

# AN EFFICIENT AND LOW-DELAY MCTF PARTITIONING

Liyang Xu<sup>1</sup> and Sunil Kumar<sup>2</sup>, Senior Member, IEEE

<sup>1</sup>Omneon Video Networks, Beaverton, OR 97006, USA

<sup>2</sup>Electrical and Computer Engineering Department, San Diego State University, San Diego, CA 92182, USA

Email: lxu@omneon.com, skumar@mail.sdsu.edu

## ABSTRACT

In this paper, we compute the residual energy of predictive frames (i.e., unidirectional and bidirectional motion prediction) by using the autocorrelation between successive video frames at different frame lags (or prediction terms). A new low-delay motion compensated temporal filtering (MCTF) structure is proposed based on the residual energy computations, which outperforms the ‘low-delay MCTF’ scheme suggested in the Joint Scalable Video Model (JSVM) in terms of delay, number of badly matched blocks and PSNR performance. The PSNR performance of the proposed scheme is only slightly lower than the ‘conventional 5/3 MCTF’ scheme. This scheme can also be used with other types of wavelet filters (e.g., Haar) and can substitute the MCTF in 3D wavelet-based video codecs to achieve low delay with only minor reduction in coding efficiency.

**Keywords:** MCTF, Video coding, JSVM, Scalable.

## 1. INTRODUCTION

Scalable video coding (SVC) schemes encode the video once at higher resolution, but enable decoding from partial bitstream at different reconstructed picture qualities, such as signal-to-noise ratio (i.e., SNR scalability), picture size (i.e., spatial scalability) and frame rate (i.e., temporal scalability). This allows for simple and flexible adaptation to network (in terms of bandwidth variations and error conditions) and terminal (in terms of frame rate, picture size, computational complexity) capabilities [1]. The motion compensated temporal filtering (MCTF) enables temporal scalability in SVC schemes and also avoids drift between successive frames. The MCTF scheme has been included in 3D wavelet-based video codecs [2-4] and the JSVM encoder, which is an extension of H.264 to provide SNR, spatial and temporal scalabilities [5].

However, the MCTF scheme used in these codecs introduces long end-to-end delay due to large GOP (group

of pictures) size. For example, the end-to-end delay for a 16-frame GOP would be more than 533ms at 30 frames/second, which is not acceptable for real-time conversational video applications. We denote the 5/3 MCTF scheme without low delay feature as ‘conventional 5/3 MCTF’. Reducing the GOP size to obtain lower delay in MCTF, however, reduces the coding efficiency and temporal scalability. A ‘low delay MCTF’ scheme which modifies the ‘conventional 5/3 MCTF’ scheme by partitioning a 16-frame GOP has been proposed in JSVM [5], whose PSNR performance is lower by about 1 dB as compared to the ‘conventional 5/3 MCTF’ scheme in JSVM for our test conditions. Moreover, it introduces non-uniform delay as described in Section 3.

In this paper, we study the residual energy of predictive video frames by using wide-sense stationary (WSS) properties (Section 2) and describe an ‘improved low-delay MCTF structure’ (Section 3) to overcome the aforementioned shortcomings. The proposed scheme is also based on the ‘conventional 5/3 MCTF’ scheme and provides (i) uniform low-delay, and (ii) better coding efficiency than the ‘low-delay MCTF’ scheme (Section 4). The PSNR performance of the proposed scheme is only slightly lower (<0.3dB for our test conditions) than the ‘conventional 5/3 MCTF’ scheme with full bi-directional prediction. Moreover, this scheme can be used with other types of wavelet filters and could operate in closed-loop structure at different temporal decompositions.

## 2. MOTION COMPENSATED RESIDUAL ENERGY

Let  $x[l, k, n]$  represent magnitude of a pixel with coordinate  $(l, k)$  in the  $n^{\text{th}}$  frame. For a video sequence of  $N$  frames with  $L \times K$  pixels, the expectation of motion compensated residual energy can be represented as [6],

$$\begin{aligned} & \frac{1}{N - \tau} \sum_{n=\tau+1}^N \sum_{l=1}^L \sum_{k=1}^K (x[l, k, n] - x[l + a_l, k + b_k, n - \tau])^2 \\ & = 2R_{xx}(0) - 2R_{xx}(\tau) \quad \dots(1) \end{aligned}$$

where  $(a_l, b_k)$  is the motion vector which minimizes the mean squared error (MSE) of the corresponding coding block, and  $\tau$  is the prediction term (or frame lag) which

represents the distance (in frames) between the predicted frame and its reference frame.  $R_{xx}(\tau)$  is the autocorrelation function with lag of  $\tau$ . The prediction of a badly matched block would result in large MSE due to residual energy in eq. (1) being larger than the block energy  $\{R_{xx}(0)\}$ , i.e.,  $2R_{xx}(0) - 2R_{xx}(\tau) > R_{xx}(0)$ . We therefore encode such blocks in intra mode.

We define the normalized motion compensated residual energy for unidirectional prediction as:

$$e^{uni}(\tau) = \frac{2R_{xx}(0) - 2R_{xx}(\tau)}{R_{xx}(0)} = 2(1 - r_{xx}(\tau)) \quad \dots(2)$$

where  $r_{xx}(\tau)$  is the normalized autocorrelation function.

For bidirectional prediction, we have:

$$\begin{aligned} & \frac{1}{N-2\tau} \sum_{n=\tau+1}^{N-\tau} \sum_{l=1}^L \sum_{k=1}^K \left( \frac{x[l, k, n] - x[l + a_l, k + b_k, n - \tau] + x[l + c_l, k + d_k, n + \tau]}{2} \right)^2 \\ &= \frac{3}{2} R_{xx}(0) - 2R_{xx}(\tau) + \frac{1}{2} R_{xx}(2\tau) \quad \dots(3) \end{aligned}$$

For a badly matched block which can not be efficiently predicted in the forward as well as backward direction, we let  $x[l + a_l, k + b_k, n - \tau] = x[l + c_l, k + d_k, n + \tau] = 0$ . In other words, such blocks are intra coded. If an appropriate match is found for a block in either forward or backward direction, we let  $x[l + a_l, k + b_k, n - \tau] \equiv x[l + c_l, k + d_k, n + \tau]$ . In other words, the forward (or backward) unidirectional prediction mode is used to encode such a block instead of using the bi-directional prediction mode.

We define the normalized motion compensated residual energy for the bidirectional prediction as:

$$e^{bi}(\tau) = \frac{\frac{3}{2} R_{xx}(0) - 2R_{xx}(\tau) + \frac{1}{2} R_{xx}(2\tau)}{R_{xx}(0)} = \frac{3}{2} - 2r_{xx}(\tau) + \frac{1}{2} r_{xx}(2\tau) \quad \dots(4)$$

### 3. DESIGN OF THE PROPOSED MCTF SCHEME

In this section, we shall study the *delay* and *coding efficiency* performance of the ‘low delay MCTF’ and the proposed MCTF schemes. The coding efficiency is strongly related to the residual energy after motion compensation. However, the accuracy of motion prediction depends on correlation between the predicted and reference frames. Since the  $r_{xx}(\tau)$  generally monotonically decreases with frame lag ( $\tau$ ) [7], the coding efficiency would decrease with  $\tau$  increasing. This will introduce many ‘badly matched blocks’ in H frames, which must be intra coded and therefore require relatively higher bit budget. The motion

prediction term ( $\tau$ ) generally increases in MCTF with the temporal decomposition level. For example,  $\tau$  is 4(8) frames at the LLH (LLLH) level.

To reduce the delay, the ‘low delay MCTF’ scheme (as shown in Fig. 1, which was proposed in JSVM [5]) divides the 16-frame GOP into two 8-frame partitions, each of which is further split into 3-frame and 5-frame sub-partitions. However, this structure gives *non-uniform delay* due to the use of 3-frame and 5-frame sub-partitions. Moreover, because of the bi-directional prediction used in predicting the LLH frames and the 8-frame lag in predicting the LLLH frame in a GOP, the delay at the LL and LLL levels is equivalent to 8-frame lag. Using  $e^{uni}(\tau)$  and  $e^{bi}(\tau)$  from eq. (2) and (4), the total residual energy in the ‘low delay MCTF’ scheme can be expressed as the sum of the residual energy of all motion compensated frames in a GOP, i.e.,

$$\begin{aligned} E_{low\_delay} &= 6e^{bi}(1) + 2e^{uni}(1) + 2e^{bi}(2) + 2e^{uni}(2) + 2e^{bi}(4) + e^{uni}(8) \\ &= 6e^{bi}(1) + 2e^{uni}(1) + 2e^{bi}(2) + 2e^{uni}(2) + \\ & \quad 2 \left[ \frac{3}{2} - 2r_{xx}(4) + \frac{1}{2} r_{xx}(8) \right] + 2(1 - r_{xx}(8)) \quad \dots(5) \end{aligned}$$

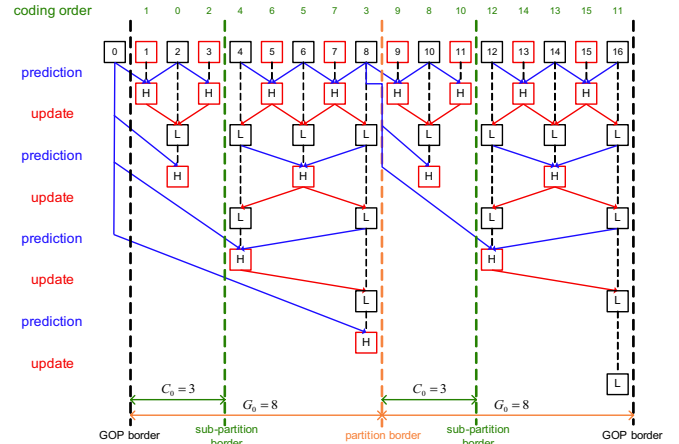


Fig. 1: Illustration of the low delay MCTF scheme [5].

To obtain the *uniform low delay* and *better coding efficiency*, we propose a *new MCTF* structure as shown in Fig. 2. The proposed structure uses 4-frame sub-partitions and 3 unidirectional predictions with maximum of 4-frame lag at LLH level in a 16 frame GOP. In the proposed scheme, all the update steps across the sub-partition boundaries are cancelled in order to obtain low delay. Unlike the ‘low-delay MCTF’ scheme of JSVM, the motion predictions across sub-partition boundaries are retained in the proposed scheme because they use reference frames only from previous sub-partitions. As to the proposed scheme, the corresponding residual energy can be expressed as,

$$\begin{aligned} E_{proposed} &= 8e^{bi}(1) + 4e^{bi}(2) + 3e^{uni}(4) \\ &= 8e^{bi}(1) + 4e^{bi}(2) + 6(1 - r_{xx}(4)) \quad \dots(6) \end{aligned}$$

There is no doubt that  $e^{bi}(\tau) \leq e^{uni}(\tau)$ . Therefore, due to the use of bidirectional motion prediction for each frame at H and LH levels, the coding efficiency of the proposed MCTF scheme for these two levels should be better than that of the ‘low delay MCTF’ scheme. Therefore, from eq. (5) and (6), the energy difference of the two schemes can be represented as,

$$E_{low-delay} - E_{proposed} \geq \frac{1}{2}2(1 - r_{xx}(8)) - 2(1 - r_{xx}(4)) \quad \dots(7)$$

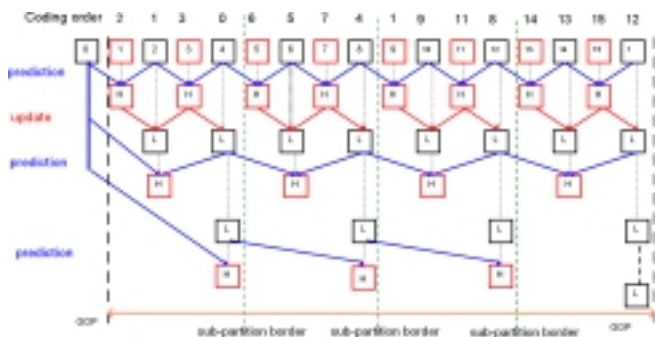


Fig. 2: Illustration of the Improved MCTF.

Here  $2(1 - r_{xx}(8))$  and  $2(1 - r_{xx}(4))$  represent the residual energy of the motion compensated frame with lag of 8 and 4 frames, respectively. Our experiments on the test video sequences show that generally we have  $\frac{1}{3}(1 - r_{xx}(8)) \gg (1 - r_{xx}(4))$ . This indicates that our proposed scheme has lower residual energy in the corresponding temporal levels than in the ‘low-delay MCTF’ scheme. This is also validated by the PSNR results in the next section. Table I shows the delay properties of the three MCTF schemes at different temporal decomposition levels.

Table I: Delay performance of MCTF schemes

Frame rate	Conventional 5/3 MCTF	Low delay 5/3 MCTF	Proposed MCTF
30 fps	533ms	141ms	133ms
15 fps	533ms	150ms	133ms
7.5 fps	533ms	267ms	133ms
3.75 fps	533ms	267ms	N/A
1.875 fps	533ms	533ms	533ms

#### 4. SIMULATION RESULTS

We evaluated our proposed MCTF scheme in JSVM 2.0 encoder and decoder [5] for the five test video sequences (e.g., Foreman, Coastguard, Akiyo, Mother&Daughter, and Football) and compared with the ‘conventional 5/3’ and the ‘low-delay 5/3’ MCTF schemes incorporated in JSVM. The GOP size of 16 is used in the simulation.

Fig. 3 shows the badly matched blocks in the LLLH frame of the 1<sup>st</sup> GOP of QCIF ‘Foreman’ sequence. Here, the proposed MCTF scheme has considerably lower number of badly matched blocks as compared to the ‘low delay MCTF’ scheme. We also observed similar results for the other test video sequences. This would result in better coding efficiency and shorter drift propagation for the proposed scheme during temporal reconstruction at the decoder [8].

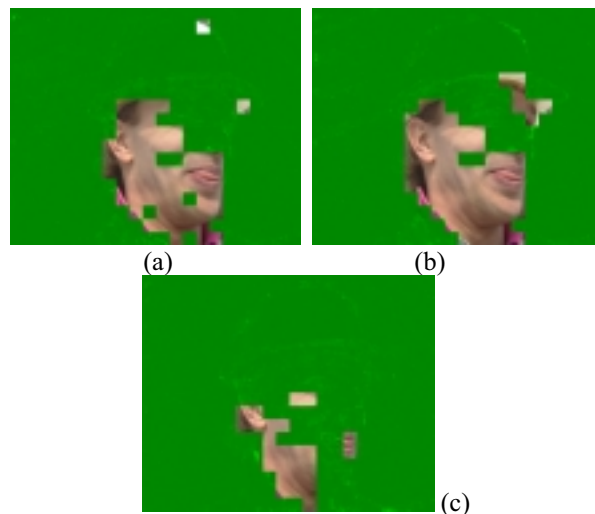


Fig. 3: The illustration of badly matched blocks in an H-frame of ‘Foreman’ sequence by using (a) conventional 5/3, (b) low delay 5/3, and (c) the proposed MCTF schemes.

In Fig. 4, we compare the PSNR performance of the three MCTF schemes for coding the ‘Foreman’ and ‘Coastguard’ video sequences at full temporal resolution of 30 fps for various target bitrates. The proposed MCTF scheme achieves 0.4 to 0.8dB improvement as compared to the ‘low delay MCTF’ scheme. Fig. 5 shows PSNR (for Y-component) for the Y component) for encoding CIF ‘Foreman’ sequence by the three MCTF schemes at frame rates of 1.875, 3.75, 7.5, 15 and 30 fps for target bit rate of 384 Kbps. Again, the proposed scheme outperforms the low-delay 5/3 MCTF scheme by 0.5 to 0.8 dB and is comparable to the ‘conventional 5/3 MCTF’ scheme which employs full bidirectional prediction. We observed similar PSNR improvements for Y and YUV components at different frame rates and target bit rates for the other test video sequences.

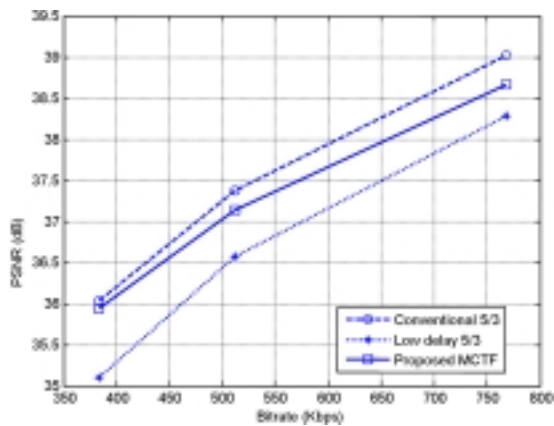
#### 5. CONCLUSION

In this paper, we computed the residual energy of predictive frames (i.e., unidirectional and bidirectional motion prediction) by using the autocorrelation between successive video frames at different frame lags. An improved MCTF structure was proposed and compared to the ‘low delay MCTF’ scheme of JSVM based on the residual energy

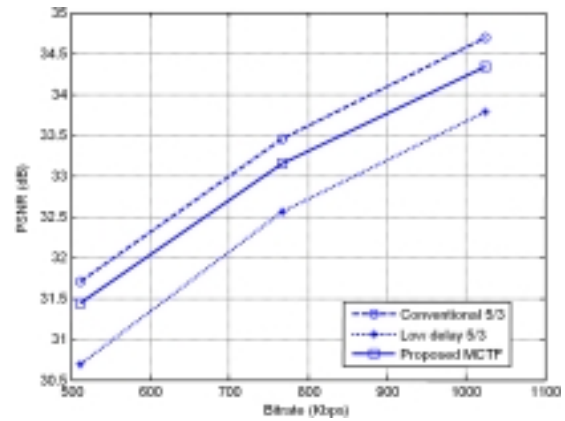
computations. The proposed scheme outperforms the ‘low-delay MCTF’ scheme in terms of delay, the number of badly matched blocks and PSNR (0.4 to 0.8 dB improvement) performance. Our experiments also showed that the PSNR performance of the proposed scheme is only slightly lower than the ‘conventional 5/3 MCTF’ scheme. It should be noted here that the proposed MCTF structure can be easily used with other wavelet filters, such as Haar.

## 6. REFERENCES

1. J.-R. Ohm, “Advances in scalable video coding,” *Proc. IEEE*, Vol. 93(1), pp. 42-56, 2005.
2. R. Xiong, F. Wu, J. Xu, S. Li and Y.-Q. Zhang, “Barbell lifting wavelet transform for highly scalable video coding,” *Picture Coding Symposium*, San Francisco, CA, USA, Dec. 2004.
3. P. Chen and J. W. Woods, “Bidirectional MC-EZBC with lifting implementation,” *IEEE Trans. Cir. Syst. Video Technol.*, Vol. 14(10), pp. 1183-1194, 2004.
4. M. Wien, T. Ruster and K. Hanke, “RWTH proposal for scalable video coding technology,” *ISO/IEC JTC1/WG11/M10569/S16*, March 2004.
5. ITU-T and ISO/IEC JTC1, “Joint Scalable Video Model JSVM-2,” *ISO/IEC MPEG and ITU-T VCEG, JVT-O202*, Busan, Korea, April, 2005.
6. J.-R. Ohm, *Multimedia Communication Technology: Representation, Transmission and Identification of Multimedia Signals*, Springer, 2004.
7. M.M. Krunz and A.M. Makowski, “Modeling video traffic using M/G/ $\infty$  input processes: a compromise between Markovian and LRD models,” *IEEE JSAC*, Vol. 16(5), pp. 733-748, 1998.
8. R. Xiong, J. Xu and F. Wu, “Coding performance comparison between MSRA wavelet video coding and JSVM,” *ISO/IEC JTC1/WG11/m11975*, Apr. 2005.



(a)



(b)

Fig. 4: PSNR performance (Y-component) for encoding CIF sequences at 30 frames/second by the three MCTF schemes, (a) ‘Foreman’ at target bit rates of 384, 512 and 768 Kbps; and (b) ‘Coastguard’ at target bit rates of 512, 768 and 1024 Kbps.

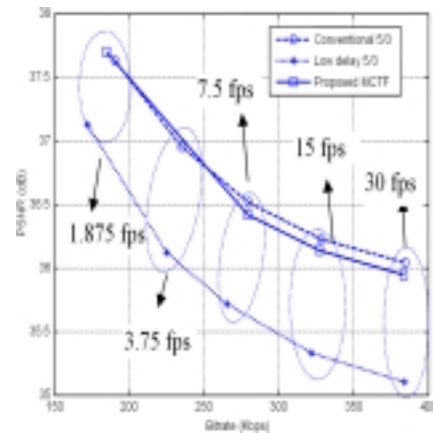


Fig. 5: PSNR (Y-component) for encoding CIF ‘Foreman’ sequence by the three MCTF schemes at frame rates of 1.875, 3.75, 7.5, 15 and 30 fps for target bit rate of 384 Kbps.