high-fidelity and low-fidelity frames from the two descriptors to produce a better reconstruction with PSNR even higher than SDC's. Even though default QP achieves the best PSNR performance, it suffers from substantial BR increase because it duplicates the entire sequence to two descriptions and hence, its bitrates is almost twice the bitrates of SDC. In Figure 5.6, 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 keeps the same PSNR as default QP as shown in Table 5-2. This explains why the modified QP has the best rate-distortion performance in Figure 5.6. The improvement of error-free performance is, however, at the cost of reducing the error robustness as shown in Figure 5.5, where modified QP has the worst packet-loss performance. By using splitting techniques, Table 5-2 shows that the three proposed methods can further reduce BRs, but also decrease the PSNRs, resulting in slightly worse or equal rate-distortion performance in Figure 5.6, when compared with modified QP. To summarize, the overall results in Figure 5.5 and Figure 5.6 demonstrate that, with slight degradation in error-free performance, the hybrid methods can improve the packet-loss performance significantly, especially when the hybrid (S+T+D) is employed.



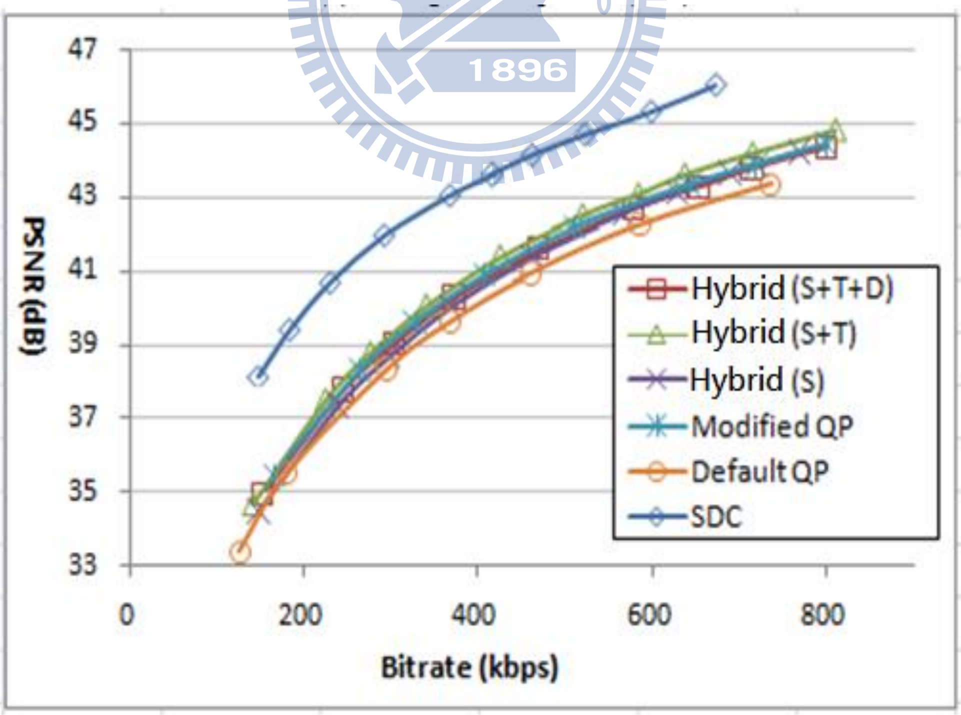
(a)Foreman sequence (CIF)



(b) Coastguard sequence (CIF)



(c) Mobile sequence (CIF)



(d) News sequence (CIF)

Figure 5-6 Rate-distortion performance comparison in error-free environments

The Figure 5-7 shows the side and center performance, the hybrid model(S+T+D) has better side performance than hybrid(S) and hybrid(S+T). The Side-Default QP method is SDC, that it has best PSNR in error free environment.

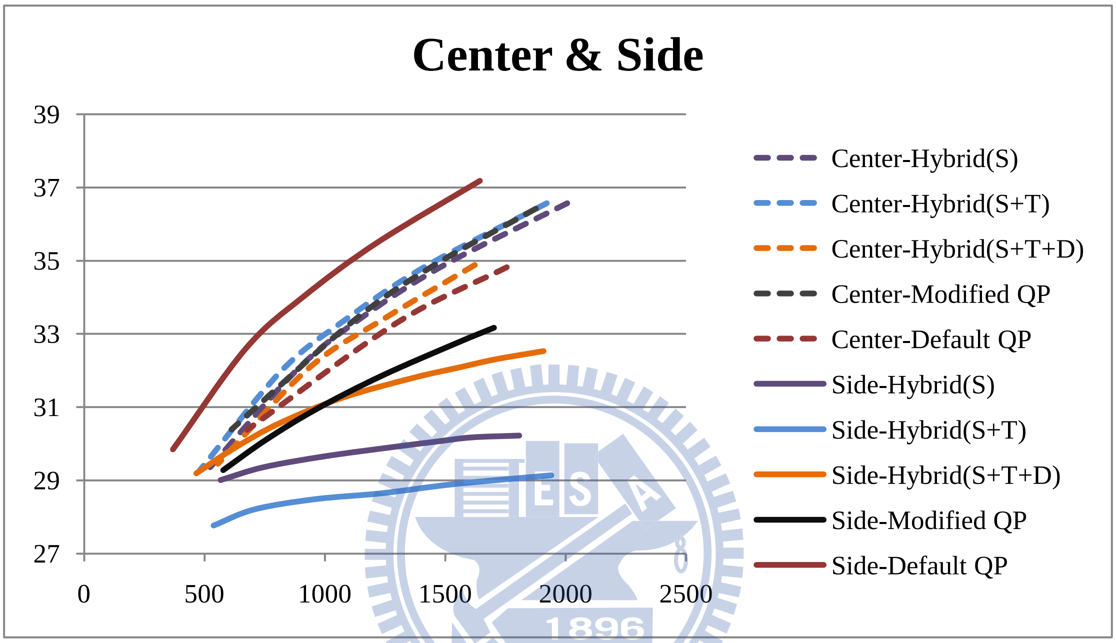


Figure 5-7 Center and side performance (@mobile_cif)

5.2.3 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 non-dyadic hierarchy structure and 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 non-dyadic structures will start from the same key frames. Let B^i

denote the B-frame at level *i*. As illustrated in Figure 5.7, four frames at level 3 of the non-dyadic 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 Figure 5.7, sequence A shows the frame numbers in the original sequence, while sequence B lists the selected frames for frame-by-frame comparison in Figure 5.8.

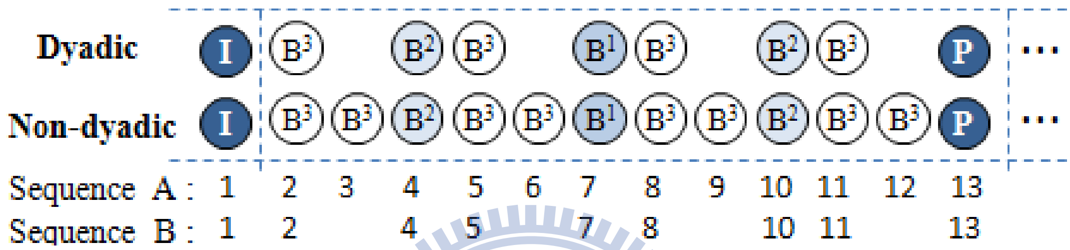
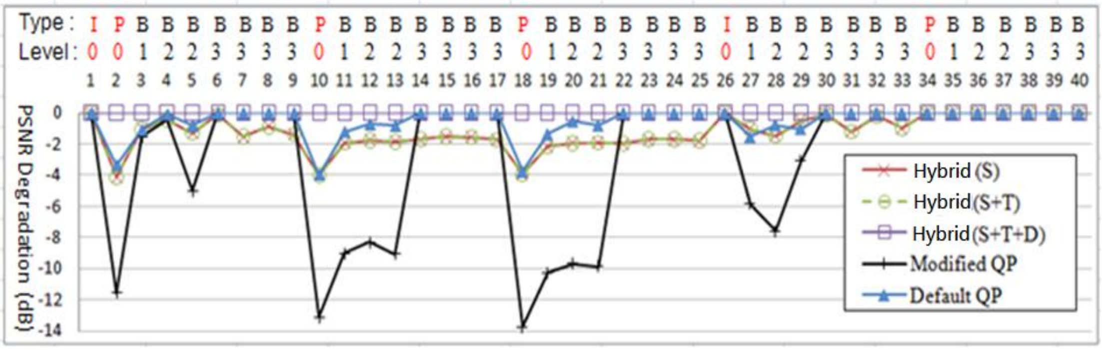


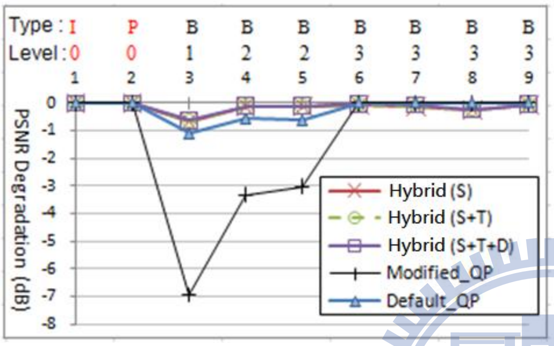
Figure 5-8 Frames used in frame-by-frame comparison

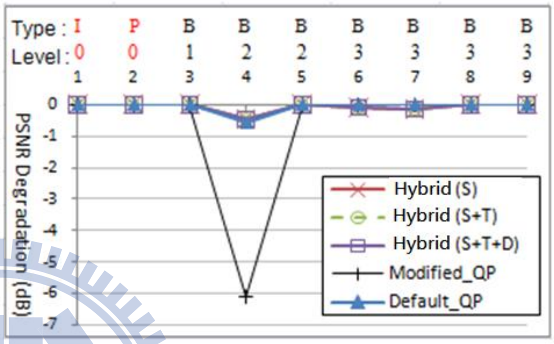
We use one packet for each frame in each descriptor and the error propagation results of a single packet loss at different hierarchical levels are shown in Figure 5.7, where we renumber the selected frames according to decoding order. Figures 5.7 (a), (b) and (c) show the results of the frame loss at levels 0 (the 2nd frame), 1(the 3rd frame) and 2(the 4th frame), respectively. In these figures, y-axis denotes PSNR degradation and x-axis the frame number (in decoding order). From Figure 5.8 (a) it is observed that almost all the methods suffer from severe error propagation for the P-frame loss, except the hybrid (S+T+D). This is due to that the hybrid (S+T+D) duplicates key-frames to two descriptors and therefore, when only one of them is loss, the other one can be used to reconstruct the frame without quality degradation and error propagation. In both hybrid (S) and hybrid (S+T) methods, key-frames are spatially split to 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 to mention that even though the quality degradation of default QP in Figure 5.8 (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 hybrid (S) and hybrid (S+T) as shown in Figure 5.5. 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 Figure 5.8 (a), the results in Figure 5.8 (b) and (c) show that when the fame 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 won't cause error propagation, the result for level 3 is not presented. To summarize, the results in Figure 5.8 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 unequal error-protection, the hybrid (S+T+D) adopts duplication (high bit-rate redundancy) for key-frames; spatial splitting (low redundancy) for level-1 and level2 RB frames; and temporal splitting (even lower redundancy) for NRB frames. As a result, the hybrid (S+T+D) optimizes the trade-off between coding efficiency and the error resilience and achieves the overall best performance.



(a) Frame loss at level 0 (the 2nd frame is lost)





(b) Frame loss at level 1 (the 3rd frame is lost) **(c) Frame loss at level 2 (the 4th frame is lost)**

Figure 5-9 Frame-by-frame comparison

Chapter 6 Conclusion

Conclusion

A hybrid 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 hybrid model is able to optimize the tradeoff between coding efficiency and error resilience. In case of data loss, the model takes advantages 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 (hybrid(S), hybrid (S+T), and hybrid (S+T+D)) and two methods (default QP and modified QP) in [8]. The experimental results show that the hybrid (S+T+D) achieves the overall best performance among these five methods.

Reference

[1] A. Vetro, J. Xin, H. Sun, "Error Resillence Video Transcoding for Wireless Communications", IEEE Wireless Communications , Vol. 12, Issue 4, pp. 14-21, Aug. 2005.

[2] V. K. Goyal, "Multiple Description Coding: Compression Meets the Network," IEEE Signal Processing Magazine, vol. 18, no. 5, Sept. 2001.

[3] Y. Wang, A. R. Reibman, and S. Lin, “Multiple Description Coding for Video Delivery,” Proceeding IEEE, vol. 93, no. 1, Jan. 2005.

[4] V.A. Vaishampayan, “Design of Multiple Description Scalar Quantizers,” IEEE Transaction on Information Theory, vol. 39, 1993.

[5] H. Schwarz, D. Marpe , T. Wiegand, “Analysis of hierarchical B pictures and MTCF,” IEEE International Conference on Multimedia & Expo, ICME ’06, pp. 1929-1932 .

[6] Heiko Schwarz, Detlev Marpe, Member, IEEE, and Thomas Wiegand,” Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” IEEE Trans. Circuits and Systems for Video Technology, vol.17, No.9, Sept. 2007.

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

[8] C. Zhu and M. Liu, “Multiple description video coding based on hierarchical B pictures,” IEEE Trans. Circuits and Systems for Video Technology, vol.19, No.4, April 2009.

[9] J. Apostolopoulos, W. Tan, S.J.Wee, and G.W. Wornell, “Modeling Path Diversity for Multiple Description Video Communication,” IEEE International

Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2002.

[10] O. Campana, R. Contiero, “An H.264/AVC Video Coder Based on Multiple Description Scalar Quantizer,” IEEE Asilomar Conference on Signals, Systems and Computers(ACSSC), 2006.

[11] R. Bemardini, M. Durigon, R. Rinaldo, L. Celetto, and A. Vitali, “Polyphase Spatial Subsampling Multiple Description Coding of Video Streams with H.264,” Proceedings of IEEE International Conference on Image Processing(ICIP), Oct. 2004.

[12] A. Reibman, H. Jafarkhani, Y. Wang, M. Orchard, “Multiple Description Video Using Rate-Distortion Splitting,” Proceedings of IEEE International Conference on Image Processing(ICIP), 2001.

[13] Matty, K.R. and Kondi, L.P., “Balanced multiple description video coding using optimal partitioning of the DCT coefficients,” IEEE Transaction on Circuits and Systems for Video Technology, vol. 15, no. 7, July 2009.

[14] Nicola Conci, Francesco G.B. Natale, “Multiple Description Video Coding Using Coefficients Ordering and Interpolation,” Signal Processing: Image Communication, 2007.

[15] J. Jia and H. K. Kim, “Polyphase Downsampling Based Multiple Description Coding Applied to H.264 Video Coding,” IEICE Transactions, June 2006.

[16] H.264/AVC Ref. Software – JM, http://iphome.hhi.de/ suehring/tml/.

[17] X. Ji, D. Zhao, and W. Gao,” Concealment of Whole-Picture Loss in Hierarchical B-Picture Scalable Video Coding”, IEEE Trans. Circuits and Systems for Video Technology, vol.11, No.1, pp.11-22, Jan. 2009.

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