Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Signal Processing: *Image Communication* 19 (2004) 675–684

SIGNAL PROCESSING:
IMAGE
COMMUNICATION

www.elsevier.com/locate/image

Motion information scalability for MC-EZBC☆

Sam S. Tsai, Hsueh-Ming Hang*

Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan, ROC

Abstract

Interframe wavelet video coding algorithm has received much attention lately because of its high coding efficiency and flexible temporal and spatial scalability. It combines the motion-compensated temporal filtering technique together with the wavelet embedded zeroblock coding technique. Blending these two techniques in a nice manner enables it achieving three types of scalability: SNR, temporal and spatial, in one single bit-stream. However, the performance of the interframe wavelet video coding at low bit-rates is less satisfactory because the bit-rate of the non-scalable motion information is too high.

In this paper, we extend the framework of Motion Compensated Embedded Zero Block Coding (MC-EZBC) proposed by RPI (Improved MC-EZBC with Quarter-pixel Motion Vectors, ISO/IEC/JTC1 SC29/WG11 doc. No. m8366, Fairfax, May 2002). Our major contribution is splitting the motion information in MC-EZBC into a few layers. At very low bit-rates only the "coarse" motion vectors are transmitted. Therefore, we are able to produce compressed bit streams at a lower bit-rate and the associated picture quality is significantly better than that of the original scheme. The overhead due to motion information partitioning is negligible at higher rates. Hence, the rate-distortion performance at high rates is about the same as that of the original scheme. In addition, a Hilbert curve scan order is proposed to increase the efficiency up to 5% in encoding the differential motion vectors.

1. Introduction

Digital video coding technology has been rapidly developed in the past 20 years. It enables the new generation of video products and services, ranging from VCD, DVD to Internet video streaming. Due to the varying capability of the video receivers in a network, many methods have been investigated to solve this problem. Lately, the technique of fine-granularity scalability is introduced [1]. Particularly, the so-called interframe wavelet coding technique is able to offer fine-granularity SNR, temporal and spatial scalability at the same time, while it still maintains high compression efficiency.

Interframe wavelet video coding combines the motion-compensated temporal filtering technique together with the wavelet embedded zeroblock coding technique. A rather successful interframe wavelet coding scheme, Motion Compensated Embedded Zero Block Coding (MC-EZBC) is proposed by Woods et al. [2]. This particular

☆This work was partially supported by National Science Council, Taiwan, R.O.C., under Grant NSC-92-2219-E-009-008.

*Corresponding author.

E-mail address: hmhang@mail.nctu.edu.tw (H.-M. Hang).

0923-5965/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2004.05.008

scheme will be introduced in Section 2. Unless specifically noted, the interframe wavelet video coder in this document refers to the MC-EZBC scheme in [2].[1]

The current interframe wavelet coding scheme collects all the motion information together and places it at the beginning of the compressed bit-stream. Because it contains the motion vectors at the highest frame rate and for the best image quality, this portion of data takes a large amount of bits. Hence, the coding performance (image quality) at low bit-rates is relatively poor because the truncated bit-stream contains mostly the motion information only. Furthermore, the motion information imposes a lower achievable bit-rate bound of this scheme. We cannot encode a video sequence at a rate lower than that of the motion information portion.

This paper describes a first attempt that implements the concept of motion information scalability. As an initial implementation, the scheme presented here is rather straightforward. More sophisticated algorithms and analysis are necessary to fully explore this subject. Several possible directions for further study are discussed in Section 6.

Two algorithms are presented in this paper. First, aiming at lowering the motion information bit-rate, the Hilbert scan order is used in producing the differential motion vectors for coding purpose. Second, a motion information partitioning scheme is proposed to split motion information into layers and packets. Only the necessary motion information is included in a low bit-rate bit-stream. Therefore, the coding performance at lower rates is significantly improved.

In Section 2, an overview of the MC-EZBC coding system is given. The Hilbert scan-order prediction method is described in Section 3. The motion information partitioning method and how it is incorporated into the MC-EZBC is described in Section 4. Experimental results are given in Section 5.

2. MC-EZBC

Interframe wavelet video coding is developed based on the concepts of motion-compensated temporal filtering and subband coding. It removes the temporal redundancy in an image sequence by using the motion-compensated (wavelet) filtering technique along the temporal axis [3,4]. Then, the spatial wavelet decomposition is applied to the temporal filtered outputs (frames). By imposing a hierarchical tree structure on the wavelet coefficients and quantizing them using the embedded quantizer [5], we can achieve fine-granularity scalability with a coding efficiency comparable to that of single-layer coding. The architecture of the MC-EZBC interframe wavelet video coder is shown in Fig. 1.

A group of pictures (GOP) are processed together as an integrated unit in the coding process. Each GOP contains 2^n frames, where n equals to the levels of temporal subband decompositions in a GOP. The temporal subband decomposition process is done in two sequential steps. We first construct the motion field (motion vectors) between two consecutive frames using, for example, the hierarchical variable size block matching [6]. Then, we apply temporal filtering to these two frames using the lifted scheme [4] to generate a temporal high-pass frame and a temporal low-pass frame. The decomposition process continues iteratively until the highest temporal level has been reached. A temporal filtering pyramid is thus constructed as shown in Fig. 2.

When the temporal pyramid is generated, the retained frames consist of one temporal low-pass

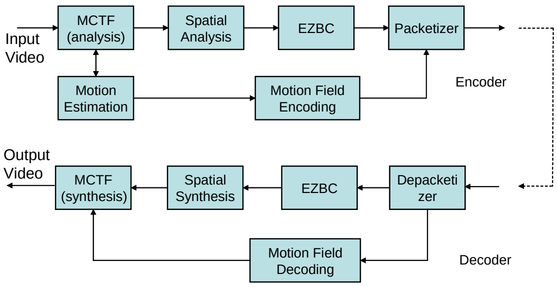


Fig. 1. The interframe wavelet video coder.

[1] We would like to thank Woods and Chen for providing us the source codes of their MC-EZBC algorithm. The work in this contribution is built on the top of these codes.

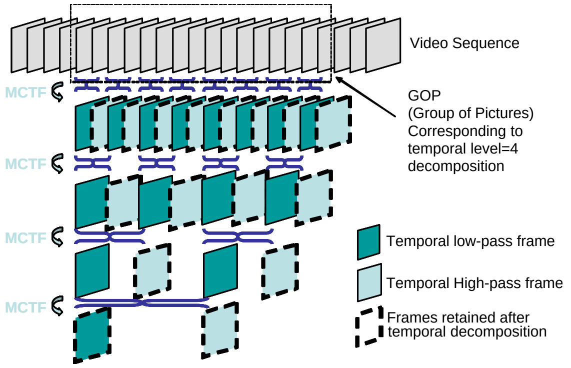


Fig. 2. Temporal filtering pyramid.

frame, and $(1 + 2 + \cdots + 2^{n-1})$ temporal high-pass frames. Each frame is then spatially decomposed into subband coefficients through the use of wavelet transform. The wavelet coefficients are then processed by the embedded zeroblock coding method, and then entropy-coded using arithmetic coding with context modeling [5].

3. Hilbert curve scan-order prediction for MC-EZBC

The motion vector encoding method in MC-EZBC follows the raster scan order with the additional level of recursive raster scan within the quad-tree. This raster scanning order is easy to implement, but it does not provide the best coding efficiency when the motion vector prediction technique is in use. This is mainly because the motion vectors are not always predicted from their nearest neighboring blocks.

In [7], an extensive study was done in finding the optimal scanning order for quad-tree-based motion estimation. The Hilbert curve described in [7] is, therefore, adapted here to improve coding efficiency. It matches the variable block structure of MC-EZBC well. The adapted scheme is described below.

In a quad-tree, a higher-level image block contains four lower-level child blocks. For these four child blocks, there exist four possible types of nearest-neighbor scan orders as shown in Fig. 3. We call the parent block a U-block if the four child blocks follow the scan order in Fig. 3(a). Other block types are similarly defined. Using these four scan types as the basic elements, we construct a nearest-neighbor scanning path for the hierarchically structured motion vectors.

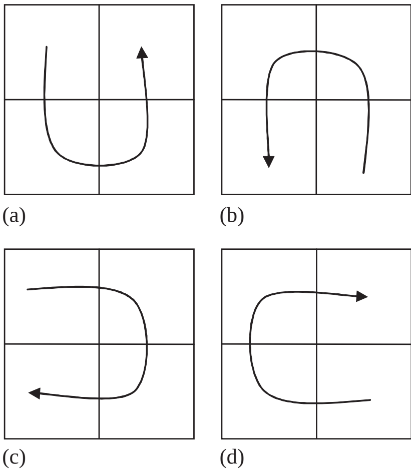
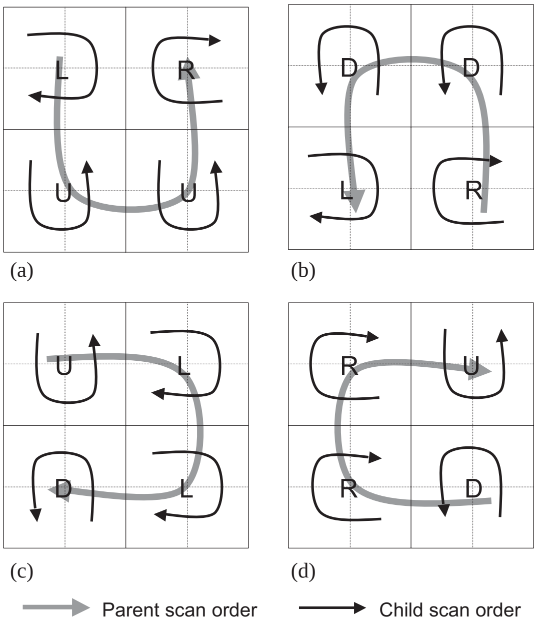


Fig. 3. Four basic types of blocks: (a) U-block; (b) D-block; (c) L-block; and (d) R-block.

When the parent block is chosen to be one specific scan type, in order to preserve the nearest-neighbor property, the scan types of its four children are uniquely determined as showed in Fig. 4 [7]. For example, if the parent block is a U-block, then its four children should have the block type defined in Fig. 4(a). If the four 32 × 32 child blocks are the sub-division of a 64 × 64 U-block, their scanning path is decided by the scan order of the U-block. Similarly, if a 32 × 32 block is replaced by its four (16 × 16) children, the children's scan path is decided by the scan type of its parent (32 × 32) block.

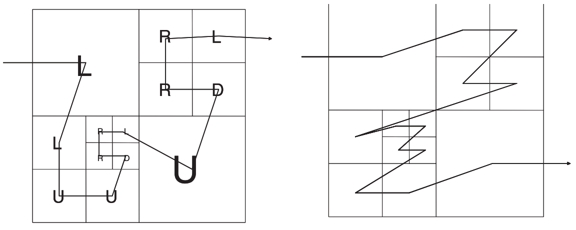
In a 64 × 64 block, the aforementioned scanning path can preserve the neighborhood connection among different level (size) blocks as shown in the example of Fig. 5(a). By setting all the upper level 64 × 64 blocks to U-blocks, the path between 64 × 64 blocks are also connected.

The Hilbert curve scan-order provides a more effective way for motion vector prediction. Hence, the (variable-length coded) bits representing motion information are reduced. Some examples are given in Table 1. The saving in bits could be more significant if a more efficient entropy coding method is in place.



(a) (b) (c) (d)

Fig. 4. Scan path decomposition: (a) U-block; (b) D-block; (c) L-block; and (d) R-block.



(a) (b)

Fig. 5. The Hilbert curve scanning path (a) and the recursive raster-scan scanning path (b) within a 64 × 64 block.

Table 1
Coded bit savings by adopting the Hilbert scan order

	Total bytes using recursive raster scan	Total bytes using Hilbert curve scan	Savings (%)
Foreman	143,258	138,429	3.37
Flower	173,269	164,631	4.99
Mobile	133,675	127,383	4.71
Bus	116,620	114,386	1.92

4. Motion information partitioning

In the motion estimation process of MC-EZBC, a full and complete 5-level motion vector quad-tree is constructed for every 64 × 64 block in an image. The motion vector pruning process then removes unnecessary leaf nodes from the quad-tree, producing a smaller set of motion vectors, which is then sent as the final motion information. This set of motion information contains two parts: (1) the map information, which describes the quad-tree structure after motion vector pruning, and (2) the motion vector information, that is, the motion vectors of each leaf node in the tree.

The preceding two types of motion vector information are both essential to the decoding process. However, the motion information altogether can take up a fairly significant part of the interframe wavelet coded bitstream. For example, in the Harbour sequence, the motion information bit-rate has an average around 400 kbps, Fig. 6(a). When the transmitted bitstream is truncated to a bit-rate close to this rate, the image quality (PSNR) drops substantially as shown in Fig. 6(b). Clearly, the bit-rate cannot go lower. This rate becomes the lower bound of the interframe wavelet coding scheme (for this sequence). Comparing to the other conventional video coding algorithms, this lower bound is too high.

Examining the coded data, we find that a large number of bits are used in representing the leaf nodes with block sizes of 16 × 16 and 8 × 8. Thus, if we divide the motion vectors into two groups: (a) the 16 × 16 and larger block sizes and (b) the block sizes smaller than 16 × 16, we can divide the motion vector information roughly into two equal sets in bits.

Essentially, the motion vector information of each image is grouped into the base layer and the enhancement layer. The *base layer* contains the motion vectors of block sizes of 16 × 16 and above. The *enhancement layer*, on the other hand, contains the refinement motion vectors with sizes below 16 × 16.

In Fig. 7, we can see the difference in the motion vector fields between the base layer and the base plus enhancement layers. When used in interframe prediction, the mean square prediction errors (MSE) of these two cases are displayed in Table 2. Clearly, the addition of enhancement layer (finer spatial resolution motion vectors) can improve prediction accuracy.

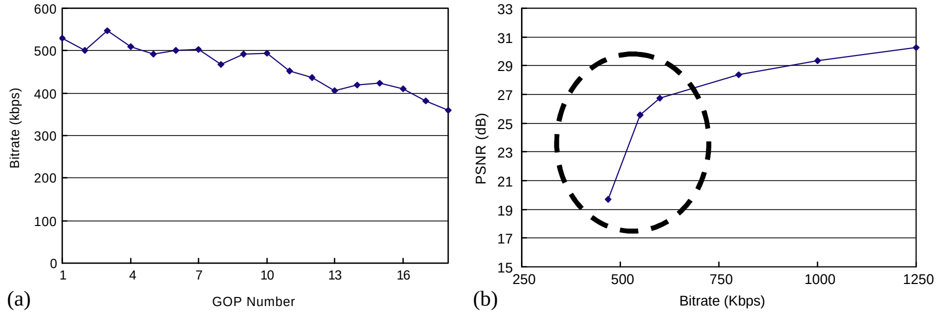


Fig. 6. The motion information bit-rate (a) and the R–D curve (b) of the Harbour sequence.

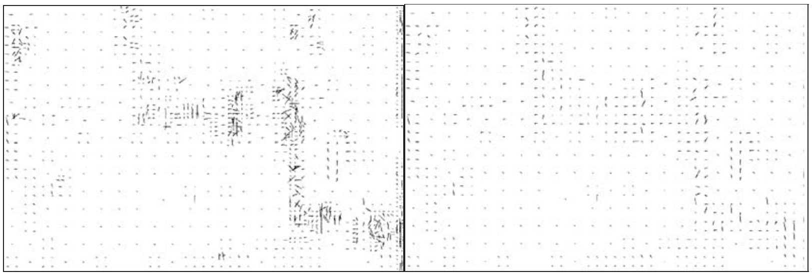


Fig. 7. Motion vector fields of (left) the base plus enhancement layers, and (right) the base layer.

Table 2
MSE values of using the different layers of motion vector

Motion information layers	MSE of prediction
Base	113.8
Base + Enhancement	95.8

Our modified motion information encoding algorithm encodes motion vectors in two passes. The first pass scans, searches and finally produces the (hierarchical) motion vectors of different block sizes following the scan order described in Section 2. The motion vectors of block sizes of 8 × 8 and 4 × 4 are replaced by the 16 × 16 block motion vectors, which were found in the hierarchical motion search process as shown in Fig. 8. Then, the base layer motion vectors (blocks of 16 × 16 and above) are first coded.

The second pass encodes the motion vectors of block sizes of 8 × 8 and 4 × 4 that were replaced by the 16 × 16 blocks during the first pass. Since blocks of these motion vectors are not always connected, the prediction of motion vectors (for coding) is based on the intermediate 16 × 16 motion vector used in the first encoding pass. In this manner, the enhancement layer is coded.

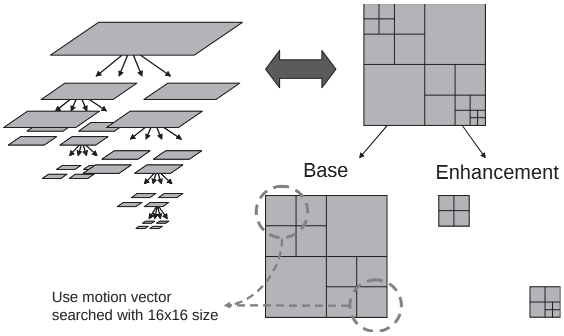


Fig. 8. Base and the enhancement layer motion vectors.

In a GOP packet of the final bit-stream, the motion information bits are organized in three parts: (1) the Map, (2) the base layer motion vectors, and (3) the enhancement layer motion vectors as shown in Fig. 9.

The above idea can be further extended to the GOP level. In each GOP with four levels of temporal decomposition, the first-temporal-level motion-compensated temporal filter performs motion estimation on eight pairs of image frames, generating eight frames of motion vectors. The second-level temporal filter generates four frames of motion vectors. When all four levels of temporal decomposition are completed, there are in total 15 frames of motion vectors.

Each frame of motion vectors in the temporal decomposition is divided into base layer and enhancement layer as described earlier. Therefore, the entire motion vector information is organized in groups as shown in Fig. 10. The base layers of all temporal levels are necessary to reproduce the full-temporal resolution sequence. Therefore, they (base layers) are the highest priority motion vector information data. The enhancement layers at different temporal levels, on the other hand, are transmitted only when the bandwidth is available. Because the temporal motion compensation is performed based on the complete set of motion vectors (base and enhancement layers), the "mismatch" errors appear when the enhancement layers are not completely transmitted to the receiver. We will elaborate on this shortly.

One step further, we split the enhancement layers into 4 sub-groups according to their importance as illustrated in Fig. 11. The higher the temporal level, the greater their importance. This is due to the fact that the higher temporal level frames affect more frames in a GOP.

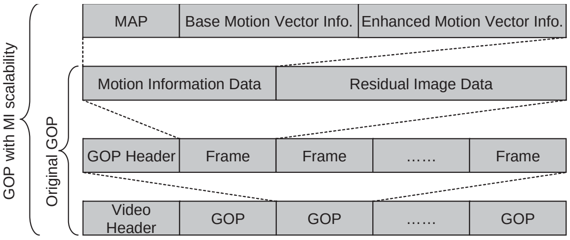


Fig. 9. Partitioned motion information segments in a GOP.

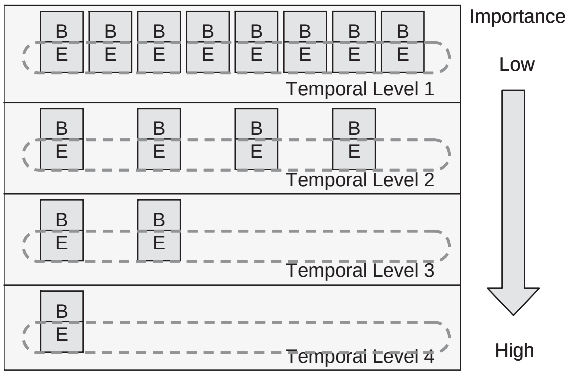


Fig. 11. Importance of enhancement layers (MI: motion information).

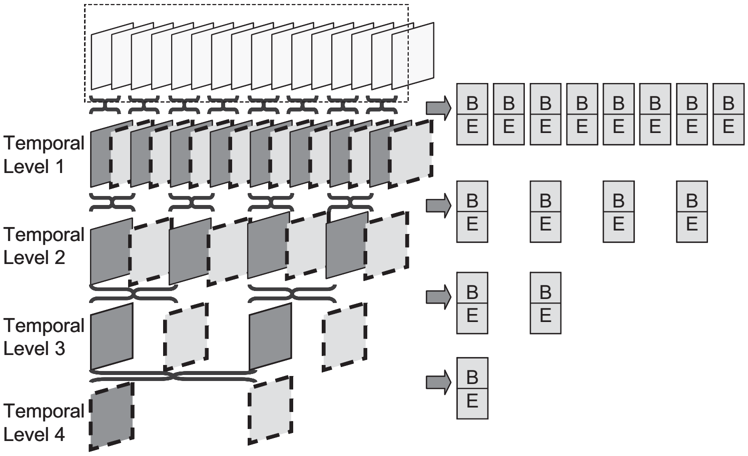


Fig. 10. Motion information of a GOP.

Fig. 12 shows two examples of the amount of bits of various layers of motion information. The notation of "0 Truncated" means all the enhancement layers are included. The "1 Truncated" means the first-temporal-level enhancement layers are dropped. When all enhancement layers are discarded ("4 Truncated"), the remaining bits (base-layer) are about half of the original.

At the end, the motion information partitioned compressed bit-stream contains: (1) motion information with scalability, and (2) residual wavelet-transformed image data with scalability. Thus, there is a compromise in rate-distortion performance between these two types of information. To maintain a decent picture quality and PSNR, a certain amount of the residual image data must be transmitted. On the other hand, at higher bit-rates, to reach the best picture quality, all motion vectors are needed.

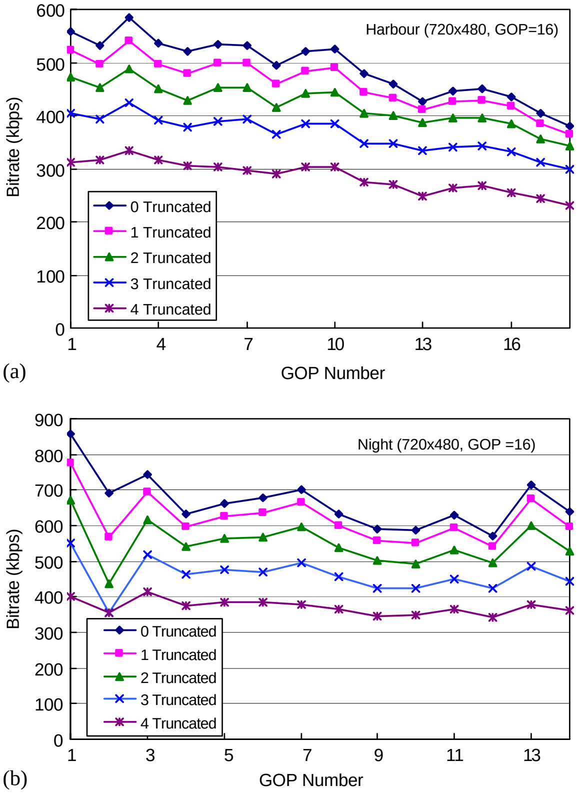


Fig. 12. Motion information (including Map) bit-rate vs. GOP for (a) the Harbour and (b) the Night sequences with different temporal levels of enhancement layer motion information.

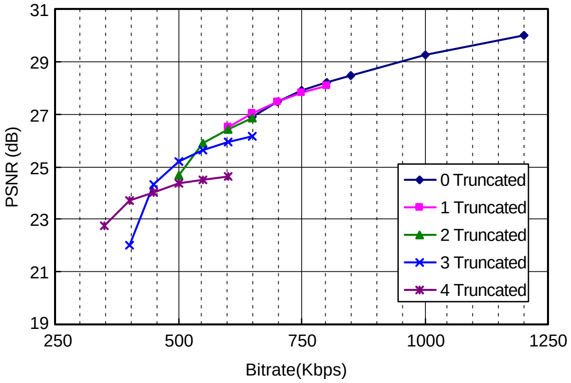


Fig. 13. The R–D performance of different layers of motion enhancement layer truncated for the Harbour sequence.

If not all the motion vectors are used in reconstruction, the mismatch errors would occur. That is, the residual image data calculated at the encoder are based on the complete set of motion vectors but only "partial" motion vectors are available at the decoder if they are truncated for lowering bit-rates. When inexact motion vectors are used in reconstruction, the additional reconstruction errors, so-called "mismatch errors", appear. In the MC-EZBC structure, these mismatch errors only propagate inside a GOP and will not across the GOP boundary.

From Fig. 13, we can see that for high bit-rates, the "4 truncated" sequence suffers from the mismatch errors and thus has the smallest PSNR. On the other hand, at very low rates, the "4 truncated" sequence reserves more bits for sending the residual image data and thus, has the best PSNR. In theory, we could achieve the optimal overall (PSNR) performance at any rate by selecting the best motion information scalable bitstream at each chosen bit-rate. That is, the upper envelope of all the five curves in Fig. 13. All we need to do is to save the crossing (switch) points (in Fig. 13) of the coded sequence for bitstream truncation. Then, in the truncation process, we simply retain the necessary layers of motion vectors based on the crossing point information and the bit budget.

5. Experiments and results

In this section, the coding performance of the interframe wavelet video coding algorithm combined with our multi-layer motion information will be presented. The rate–distortion curves of different sequences are shown. Results of the Hilbert curve scanning method described in Section 3 are also given. In Fig. 14, the motion scalable case is formed by picking up the best performance from the curves of the five truncated bitstreams (e.g., Fig. 13) by choosing the optimal switching points. These points are obtained from Fig. 13 by inspection.

The curves in Figs. 14 and 15 show that the overhead of the proposed algorithm is quite small. For the "Harbour" sequence say, the PSNR drop at high bit-rates is approximately 0.05 dB. The loss for "Night" sequence is a bit higher, approximately 0.12 dB. However, the visual improvement over the original MC-EZBC is significant at low rates as shown in Fig. 16.

After incorporating our algorithm into the MC-EZBC, we find that the compressed bit-rate range now can be extended to lower bit-rates. Also, the combined scheme outperforms the original MC-EZBC substantially at the lower-end of the compressed bit-rate range.

Figs. 17–20 are examples of coded PSNR vs. sequence frame number. Again at high bit-rates (Figs. 17 and 19), because of the additional overhead used in motion information scalability, the new scheme has a slightly lower PSNR. But they are almost identical. They seem to overlap with each other completely when plotted together on the same figure. At lower rates (Figs. 18 and 20), the new scheme is much better.

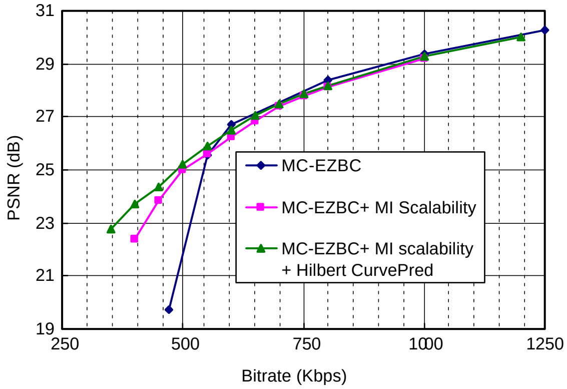


Fig. 14. The R–D curve of the combined algorithm for the Harbour (720 × 480, 30 Hz) sequence.

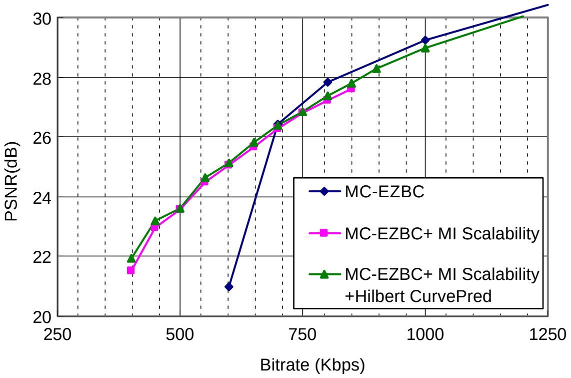


Fig. 15. The R–D curve of the combined algorithm for the Night (720 × 480, 30 Hz) sequence.

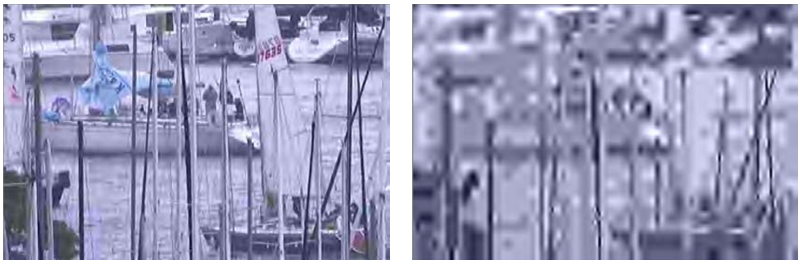


Fig. 16. Frame 15 of the Harbour sequence at 470 kbps (right: original MC-EZBC, left: MC-EZBC with motion information scalability).

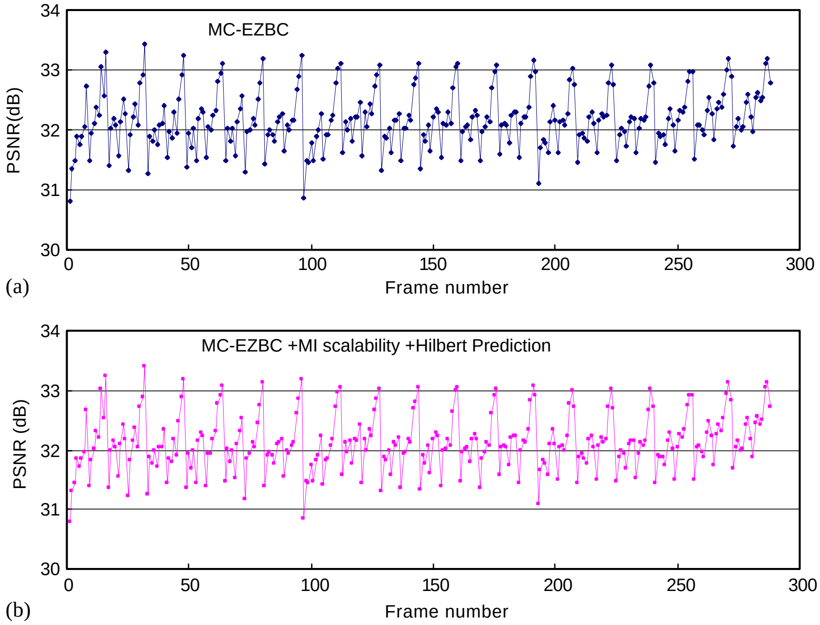


Fig. 17. PSNR vs. frame number of the Harbour sequence at bit-rate 800 kbps using (a) MC-EZBC (b) MC-EZBC combined with motion information scalability and Hilbert curve prediction.

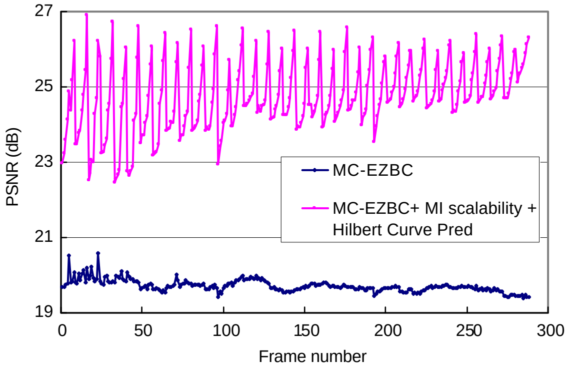


Fig. 18. PSNR vs. frame number of the Harbour sequence at bit-rate 470 kbps.

6. Conclusions

In this paper, two algorithms are proposed to improve the motion information scalability and coding efficiency for interframe wavelet video coding. The Hilbert curve scan order lowers the motion information bit-rate. The motion information partitioning scheme produces a multiple-layer structure of the motion information, which enables the motion information scalability in the inter-frame wavelet video coding. At low bit-rates, only the necessary motion information is included in a transmitted bitstream. Therefore, the coding performance at lower rates is significantly improved.

In this initial attempt to construct a layered structure of motion information, we partition the motion vectors based only on the block sizes. There are several related subjects are under active investigation. For example, the pixel accuracy of motion vector can be another scalable dimension. When an image is down-sampled to 1/4 of its original size, its associated motion vectors are down-scaled by a factor of 2. Therefore, the low-resolution motion vectors may also have lower pixel accuracy. Another topic is the mismatch problem appears in our motion vector truncation process. It causes uneven error propagating along temporal axis as shown in Figs. 18 and 20. The averaged image subjective quality would be much improved if we can compensate the mismatch errors and/or redistribute the distortion. Finally, it would be a rather complex but interesting topic to

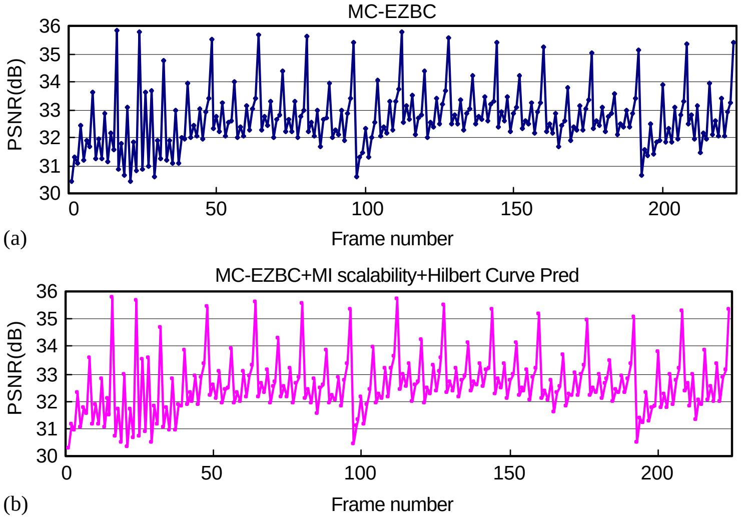


Fig. 19. PSNR vs. frame number of the Night sequence at bit-rate 2000 kbps using (a) MC-EZBC and (b) MC-EZBC combined with motion information scalability and Hilbert curve prediction.

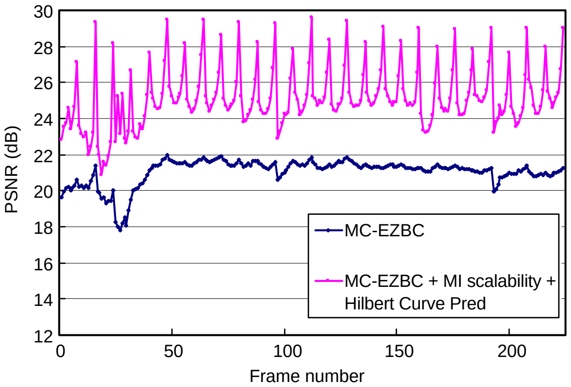


Fig. 20. PSNR vs. frame number of the Night sequence at bit-rate 600 kbps.

look for the optimal R–D representation and partition of motion information. Because the relationship between motion information bits and coding distortion is indirect (depending on residual representation and coding), this topic is very challenging.

References

[1] W.P. Li, Overview of fine granularity scalability in MPEG-4 video standard, IEEE Trans. Circuits Systems Video Technol. 11 (2001) 301–317.

[2] J.W. Woods, P.S. Chen, Improved MC-EZBC with Quarter-pixel Motion Vectors, ISO/IEC/JTC1 SC29/WG11 doc. No. m8366, Fairfax, May 2002.

[3] J.-R. Ohm, Three-dimensional subband coding with motion compensation, IEEE Trans. Image Process. 3 (1994) 559–571.

[4] B. Pesquet-Popescu, V. Bottreau, Three-dimensional lifting schemes for motion compensated video compression, in: Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing, Vol. 3 (2001) pp. 1793–1796.

[5] S.T. Hsiang, J.W. Woods, Embedded image coding using zeroblocks of subband-wavelet coefficients and context modeling, in: Proceedings of IEEE International Symposium on Circuits and Systems, Vol. 3 (2000) pp. 662–665.

[6] S.J. Choi, J.W. Woods, Motion-compensated 3-D subband coding of video, IEEE Trans. Image Process. 8 (1999) 155–167.

[7] G.M. Schuster, A.K. Katsaggelos, An optimal quad-tree based motion estimation and motion-compensated interpolation scheme for video compression, IEEE Trans. Image Process. 7 (1998) 1505–1523.