# Joint Temporal and Spatial Multiple Description Coding for H.264 Video

*Jian-Yu Chen and Wen-Jiin Tsai**

Department of Computer Science, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C.
wjtsai@cs.nctu.edu.tw

## ABSTRACT

Multiple description video coding (MDC) is one of the techniques used to reduce the detrimental effects caused by transmission over error-prone networks. This paper presents a hybrid MDC method which segments the video along spatial and temporal dimensions, and provides efficient estimation methods for missing description reconstruction. The experimental results confirm the improved error resilience capability achieved by the proposed hybrid MDC in lossy networks.

***Index Terms***— Multiple description coding, description estimation, temporal segmentation, spatial segmentation.

## 1. INTRODUCTION

Multiple description coding is a technique that encodes a single information source into two or more output streams, called *descriptions*, and each description can be decoded independently and has an acceptable decoding quality; in addition, the decoding quality will be better if more descriptions were received. The first MD video coder, called multiple description scalar quantizer (MDSQ) [1], has been realized in 1993. Afterwards, researches on various MDC approaches had been proposed. These approaches can be intuitively classified through the stage where it split the signal, such as, spatial, frequency, and temporal domains. To be more precise, Wang [2] had come up with another classification scheme which is based on the type of predictor a MDC approach had adopted: class A focuses on the prediction efficiency and class B on the mismatch control.

Class-A model applies the MDC after motion-compensation, namely, there is only one prediction loop for motion estimation. As the predictor used in the MDC encoder is in accordance with that used in standard encoder, class-A model is characterized by its high prediction efficiency. There are a number of MDC approaches using class-A model. These approaches split video signal either on motion-compensated residue or on frequency coefficients [1,3,4]. Class B model is characterized by the prediction mismatch control, which is achieved by applying MDC before motion compensation and then encoding each description separately. Since the prediction loop in encoder is the same as that in the side decoder of each description, prediction mismatch no longer exists and a better side-decoder performance can be achieved in the case of description loss. However, since the encoder of each description uses incomplete frame information for motion prediction, class-B model has relatively worse coding efficiency, in comparison to class-A model. A variety of MDC approaches of class B have been proposed, from simple to complex architectures [5,6,7].

Most MDC methods are limited to produce two descriptions [8]. This is a heavy constraint for a scalable environment when more than two levels of reconstruction are required or for high bit-rate applications where having a multilayer representation of the source is useful. The polyphase spatial sub-sampling (PSS) in [9] is a class B method worth mentioning because it is designed for four descriptions. The PSS method sub-samples each frame of original sequence by factor 2 row-by-row and then column-by-column, resulting in four sub-frames, each of them has half size of width and height. In PSS, the MDC is applied before motion estimation, so four motion prediction loops are required, one for each description. This paper presents a hybrid MDC method which is also designed for producing four descriptors. The novelty of the hybrid MDC proposed in this paper is that it combines class-A and class-B methods to achieve better coding efficiency and it segments the video in both spatial and temporal domains to have better estimation of lost descriptions.

## 2. HYBRID MDC MODEL

In this section, the encoder of the proposed Hybrid model is presented first, and then is the decoder.

### 2.1. Hybrid Encoder

The Hybrid encoder has a two-level splitting process: 1) *Temporal splitter*, and 2) *Residual Splitter*; the former one which splits the video sequence in temporal domain before motion estimation is a class-B method; while the latter one which splits the motion compensated residual in spatial domain is a class-A method.

978-1-4244-7994-8/10/$26.00 ©2010 IEEE

The temporal splitter segments a sequence along temporal dimension into two subsequences: one for all the even frames and the other for all the odd frames. Even frames are predicted from even ones, and odd frames from odd ones, resulting in two motion-estimation prediction loops. We refer to one of the prediction loops as $T_0$ and the other as $T_1$. After motion estimation and compensation in each loop, the residual splitter is performed on an 8x8-block basis using *polyphase permuting and splitting* in the residual data. Each 8x8 residual block is first polyphase permuted inside the block and then split to 2 blocks as shown in Fig. 1. The middle of Fig. 1 shows the polyphase permuting results, where label-0 pixels are re-arranged to the top-left 4x4 block, label-1 pixels to the top-right 4x4 block, and etc. The purpose of permuting pixels before splitting is to take into account the estimation of lost description, which will be discussed later. After polyphase permuting, the splitting process 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. For each 8x8 block, the remaining two 4x4 blocks with pixels all labeled with ‘x’ in Fig.1 are given residual pixels all set to zero. The encoder has no need to encode the coefficient of these two all-zero blocks. Briefly, the encoding path of Hybrid MDC is split into two after temporal splitter, and then four after residual splitter. The resulting four descriptions are called $T_0R_0$, $T_0R_1$, $T_1R_0$, and $T_1R_1$, respectively.

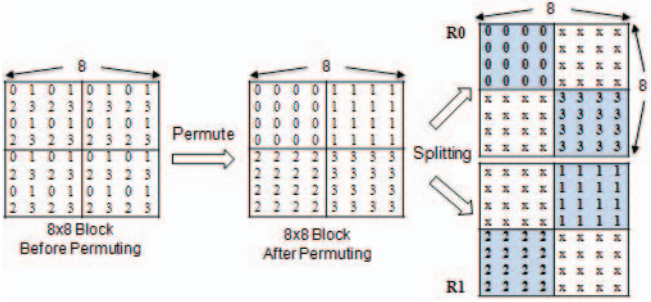


**Fig. 1**. Polyphase permuting and splitting of an 8x8 residual block.

## 2.2. Hybrid Decoder

In Hybrid decoder, if all the four descriptions are received correctly, these descriptions are separately entropy decoded, dequantized, inversely transformed, and then a *Residual Merger* is applied to merge every two descriptions from the same prediction loops. The Residual Merger adopts residual merging and polyphase inverse permuting in a reversed way of Fig.1. Then, motion compensation is applied on each prediction loop and a *Temporal Merger* is used to combine even and old frames so the whole sequence is reconstructed.

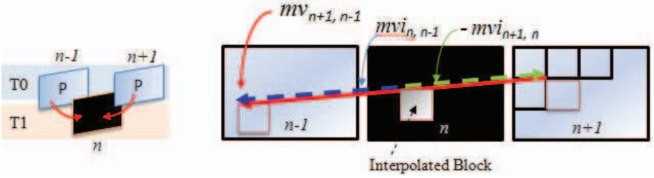
However, if not all the descriptions are received intactly, the decoder may apply spatial estimation after Residual Merger, or temporal estimation after Temporal Merger to reconstruct the lost descriptions. We first describe the temporal and spatial estimation methods in the context of the proposed Hybrid MDC, and then the criterion for estimation method selection is presented.

### *2.2.1. Temporal Estimation Method: B-PMVI*

When two descriptions from the same prediction loop are lost, it will result in whole-frame loss. Since the Hybrid method splits consecutive frames into different prediction loops, the lost frame can be estimated by its previous frame and next frame in the other prediction loop as illustrated in Fig. 2(a), where assume frame $n$ from prediction loop $T_1$, is lost. Since all the MVs in frame $n$ are lost, we propose a Bi-directional pixel-based motion vector interpolation method (B-PMVI) to calculate the motion compensated locations for each lost pixel. Let $mv_{i,\,j}$ denote the motion vector pointing to frame $j$ from frame $i$. In Fig. 2(b), by interpolating the $mv_{n+1,\,n-1}$ (obtained from prediction loop $T_0$), an *interpolated block* on the missing frame $n$, and its two interpolated motion vectors, $mvi_{n,\,n-1}$ and $mvi_{n+1,n}$ , can be obtained (Note the interpolated block is unnecessary to be aligned on MB positions). By inversing $mvi_{n+1,n}$ , we yield two vectors for each pixel of the interpolated block: one is forward vector, $(f_x\,, f_y) = mvi_{n,\,n-1}$, and the other is backward vector, $(b_x,b_y) = -mvi_{n+1,\,n}$ . The interpolation is performed for every MV in frame $n+1$. For each pixel location in the lost frame $n$, if it is covered by at least one interpolated block, its forward and backward vectors are estimated by averaging the motion vectors of all the overlapped blocks; otherwise, its vectors are simply set to zero. As a consequence, for a pixel (x, y) in the lost frame $n$, with its two motion vectors, $(f_x\,, f_y)$ and $(b_x,b_y)$, its value $P_n(x, y)$ can be estimated as follows:

$$P_n(x, y) = w \times P_{n-1}(x+f_x,\ y+f_y) + (1-w) \times P_{n+1}(x+b_x,\ y+b_y) \quad (1)$$

where $w$ is the weighting factor of forward and backward motion compensated pixels. In this paper, $w = 0.5$ is used.



(a) Bidirectional estimation (b) Motion vector interpolation

**Fig. 2**. Whole frame estimation with B-PMVI

### *2.2.2. Spatial Estimation Method*

Spatial estimation method explores the spatial correlation between motion compensated residual pixels to estimate the lost description. It is only adopted for the case of partial frame loss. Since a lost description can obtain its missing motion vectors from its counterpart in the same prediction loop, motion compensation still can be performed. Spatial estimation described here is used to recover the lost residual data. Assuming that $T_0^nR_0$ and $T_0^nR_1$ are two descriptions split from frame $n$ of prediction loop $T_0$, and only $T_0^nR_0$ is received. By polyphase inversely permuting the residual pixels of $T_0^n R_0$, they are distributed like a checkerboard within a macroblock as shown in Fig.3, where black area denotes the lost residual pixels of $T_0^n R_1$. Our spatial method uses *bilinear interpolation* to estimate these lost residual pixels, as shown in Equation (3). Since neighboring pixels have high spatial correlation, this method should be efficient.

$$\widetilde{f_{j,i}} = (f_{j+1,i} + f_{j-1,i} + f_{j,i+1} + f_{j,i-1})/4 \qquad (3)$$

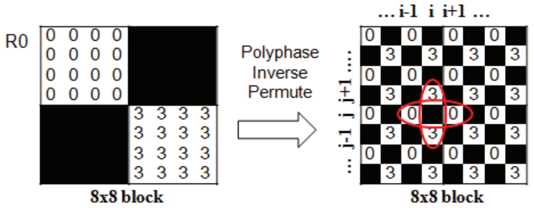
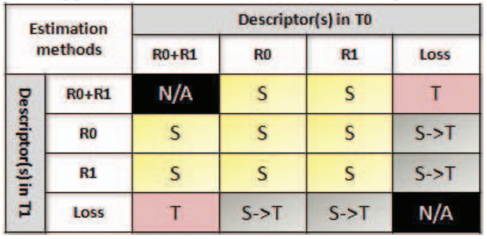


**Fig. 3.** Spatial estimation by bilinear interpolation.

### 2.2.3. Estimation Method Selection

The proposed Hybrid MDC segments a video sequence into four descriptions. There are 16 states of the four descriptions as listed in Table I, where the columns describe the four possible cases for the two descriptions split from prediction loop $T_0$; while the rows describe those for $T_1$. The estimation method to be applied for each case are also shown in this table, where 'T' denotes the temporal estimation, 'S' the spatial estimation, and 'S→T' denotes that spatial method will be performed first and then temporal method.

**Table I**. Mapping of estimation methods and description-loss cases



| Estimation methods | | Descriptor(s) in T0 | | | |
|---|---|---|---|---|---|
| | | R0+R1 | R0 | R1 | Loss |
| Descriptor(s) in T1 | R0+R1 | N/A | S | S | T |
| | R0 | S | S | S | S->T |
| | R1 | S | S | S | S->T |
| | Loss | T | S->T | S->T | N/A |

To illustrate the cases that 'S→T' will be applied, Fig. 4 depicts one of the four possible cases that three descriptions are lost. The descriptions marked with '(x)' mean they are lost. In Fig. 4, since $T_0^n R_0$ of $T_0$ is received, spatial method can be applied to reconstruct its counterpart, $T_0^n R_1$, as denoted by the dotted arrow labeled with S. Then, after merging $T_0^n R_0$ and $T_0^n R_1$, the reconstructed frame $T_0^n$, together with the frame $T_0^{n+2}$, are used by temporal method B-PMVI to recover the lost whole frame $T_1^{n+1}$, as denoted by the dotted arrow labeled with T. The right side of Fig. 4 shows how the 'S→T' is performed.

To illustrate the cases that 'T' will be applied, Fig.5 depicts two cases that two descriptions from the same prediction loop are lost. In each case, the lost description has no counterpart in the same prediction loop available for spatial estimation and hence, temporal method will be applied. As an example in Fig.5(a), after frame *n*+1 from prediction loop $T_1$ is obtained, it together with frame *n*-1 are adopted by temporal estimation to recover the lost frame *n* belonging to prediction loop $T_0$.

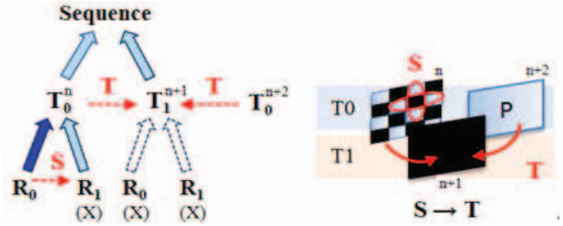
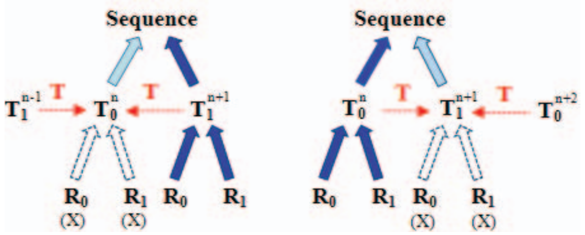


**Fig. 4.** 'S→T' for three missing descriptions



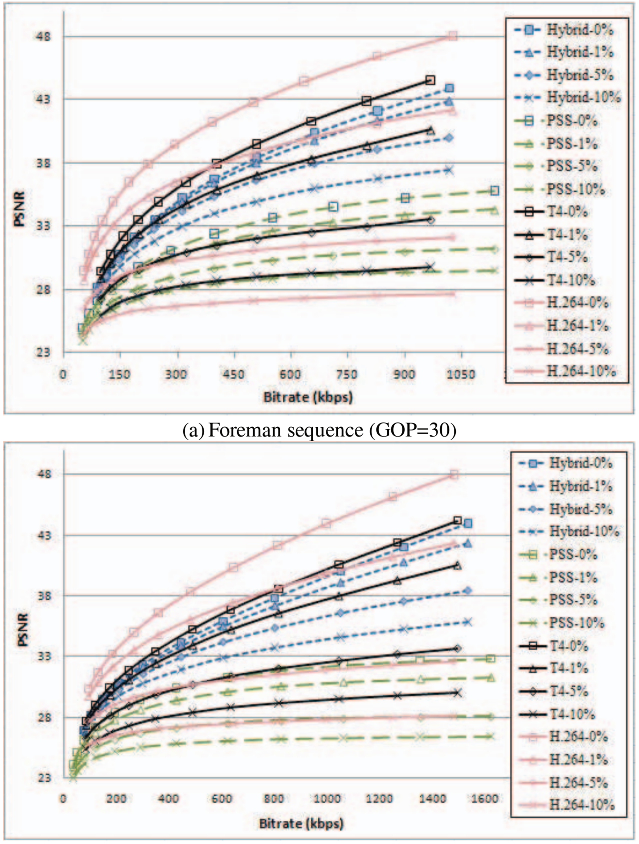
(a) Two descriptions of T1 are lost (b) Two descriptions of T0 are lost.

**Fig. 5**. Temporal estimation for two missing descriptions

## 3. EXPERIMENTAL RESULTS

The performance of the proposed Hybrid MDC is examined in a packet-loss scenario with various packet-loss rates. We compare it with T4, PSS [9] and H.264/AVC, where the T4 splits the sequence into four descriptions along temporal dimension; PSS splits the sequence into four descriptions along spatial dimension; and H.264/AVC is a standard video coder. These methods are implemented by modifying H.264/AVC reference software, JM 13.2 [10]. Two QCIF (176x144) test sequences: *foreman*, and *coastguard* are used, where the GOP size is 30 frames, and the structure of GOP is IPPPP. The results are shown in Fig.6.

From Fig.6, it can be seen that H.264/AVC has a better rate-distortion (R-D) performance than all MDC methods for $P_{loss} < 1\%$, showing that for very low packet-loss rates, the PSNR gain from MDC methods cannot compensate for the loss in coding efficiency. Among the MDC methods, T4 performs best and PSS performs worst for $P_{loss} < 1\%$. This is due to that T4 has the best and PSS has the worst coding efficiency among the three methods. As $P_{loss}$ increases, however, the R-D curve of H.264/AVC drops quickly but the curves of the MDC methods drop gradually, confirming the advantage of the error resilience capability of MDC methods. Among the MDC methods, Hybrid performs better than T4 for $P_{loss} > 1\%$ because T4's performance drops much more quickly than Hybrid's as $P_{loss}$ increases. The results show that the estimation methods adopted in Hybrid are superior to the methods used in T4. On the other hand, due to poor coding efficiency, PSS outperforms H.264/AVC only when $P_{loss}>10\%$ in *foreman* sequence and has the worst R-D performance among the three MDC methods.



(a) Foreman sequence (GOP=30)

(b) Coastsguard sequence (GOP=30)

**Fig. 6**. Performance comparison in packet-loss environment (GOP = 30).

To see the error propagation effect of various MDC methods, we enlarged the GOP size of *coastguard* sequence to 300 and produced one-packet loss at frame 39. The results are depicted in Fig.7, where the PSNR relative to error-free environment are presented. The results in Fig.7 show that the PSNR of T4 drops periodically for every fourth frame. This is due to that T4 uses four prediction loops for every four consecutive frames and that the error happened on one loop will not propagate to the other three. As a result, there is an unbalanced quality among frames for T4. Hybrid uses two prediction loops for every two consecutive frames and hence the quality is degraded for every two frames. But, the PSNR difference between successive frames is quite small and the overall quality is much better than other MDC methods. PSS uses four prediction loops for every frame so each frame is almost equally affected by the packet loss. However, the overall performance of PSS is not good because the average PSNR degradation for the first 100 frames after packet loss is more than 1.5dB. Compared with MDC methods, H.264/AVC has the worst performance.

To sum up, Hybrid method not only has the best average quality, but also has a relatively stable quality among frames when packet loss occurs.

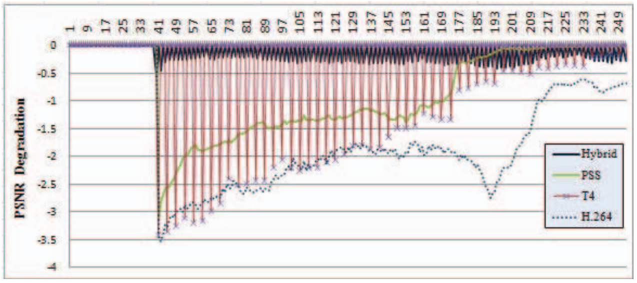


**Fig. 7**. Frame by frame comparison (packet loss at frame 39).

## 6. CONCLUSION

A hybrid MDC method which segments the video in both temporal and spatial domains can provide better estimation of lost descriptions by taking advantages of data correlation in these two domains. The overall experimental results demonstrate the improved lost-description estimation accomplished by the adopted estimation methods and the error resilience capability exhibited by the Hybrid MDC.

**Acknowledgements** This work was supported in part by grants from the National Science Council of Taiwan under Contracts NSC 98-2221-E-009 -085.

## 6. REFERENCES

[1] V.A. Vaishampayan, “Design of Multiple Description Scalar Quantizers,” IEEE Trans. on Info. Theory, vol. 39, 1993.

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

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

[4] A. Reibman, H. Jafarkhani, Y. Wang, M. Orchard, “Multiple Description Video Using Rate-Distortion Splitting,” IEEE Intl. Conf. on Image Processing, 2001.

[5] J. G. Apostolopoulos, “Error-Resilient Video Compression Through the Use of Multiple States,” IEEE Intl. Conf. on Image Processing, 2000.

[6] S. Gao, H. Gharavi, “Multiple Description Video Coding over Multiple Path Routing Networks,” Intl. Conf. on Digital Communication Proceedings (ICDT), 2006.

[7] D. Wang, N. Canagarajah and D. Bull, “Slice Group Based Multiple Description Video Coding Using Motion Vector Estimation,” IEEE Intl. Conf. on Image Processing, 2004.

[8] T. Tillo, M. Grangetto, and G. Olmo, “Redundant Slice Optimal Allocation for H.264 Multiple Description Coding,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 18, no. 1, Jan. 2008.

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

[10] H.264/AVC Software, http://iphome.hhi.de/suehring/tml/