Invertible Three-Dimensional Analysis/Synthesis System for Video Coding with Half-Pixel-Accurate Motion Compensation

Shih-Ta Hsiang and John W. Woods*

Center for Image Processing Research and
Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute, Troy, NY 12180-3590

ABSTRACT

Three-dimensional subband coding with motion compensation (MC-3DSBC) has been demonstrated [1–4] to be an efficient technique for video coding applications. With half-pixel-accurate motion compensation, images need to be interpolated for motion-compensated (MC) temporal filtering. The resulting analysis/synthesis system is not invertible. In this paper, we propose a new three-dimensional analysis/synthesis system which guarantees perfect reconstruction and has a nonrecursive coding structure. We replaced the analysis/synthesis system of [1] by the new scheme. The resulting coding system does not have distortion from the analysis/synthesis system and allocate bits among classes of 3-D subbands optimally in the sense of rate-distortion function. The experimental results show that the proposed video coding system improves [1] by PSNR .3 - 2.0 dB and TM5 MPEG [10] by PSNR 2.1 - 3.0 dB over a range of bit rates.

Keywords: 3-D subband/wavelet coding, motion compensation, optimized rate allocation, video filtering

1. INTRODUCTION

Owing to its high energy compaction and nonrecursive coding structure, three-dimensional (3-D) subband/wavelet coding with motion compensation (MC-3DSBC) has been demonstrated to outperform the conventional hybrid coders in compression efficiency [1–3] and in robustness for video transmission [4].

It is widely acknowledged that motion compensation with half-pixel accuracy is necessary in order to effectively reduce the energy of the displaced frame difference (DFD). Since the high-frequency output of the temporal Haar analysis filter bank utilized in [1–4] is the scaled difference of the previous and current frames, they adopted half-pixel accuracy for MC temporal filtering in order to reduce the energy of the high-frequency band. The images therein needed to be interpolated at both analysis and synthesis stages and the resulting systems were thus not invertible. Therefore, reconstruction error was introduced even without any coding distortion. This excluded the technique from high-quality video coding applications and also limited the number of analysis/synthesis stages allowed. In [1], two stages of temporal decomposition were applied in order to avoid build-up of reconstruction error from the analysis/synthesis system. For the HDTV application, only one stage could be used in [2]. To further enhance coding efficiency, the images of the lowest temporal band from the same GOP were encoded by temporal DPCM. Therefore, the overall system still could not fully avoid recursive coding structures and their disadvantages.

In this work, we propose an invertible 3-D or spatiotemporal subband/wavelet system with half-pixel-accurate motion compensation for video coding. We term it *invertible motion-compensated 3DSBC* or IMC-3DSBC. We look at temporal decomposition of the progressively scanned image sequence as a kind of down-conversion of the sampling lattice from the interlaced format to the progressive format, following the suggestion in [5]. We thus extended the method of [5], intended for interlaced/progressive scan conversion, to our analysis/synthesis system IMC-3DSBC. An important feature of the new system is that it guarantees perfect reconstruction while high energy compaction is retained.

It is known that optimal bit allocation for conventional hybrid coders is very complex due to the frame-to-frame dependent quantization structure resulting from the DPCM coding loop [6]. On the other hand, in a subband-based coder, coefficients of individual subbands are quantized and coded independently. Optimal bit allocation is therefore

*This work was partially supported by the National Science Foundation under grant MIP-9528312.

possible. However, since MC-DPCM was still used to encode frames of the lowest temporal band in the earlier MC-3DSBC [1], bit allocation could not be fully optimized for the GOPs. In the new system, the input video is decomposed into four temporal stages without build-up of reconstruction error. The GOP consisting of 16 frames does not contain any dependent coding structure at all. Therefore, if the effects of side information are neglected, each GOP can be optimally encoded in an operational rate-distortion sense.

This paper is organized as follows. We present our new temporal analysis/synthesis system in Section 2. Section 3 describes the overall coding system. In Section 4, simulation results are reported, including comparisons with other coding systems.

2. TEMPORAL SUBBAND ANALYSIS/SYNTHESIS SYSTEM

In this section we will demonstrate that, after half-pixel accurate motion compensation, video signals in a progressive format can be treated as a kind of *generalized interlaced video*, to be explained later. Conventional techniques for sampling rate conversion between interlaced and progressive raster can then be utilized to solve the issue of temporal decomposition of video signals.

A simplified model of the projected motion in the image plane of video is to represent the motion field by a global motion vector with the constant velocity (v_1, v_2). Then the intensities of the video can be modeled as

$$\begin{aligned} s_c(x,y,t) &= s_c(x - v_x t, y - v_y t, 0) \\ &= s_{c0}(x - v_x t, y - v_y t) \end{aligned} \tag{1}$$

and the sampled video is represented by

$$S[m,n,k] = s_c(VN) \tag{2}$$

where $N = [m,n,k]^T$ and V is the 3 x 3 sampling matrix. If all frames of the sampled video are aligned with respect to the reference frame, pixels in sub-lattice positions of the reference frame may be filled in with samples from the same lattice positions of other frames after *motion compensation*. Reconstruction of a higher-resolution image from multiple frames is possible, depending on the sampling lattice and the velocity of the global motion [8].

For example, consider the video signal in an interlaced format sampled from the video signal model (1) with the constant-velocity global motion vector $\mathbf{v}$ = (0,0), as depicted in Fig. 1(a) in two dimensions (temporal and vertical). Since the spatial sampling lattices of two adjacent odd and even fields are interlaced, as seen in Fig. 1(b), the composite frames can be built by just merging two adjacent fields, as follows:

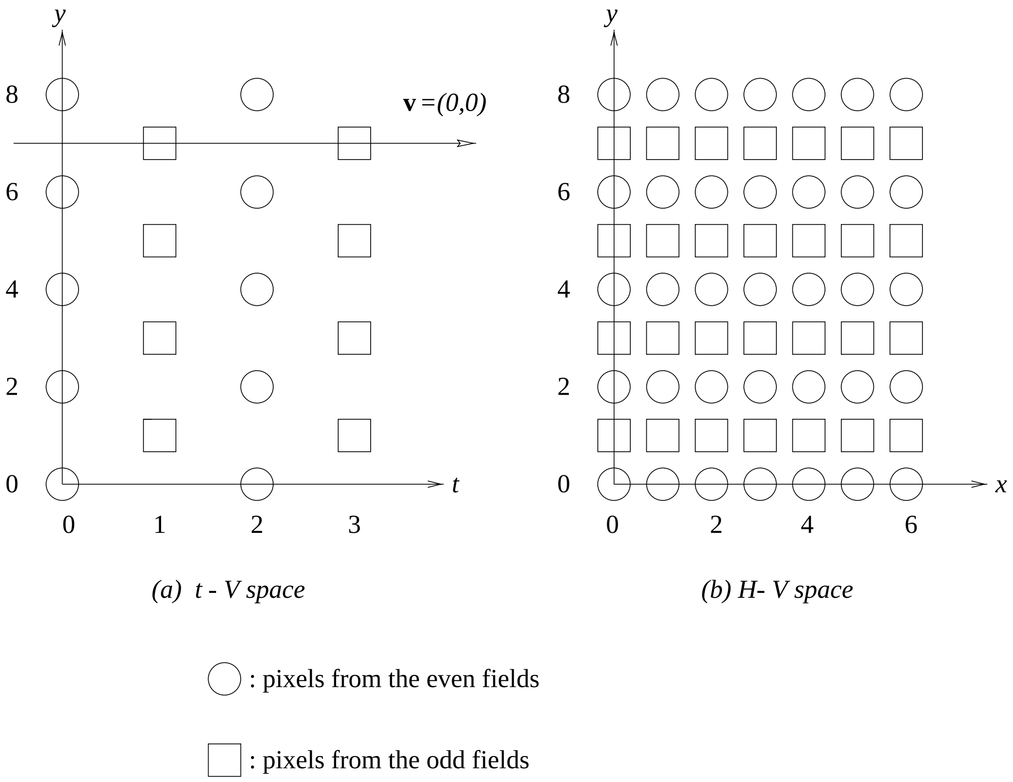
$$C[m,n] = \begin{cases} A[m,n] & n \text{ even} \\ B[m,n] & n \text{ odd} \end{cases} \tag{3}$$

where A and B respectively represent even and odd fields, and C denotes the composite frame.

Next, we look at the video signal with a general integer-valued constant velocity of global motion in a progressive format. As demonstrated in Fig. 2(a), tracking backward along the motion trajectory, each pixel in the video passes through an existing sample in the reference frame. Therefore, the subsequent frames cannot provide any new information. If we also construct the composite frame by merging a pair of two consecutive frames *after motion compensation*, the resolution of the resulting frame is unchanged since the sampling lattices of the two frames are overlapped *after motion compensation*, as seen in Fig. 2(b). This velocity corresponds to the so-called critical velocity described in [7]. Because the following frames do not contain new information, reconstruction of a higher-resolution image from the video is impossible.

When projected motion in the video signal in a progressive format is represented by the global motion model with *half-pixel-accurate* constant velocity, four patterns exist for overlapping sampling lattices of two consecutive frames after motion compensation, as depicted in Fig. 3, respectively corresponding to:

(a) class EO: $2d_m$ even, $2d_n$ odd
(b) class OE: $2d_m$ odd, $2d_n$ even
(c) class OO: $2d_m$ odd, $2d_n$ odd
(d) class EE: $2d_m$ even, $2d_n$ even


y
8
6
4
2
0
0
1
2
3
t
v =(0,0)
y
8
6
4
2
0
0
2
4
6
x
○ : pixels from the even fields
□ : pixels from the odd fields

(a) t - V space

(b) H- V space

Figure 1. (a) Sampling lattice of the video signal in the interlaced format, only vertical and temporal dimensions shown. (b) Spatial sampling lattices of two adjacent fields.

where $(d_m, d_n) = (v_x \triangle t, v_y \triangle t)$ is the displacement vector between the previous and current frames, and $\triangle t$ is the temporal sampling period. Comparing Fig. 3(a) to Fig. 1(b), it is found that the sampling lattices in Fig. 3(a) are the same as the interlaced lattices in Fig. 1(b) scaled by .5 along the vertical direction. Hence, the sampling lattices of two adjacent frames are also interlaced after motion compensation. As an extension of (3), we can construct the composite frame C with resolution doubled along the vertical direction by

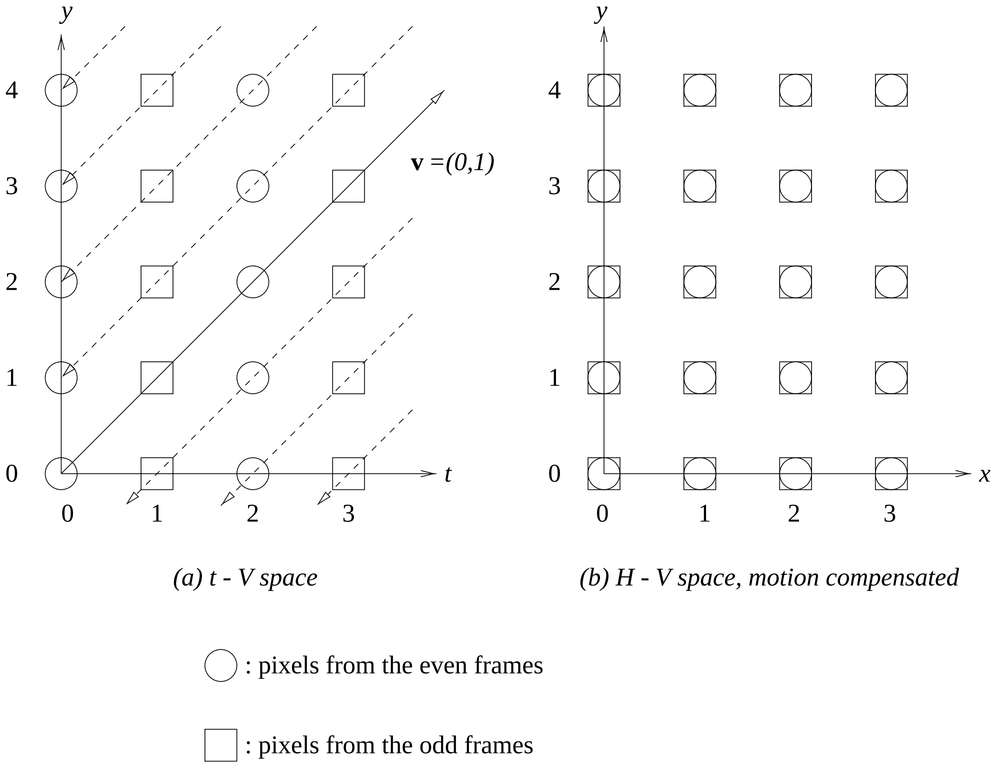
$$C[m,n] = \begin{cases} A[m, n/2] & n \text{ even} \\ B[m + d_m, n/2 + d_n] & n \text{ odd} \end{cases} \tag{4}$$

where, following the notation of [1], A and B respectively denote the previous and current frames. That is, missing pixels at the sub-lattice positions can be filled in by applying the temporal zero-order hold interpolation filter along the motion trajectory. To reduce to original resolution, we followed a suggestion of [5] and extended his algorithm, originally designed for interlace-to-progressive down conversion. The composite frame is decomposed by the two-channel subband analysis filter bank, Daubechies 9/7 filters [9], along the vertical direction. The low and high frequency bands of the analysis output are generated by

$$L_t[m,n] = \sum_k C[m, 2n-k]\; h_0[k]\,, \text{ and} \tag{5}$$

$$H_t[m,n] = \sum_k C[m - d_m, 2(n-d_n) - k]\; h_1[k], \tag{6}$$

where h_0 and h_1 are the lowpass and highpass filters, respectively. The composite frame can be perfectly reconstructed from L_t and H_t utilizing the synthesis filter bank. The reverse operation to restore frames A and B from the composite


y
4
3
2
1
0
0
1
2
3
t
v =(0,1)
y
4
3
2
1
0
0
1
2
3
x
○ : pixels from the even frames
□ : pixels from the odd frames

(a) t - V space

(b) H - V space, motion compensated

Figure 2. Demonstration of the critical velocity in the video signal in the progressive format. (a) Sampling lattice of the video with a constant-velocity global motion vector $\mathbf{v}$ = (0,1), only vertical and temporal dimensions shown. (b) Spatial sampling lattices of two consecutive frames after motion compensation.

frame C is straightforward, i.e.,

$$\begin{aligned} A[m,n] &= C[m,2n] \\ B[m,n] &= C[m-d_m, 2(n-d_n)] \end{aligned} \tag{7}$$

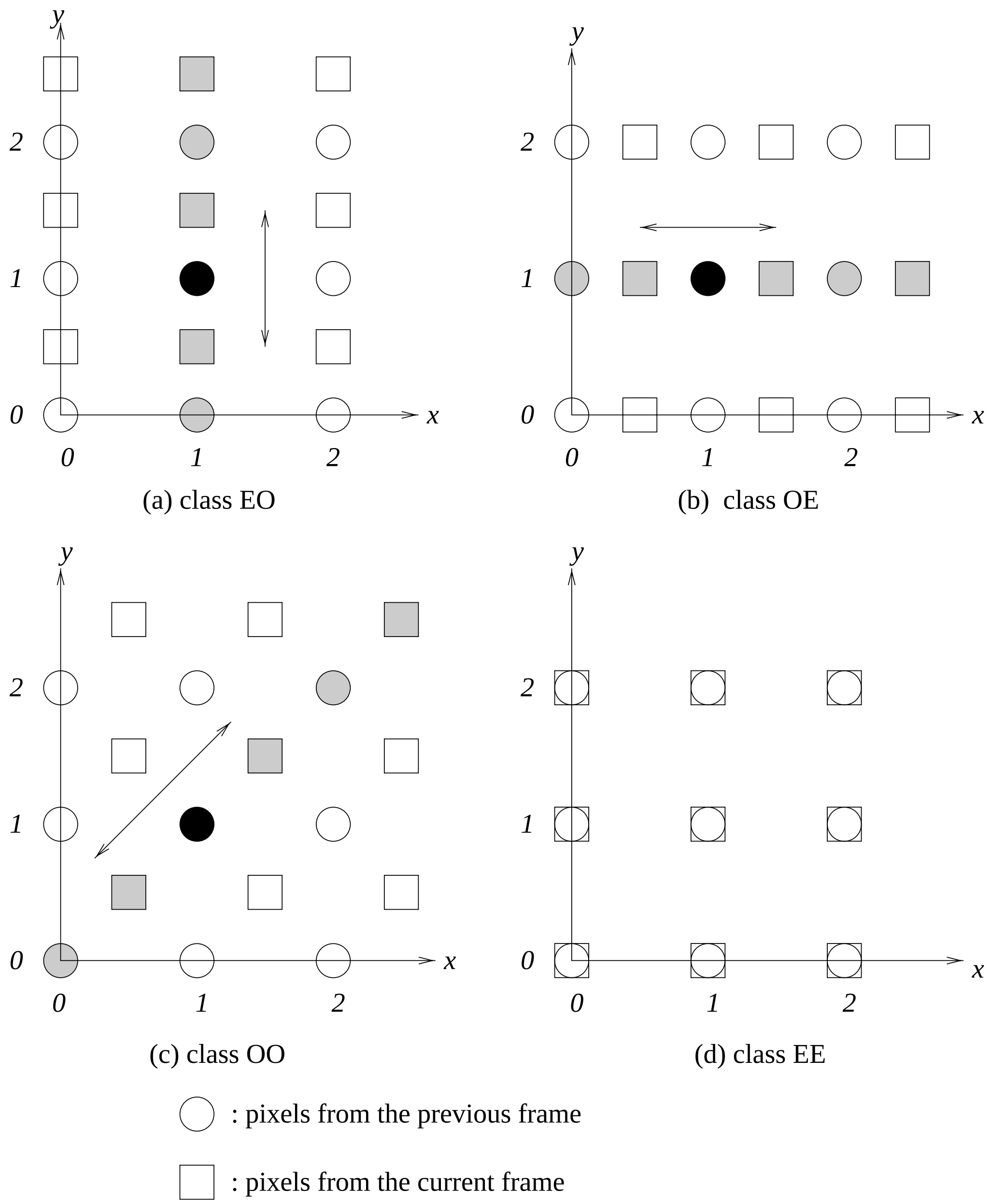
Video signals with global motion vectors corresponding to classes OE and OO can be considered as *generalized interlaced video after motion compensation*. Here, similar to Fig. 3(a), sampling lattices of the odd frames and even frames are interlaced after motion compensation, but sublattice positions are filled in the horizontal and diagonal directions, respectively. Hence, the composite frames can also be built by merging a pair of adjacent frames *after motion compensation*. Equations (4)-(7) can be applied to analyze and synthesize the video signal but along horizontal and diagonal directions, respectively. In Fig. 3(d), the lattice corresponds to the video model explained earlier in Fig. 2. Hence, composite frames with higher resolution in the spatial domain cannot be built for this model. For this case, L_t and H_t respectively represent scaled motion-compensated sum and difference of frames A and B:

$$\begin{aligned} L_t[m,n] &= (B[m+d_m, n+d_n] + A[m,n])/\sqrt{2}\,, \\ H_t[m,n] &= (B[m,n] - A[m-d_m, n-d_n])/\sqrt{2}\,. \end{aligned} \tag{8}$$

Frames A and B can be reconstructed by

$$\begin{aligned} A[m,n] &= (L_t[m,n] - H_t[m+d_m, n+d_n])/\sqrt{2}\,, \\ B[m,n] &= (L_t[m-d_m, n-d_n] + H_t[m,n])/\sqrt{2}\,. \end{aligned} \tag{9}$$

It is noted that (8) and (9) are just the temporal Haar analysis/synthesis pair adopted in [1–4].



y
x
0
1
2
○ : pixels from the previous frame
□ : pixels from the current frame

(a) class EO

(b) class OE

(c) class OO

(d) class EE

Figure 3. The spatial lattices of two consecutive frames after motion compensation. The black circle is the pixel being processed. The gray pixels and arrows indicate the direction of filtering, (a) class EO, (b) class OE, (c) class OO, (D) class EE.

Strictly speaking, only the video model with global motion vector corresponding to Fig. 3(d) utilizes temporal filtering. We use the notation L_t because it represents video at the lower frame rate in our system. The notation H_t follows similarly.

Although the assumption of a global constant-velocity motion model is rarely valid for video captured from real scenes, it can be approximated reasonably well locally as has been demonstrated in many video coding applications. In our algorithm, hierarchical variable size block matching (HVSBM) [1] is utilized for motion estimation. The resulting motion vectors are constant for all pixels from the same motion block. Therefore, if the motion vector for the current motion block is among classes EO, OE and OO, *after motion compensation* a composite block can be constructed by merging a pair of *linked* motion blocks from the previous and current frames, respectively. Then, for *connected pixels* [1], the analysis/synthesis scheme demonstrated by (4)-(7) can be performed block by block along the spatial direction decided by the class of the motion vector, as shown in Fig.3. Block boundaries are symmetrically extended for subband filtering. The low and high frequency bands of the subband output are respectively stored in the lattice positions corresponding to input motion blocks of the previous and current frames. In Fig. 4, we illustrate this temporal decomposition process for a pair of linked motion blocks. The motion vector of class EO with a motion block size 3 × 3 is used for demonstration. For motion vectors corresponding to class EE, equations (8) and (9) are applied to decompose and reconstruct a pair of motion blocks. The new analysis/synthesis system reduces to the original system of [1] for this class.

Regarding unconnected pixels [1] generated by HVSBM, a similar method is adopted, as follows:

For analysis:

$$L_t[m,n] = \sqrt{2}A[m,n], \text{ and} \tag{10}$$

$$H_t[m,n] = (B[m,n] - A[m-\overline{d_m}, n-\overline{d_n}])/\sqrt{2}\,. \tag{11}$$

For synthesis:

$$A[m,n] = L_t[m,n]/\sqrt{2}, \text{ and} \tag{12}$$

$$B[m,n] = A[m-\overline{d_m}, n-\overline{d_n}] + \sqrt{2}H_t[m,n], \tag{13}$$

where $(\overline{d_m}, \overline{d_n})$ is the integral part of the motion vector.

It is interesting to observe that for video compression, it is favorable to have projected motion with critical velocities in an image sequence, as opposed to applications of superresolution from video, where reconstruction of higher-resolution is impossible if a critical velocity occurs. When the video model (1) holds globally, with a critical velocity, the sampling lattice of the new frames overlaps with that of the reference frame after motion compensation. No new information needs to be encoded. However, due to local motion in the image sequence, sublattice positions in the previous frames are filled in by the samples from the current frame in a way that varies from region to region, depending on the class of local motion vector as shown in Fig. 3. Therefore, our algorithm is made adaptive to local motion vectors in processing the different patterns of merged lattices on a block-by-block basis.

3. CODING SYSTEM

A coding structure similar to [1] is used in our coding system, shown in Fig. 5. The input video is temporally decomposed by our new two-channel analysis system, described in the previous section. Four-stage analysis is performed to generate an octave based five-band decomposition. Three-stage spatial analysis follows this temporal stage to complete the 3-D subband decomposition. HVSBM, which helps reduce the number of unconnected pixels [1], is utilized for motion estimation. The sizes of our square motion blocks range from 4 × 4 to 64 × 64.

Our coding system divides consecutive frames into groups, similar to the GOP (group of picture) structure in MPEG. Each GOP contains 16 frames — 1 t-LLLL frames, 1 t-LLLH frame, 2 t-LLH frames, 4 t-LH frames, and 8 t-H frames. By decoding different numbers of temporal lower bands, 5 frame rates can be offered at the receiver. Uniform threshold quantizers (UTQ), based on the Laplacian model with a central dead zone, are used in 3DSB-FSSQ [1] for

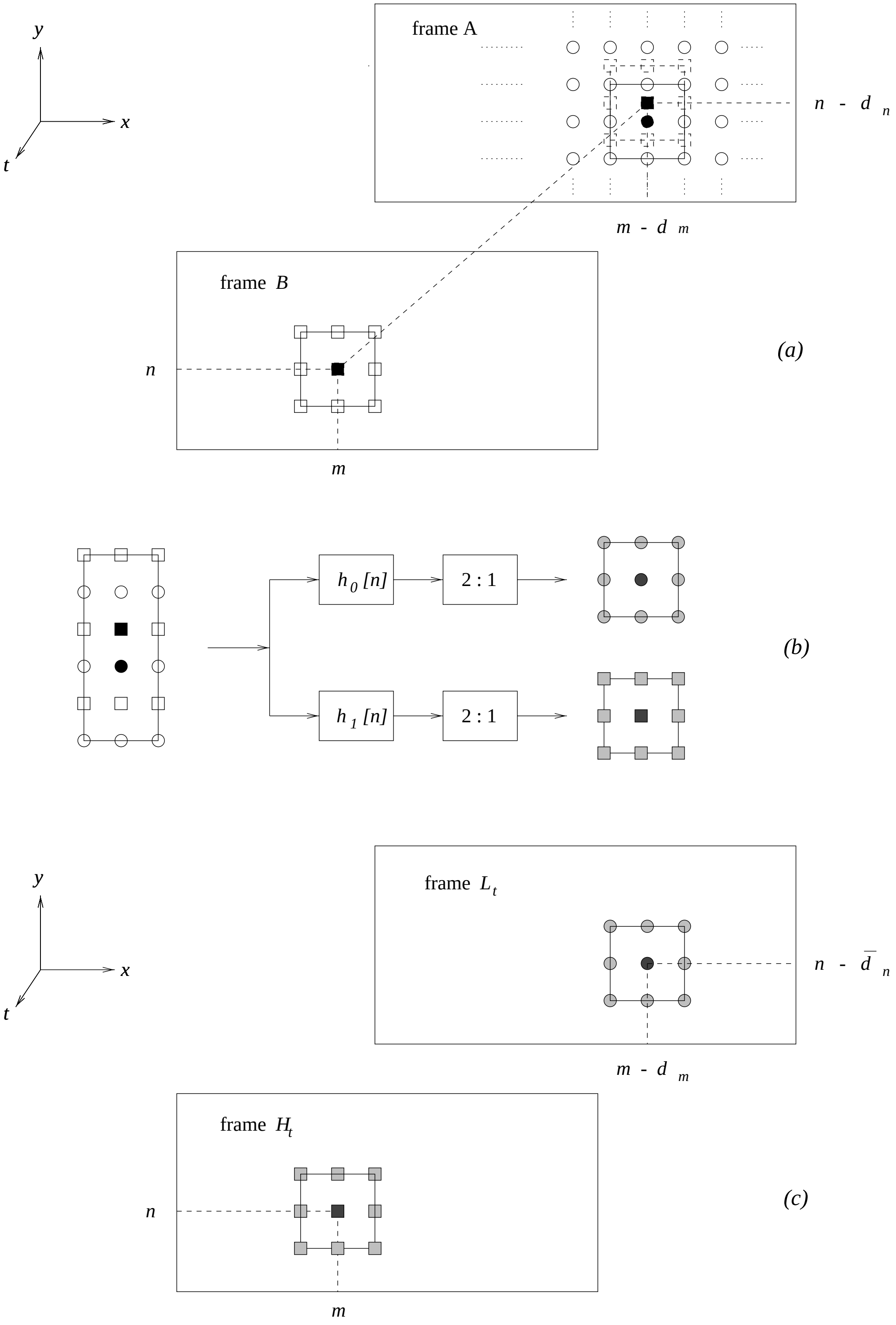


Figure 4. Decomposition of a pair of linked motion blocks, the motion vector corresponding to class EO with the block size 3 x 3. (a) A pair of linked blocks. (b) Decomposition of the composite block. (c) The resulting frames L_t and H_t.

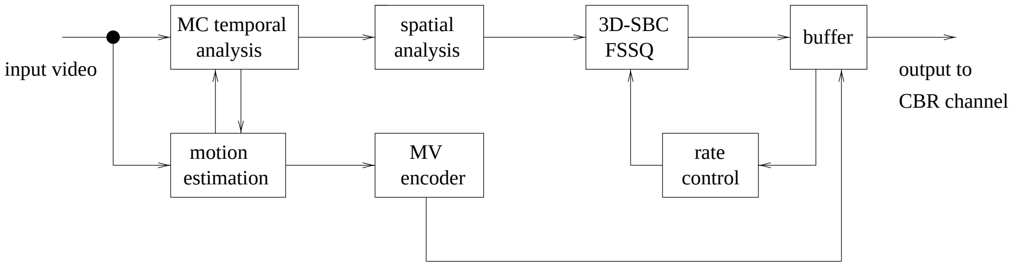


Figure 5. The block diagram of the coding system.

	IMC-3DSBC	MC-3DSBC
recursive structure	No	Yes
number of frames in GOP	16	16
number of temporal analysis stages	4	2
frame rates available for temporal scalability	5	3
frame memory	16 frs	4 frs
bit rate optimized for GOP	Yes	No

Table 1. Comparison between IMC-3DSBC and MC-3DSBC.

encoding the 3-D subbands. The bit budget for each GOP is given by

$$R_g = N_g \ \times \ r \ / \ F \ (bits) \tag{14}$$

where

N_g: the number of frames in a GOP,
r: the bit rate of coding at bits/sec,
F: the frame rate of the image sequence at frames/sec.

Bit allocation is optimized among classes of 3D subbands by the generalized BFOS algorithm as described in [1]. Table 1 gives a comparison between IMC-3DSBC and the earlier MC-3DSBC [1].

4. EXPERIMENTAL RESULTS

The new technique has been implemented in software (IMC-3DSBC). The standard test video clips *Mobile Calendar* and *Flower Garden* in SIF resolution (progressively scanned, 352 x 240, 30 fps) were used in our experiments. Performances of IMC-3DSBC, MC-3DSBC [1], and a standard MPEG [10] were compared.

In Fig. 6, we display the average PSNR of the luminance component of Mobile Calendar encoded by the three coders at various rates. It is seen that our new technique outperforms TM5 MPEG, by 2.1 - 3.0 dB and the earlier MC-3DSBC by 0.3 - 2.0 dB.

Instead of going through a closed error-feed-back loop, IMC-3DSBC uses original images to compute the DFD. Moreover, it provides high energy compaction. Both features are considered advantageous for low bit rate coding. This property is demonstrated in Fig. 6 which shows that relatively larger improvements over MPEG are achieved by IMC-3DSBC at lower rates. Because the earlier MC-3DSBC utilized temporal DPCM to encode frames of the lowest temporal band and applied only two stages of temporal decomposition, it was outperformed by IMC-3DSBC based on the same reasons.

Since the analysis/synthesis system of MC-3DSBC is not invertible, with an increase in the coding rate, the reconstruction error from the analysis/synthesis system will finally dominate the coding distortion. This phenomenon is seen in Fig. 6. However, IMC-3DSBC does not suffer this effect and still outperforms MPEG by more than 2 dB at the higher rates.

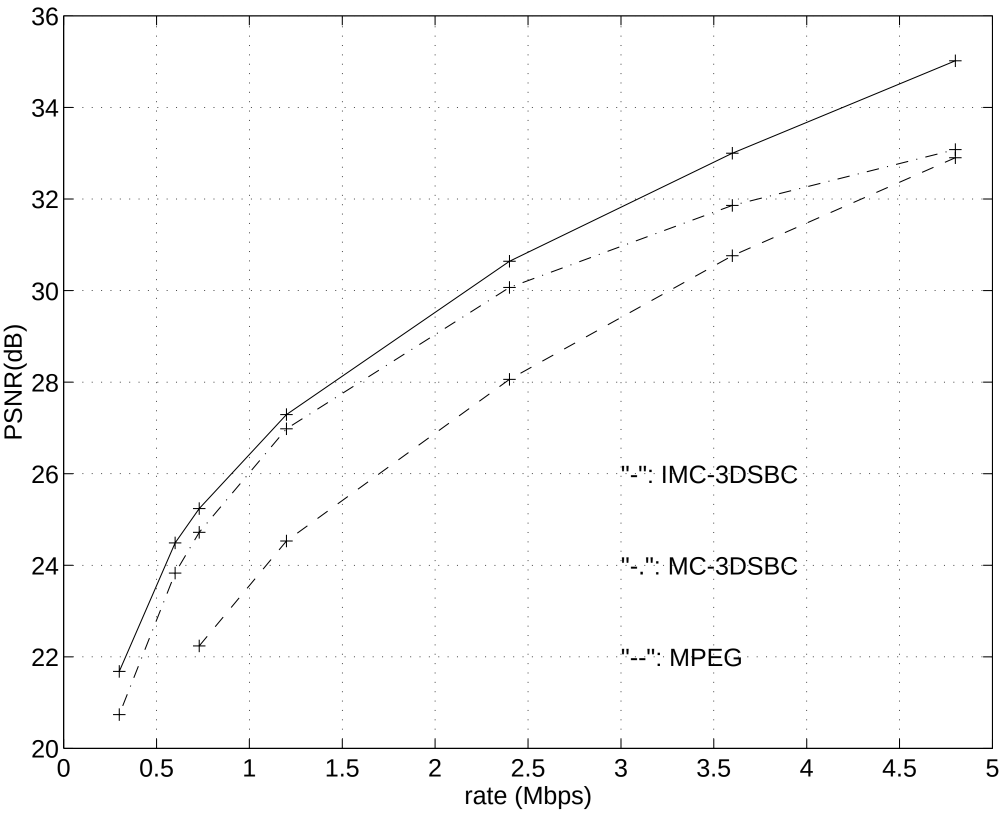


Figure 6. Average PSNRs for the luminance components of Mobile Calendar at various rates.

video sequence	coder	Y (dB)	U (dB)	V (dB)
Mobile Calendar	MPEG	28.06	31.52	31.79
	MC-3DSBC	30.07	35.01	34.79
	IMC-3DSBC	30.64	34.97	34.62
Flower Garden	MPEG	30.32	33.95	32.54
	MC-3DSBC	30.90	35.30	36.17
	IMC-3DSBC	30.80	34.80	35.78

Table 2. Average PNSRs in coding SIF Mobile Calendar and Flower Garden at 2.4 Mbps.

Table 2 summarizes the average PSNRs for Mobile Calendar and Flower Garden encoded at 2.4 Mbps. It is found that IMC-3DSBC does not perform as well for the Flower Garden sequence, while still beating TM5 MPEG. We think this performance drop is related to the larger number of unconnected pixels generated by a four-stage temporal decomposition for a fast camera pan. In our simulations, more than 20% of the pixels in the lowest temporal band are unconnected. Therefore, it turns out that the assumption of additive superposition of coding error is not valid for Flower Garden. This effect, which is not new, can be compensated by scaling the quantizer step sizes appropriately [3]. Work continues on this problem.

5. CONCLUSIONS

In this paper, an invertible 3-D analysis/synthesis system with half-pixel-accurate motion compensation was presented. This new video coding system possesses a nonrecursive structure. Therefore, it is robust for data transmission and bit rates can be more easily optimized in an operational rate-distortion sense. Our experimental results show improved performance over TM5 MPEG and the earlier MC-3DSBC at several coding rates.

REFERENCES

1. S.-J Choi and J. W. Woods, “Motion-Compensated 3-D subband coding of video,” *IEEE Trans. on Image Processing* (, submitted May 1996, accepted March 1998).
2. G. Lilienfield and J. W. Woods, “Scalable High Definition Video Coding,” *Proc. VCIP ’98*, SPIE vol 3309, pp. 158–169, San Jose, CA, Jan. 1998
3. J.-R. Ohm, “Three-dimensional subband coding with motion compensation,” *IEEE Trans. on Image Processing*, vol. 3, pp. 559–571, Sept. 1994.
4. J.-R. Ohm,“Advanced packet-video coding based on layered VQ and SBC technique,” *IEEE Trans. on Circuits and systems for video technology*, vol. 3, pp 208–221, Jun. 1993.
5. J.-R. Ohm, “Variable-raster multiresolution video processing with motion compensation techniques,”*Proc. IEEE ICIP 97*, Santa Barbara, CA, Oct. 1997.
6. K. Ramchandran, A. Ortega, am M. Vetterli, “Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders,” *IEEE Trans. on Image Processing*, vol. 3, pp. 533–545, Sept. 1994.
7. R. A. F. Belfor, R. L. Lagendijk, and J. Biemond, “Subsampling of digital image sequences using motion information,” in *Motion Analysis and Image Sequence Proc.*, M. I. Sezan and R. L. Lagendijk, eds., Kluwer, 1993.
8. M. Tekalp, *Digital Video Processing*, Prentice-Hall, NJ, 1995.
9. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, “Image coding using wavelet transform,” *IEEE Trans. on Image Processing*, vol. 1, pp. 205–220, Apr. 1992.
10. MPEG software simulation group, MPEG-2 encoder v.1.1, TM5, (http://www.mpeg.org/index.html /MSSG/#source).