ADAPTIVE MCTF BASED ON CORRELATION NOISE MODEL FOR SNR SCALABLE VIDEO CODING

Ruiqin Xiong[1*], *Jizheng Xu*[2], *Feng Wu*[2], *Shipeng Li*[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] Microsoft Research Asia

ABSTRACT

This paper proposes a subband adaptive motion compensated temporal filtering (MCTF) technique for scalable video coding and introduces a revised synthesis gain model for the quantization in this adaptive MCTF scheme. In scalable video coding, hierarchical MCTF is extensively adopted to exploit the temporal correlation across video frames. In this hierarchical MCTF structure, the strength of temporal correlation varies with the level of temporal transform and varies with the various spatial frequency components in a frame. The reconstruction noises also have diverse strength at various subbands. According to the correlation and noise characteristics of various subbands, we can adjust the strength of motion compensated prediction step in MCTF to maximally take the advantage of temporal correlation but restrict the propagation of reconstruction noise. The quantization step of each subband is also adjusted according to synthesis gain determined by the MCTF structure. In this way an adaptive MCTF scheme is formed and the proposed technique improves the coding performance of scalable video coding.

1. INTRODUCTION

In video coding schemes, the motion compensated interframe prediction for exploiting the temporal correlation across frames is the key component whose efficiency influences the overall coding performance significantly. The motion compensated temporal filtering (MCTF) technique with dyadic multi-level temporal decomposition structure is firstly employed in various 3D-wavelet-based video coding schemes, such as [1]~[5]. Due to its good performance in energy compaction and its ability to provide efficient SNR scalability, this MCTF coding framework is also incorporated into the recent state-of-the-art scalable video coding scheme, the scalable extension of H.264/AVC [6][7], with some variations such as omitting the update step. It has been shown that the hierarchical MCTF structure has significant performance gain over the traditional hybrid close-loop predictive coding structure [6][7].

In MCTF, input video frames are transformed into lowpass frames and highpass frames through motion compensated prediction and update steps. In the context of scalable video coding, the reference frames used in motion compensation at decoder are usually not identical with the ones used at encoder. The reference frames are random signals from the viewpoint of decoder, with reconstruction noises added on their original values at encoder. This is an intrinsic characteristic of many efficient SNR-scalable video coding schemes. Therefore, the goal of motion compensated prediction step is to produce a good approximation of current frame from the random reference frame.

In hierarchical MCTF structure, the temporal correlation among neighboring frames has different statistical characteristic at various MCTF levels. The correlation is usually strong at high frame-rate (HFR) MCTF level but relatively weak at low frame-rate (LFR) MCTF level, due to the changes in frame interval. In addition, the strength of temporal correlation exploited in motion compensation may vary with the various spatial frequency components in a frame. The correlation is usually strong for low-frequency components but weak for high-frequency components. Furthermore, the decoder-side reference frames have diverse signal-to-noise ratio (SNR) for various spatial frequency components.

It motivates us to differentiate the various spatial frequency components at each transform level according to their correlation-noise characteristics. The strength of motion compensated prediction can be adjusted adaptively, e.g., to apply strong temporal filtering on signal component with strong correlation and small noise but apply weak filtering or stop filtering on signal components with weak correlation and large noise. In this way, an adaptive MCTF scheme can be formed, not only to take advantage of temporal correlation but also reduce the propagation of reconstruction noise.

With the proposed adaptive MCTF scheme, the temporal transform filter is adjusted adaptively for each subband. Accordingly, the error propagation of each spatial-temporal subband in reconstruction is changed. To achieve an optimized quantization or rate allocation, the estimated synthesis gain should be modeled based on the adaptive MCTF structure.

The rest of this paper is organized as follows. Section 2 reviews the MCTF in scalable video coding and analyzes the correlation-noise characteristics of inter-frame prediction in MCTF. Section 3 describes the proposed adaptive MCTF algorithm. Section 4 proposes a model to estimate

[*] This work has been done while the author is with Microsoft Research Asia

1-4244-0367-7/06/$20.00 ©2006 IEEE

the synthesis gain based on the adaptive MCTF structure. Experimental results are given in section 5 and section 6 concludes this paper.

2. INTER-FRAME CORRELATION AND NOISE CHARACTERISTIC IN MCTF

In video coding with MCTF, input frames are decomposed into lowpass frames and highpass frames through motion compensated prediction and update lifting steps. With the most frequently used biorthogonal 5/3 filter, the prediction step is defined by (1):

$$\begin{aligned} h_k(\mathbf{x}) &= f_{2k+1}(\mathbf{x}) - p_{2k+1}(\mathbf{x}) \\ p_{2k+1}(\mathbf{x}) &= 0.5\times[f_{2k}(\mathbf{x}+\mathbf{mv}_{2k+1}^{left}) + f_{2k+2}(\mathbf{x}+\mathbf{mv}_{2k+1}^{right})] \end{aligned} \tag{1}$$

In the reconstruction process, the lifting step (1) is inversed by (2) and perfect reconstruction can be achieved.

$$\begin{aligned} \hat{f}_{2k+1}(\mathbf{x}) &= \hat{h}_k(\mathbf{x}) + \hat{p}_{2k+1}(\mathbf{x}) \\ \hat{p}_{2k+1}(\mathbf{x}) &= 0.5\times[\hat{f}_{2k}(\mathbf{x}+\mathbf{mv}_{2k+1}^{left}) + \hat{f}_{2k+2}(\mathbf{x}+\mathbf{mv}_{2k+1}^{right})] \end{aligned} \tag{2}$$

Here $\hat{X}$ represents the reconstructed signal of X at decoder.

For non-scalable video coding, both encoder and decoder can always use the same reference. However, in the context of SNR scalable video coding, the quality of reference available at decoder varies with the bit rate. If both encoder and decoder employ the same base-quality reference for motion compensation, like in MPEG-4 FGS [8], it does not fully exploit the temporal correlation in the SNR enhancement layers of neighboring frames and thereby leads to inefficient SNR-scalability performance. Therefore, the idea of open-loop structure is widely adopted for efficient SNR scalability, in which encoder takes high quality reference while decoder uses the best reference it can reconstruct. From the viewpoint of decoder, the references are random signals with reconstruction noises added on their original values at encoder.

The temporal correlation among neighboring frames is essential for video coding but it is not easily to measure it directly. Actually the correlation is realized by motion compensated prediction. It is reasonable to use the correlation between target frame and prediction as the indicator of temporal correlation, as defined in (3):

$$\rho = cor\left[f_{2k+1}(\mathbf{x}), p_{2k+1}(\mathbf{x})\right] \tag{3}$$

Here $cor\left[A, B\right]$ denotes the correlation between random signal *A* and *B*. The ρ value depends on the frame interval, motion field and motion compensation techniques employed in coding. In hierarchical MCTF structure, the temporal correlation varies with the level of temporal transform due to the changes in frame distance. Usually, the correlation becomes weaker as the frame distance increases.

Furthermore, if we divide a frame into several spatial subbands, we can find that the correlation exploited in motion compensation has diverse strength for various spatial frequency components. In motion estimation, the block matching process is usually dominated by low-frequency component since it constitutes the major part of a frame. Moreover, in motion compensation, high-frequency components are more sensitive to the accuracy of motion field than DC or low-frequency components. In addition, the widely used block-based MC technique produces some false high-frequency artifacts at block boundary. Therefore, the temporal correlation is usually weaker for spatial high-frequency subbands than for low-frequency subbands.

Define $\boldsymbol{S}$ as a subband decomposition operator and $\boldsymbol{S}_i(f(\mathbf{x}))$ is the ith subband of $f(\mathbf{x})$. We define the correlation of subband-*i* as (4)

$$\rho_i = cor[\boldsymbol{S}_i(f_{2k+1}(\mathbf{x})), \boldsymbol{S}_i(p_{2k+1}(\mathbf{x}))] \tag{4}$$

Fig. 1 illustrates the process using a DWT decomposition with 16 subbands. Both the current frame and motion compensated prediction frame are decomposed into various spatial subbands and then the correlation parameter ρ_i is calculated on each spatial subband separately.

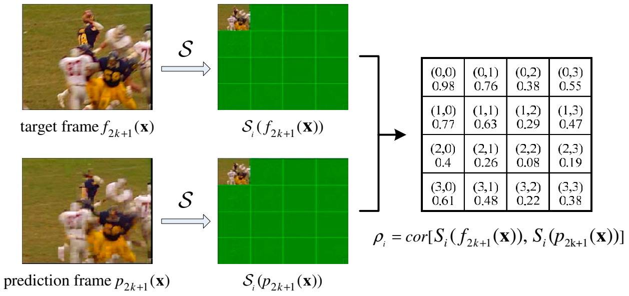

$\mathcal{S}$
$\mathcal{S}$
target frame $f_{2k+1}(\mathbf{x})$
$\mathcal{S}_i(f_{2k+1}(\mathbf{x}))$
prediction frame $p_{2k+1}(\mathbf{x})$
$\mathcal{S}_i(p_{2k+1}(\mathbf{x}))$
(0,0) 0.98
(0,1) 0.76
(0,2) 0.38
(0,3) 0.55
(1,0) 0.77
(1,1) 0.63
(1,2) 0.29
(1,3) 0.47
(2,0) 0.4
(2,1) 0.26
(2,2) 0.08
(2,3) 0.19
(3,0) 0.61
(3,1) 0.48
(3,2) 0.22
(3,3) 0.38
$\rho_i = cor[S_i(f_{2k+1}(\mathbf{x})), S_i(p_{2k+1}(\mathbf{x}))]$

Figure 1: Subband based correlation calculation.

In addition, the reference frames reconstructed at decoder have diverse SNR for various subbands. Compared with low-frequency subbands, high-frequency subbands usually contains less energy in video frames, but they are usually quantized with larger or equal quantization step, and transmitted with a lower or equal priority. Therefore, the SNR of reference frames is usually high at low-frequency subbands but low at high-frequency subbands. This conclusion is also valid for prediction frame $p_{2k+1}(\mathbf{x})$.

Table 1: Correlation and noise characteristic for *Football* (CIF)

Band index	Correlation of f_k and p_k			SNR of p_k (dB)		
	30Hz	15Hz	7.5Hz	30Hz	15Hz	7.5Hz
Band(0,0)	0.99	0.98	0.96	25.60	27.96	29.06
Band(0,1)	0.92	0.77	0.69	14.16	16.04	16.94
Band(1,0)	0.95	0.87	0.78	15.52	17.75	18.84
Band(1,1)	0.79	0.58	0.37	8.28	9.87	10.35
Band(0,3)	0.78	0.53	0.36	9.38	11.51	12.30
Band(0,2)	0.57	0.36	0.19	5.00	7.99	9.01
Band(1,3)	0.67	0.45	0.24	5.56	7.45	7.79
Band(1,2)	0.49	0.28	0.12	1.44	4.06	4.57
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Band(3,3)	0.56	0.35	0.16	2.57	4.93	5.30
Band(3,2)	0.35	0.21	0.08	-1.34	1.83	2.39
Band(2,3)	0.30	0.17	0.07	-1.90	1.66	2.17
Band(2,2)	0.16	0.08	0.02	-4.65	-0.72	-0.22

Table 1 shows the correlation and noise characteristic of the *Football* CIF 30Hz sequence. The obtained parameters are averaged over frames. Please refer to [10] for the relationship between the wavelet subband index and its position in signal spectrum.

3. SUBBAND ADAPTIVE MCTF

Based on the above observations, we proposed an adaptive motion compensated prediction technique for MCTF. It adjusts the strength of MCTF for each spatial subband, according to its correlation and noise characteristics.

We assume that, for a motion compensated prediction step, F denotes the target frame; P denotes the prediction frame produced by traditional MC at encoder; $N = \hat{P} - P$ is the reconstruction noise of P at decoder. N is a random signal which is unknown at encoder but we can assume that (1) N is independent with F and P; (2) N has zero mean; (3) The variance of N is predictable at encoder at a certain bit rate. To differentiate signal components with different correlation-noise characteristics, an invertible subband decomposition $\boldsymbol{S}$ is used to analyze the above signals: $\mathbf{F}=\boldsymbol{S}(F)=(F_1, F_2, \ldots, F_n)^T$, $\mathbf{P}=\boldsymbol{S}(P)=(P_1, P_2, \ldots, P_n)^T$, $\mathbf{N}=\boldsymbol{S}(N)=(N_1, N_2, \ldots, N_n)^T$.

We define a diagonal scaling matrix $\mathbf{A}_{n\text{x}n}$= diag{$\alpha_1, \alpha_2, \ldots, \alpha_n$}. The traditional motion compensated prediction at encoder (1) is modified to (5):

$$\begin{aligned}\mathbf{H} &= S(H) = (H_1, H_2, \ldots, H_n)^T = \mathbf{F} - \mathbf{AP} \\ &= \mathbf{F} - \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_n \end{pmatrix} \mathbf{P}\end{aligned} \quad (5)$$

The motion compensated prediction error $\mathbf{E}$ at decoder is

$$\mathbf{E} = \mathbf{F} - \mathbf{A}\hat{\mathbf{P}} = \mathbf{F} - \mathbf{A}(\mathbf{P}+\mathbf{N}) \quad (6)$$

The goal of motion compensated prediction step is to produce a good approximation for current frame during reconstruction process. The scaling matrix $\mathbf{A}$ is selected to minimize the mean square prediction error D:

$$\begin{aligned} D &= \mathrm{E}[\mathbf{E}^T \cdot \mathbf{E}] \\ &= \mathrm{E}[(\mathbf{F} - \mathbf{A}(\mathbf{P}+\mathbf{N}))^T \cdot (\mathbf{F} - \mathbf{A}(\mathbf{P}+\mathbf{N}))] \\ &= \sum_{i=1}^{n} \mathrm{E}[F_i - \alpha_i (P_i + N_i)]^2 \end{aligned} \quad (7)$$

Here E[·] denotes mathematical expectation. Let $\partial D / \partial \alpha_i = 0$ (i=1,2,…,n), it leads to:

$$\alpha_i = \frac{E[F_i(P_i+N_i)]}{E[(P_i+N_i)^2]} \approx \frac{E[F_iP_i]}{E[P_i^2]+E[N_i^2]} \approx \frac{E[F_iP_i]/E[P_i^2]}{1+E[N_i^2]/E[P_i^2]} \quad (8)$$

The equation (8) is the statistically optimal solution of scaling matrix $\mathbf{A}$. It reflects the correlation and noise characteristic of F and P in two ways. (1) When there is no reconstruction noise on P, we have $\mathrm{E}[N_i^2]/\mathrm{E}[P_i^2]=0$, the scaling parameter $\alpha_i=\mathrm{E}[F_iP_i]/\mathrm{E}[P_i^2]$. We use m_X and σ_X^2 to denote the mean and variance of a random variable X. For spatial highpass subbands, we usually have $\sigma_F^2 \approx \sigma_P^2 \approx \sigma^2$ and $m_F^2 \approx m_P^2 \ll \sigma^2$. With these approximations, we have

$$\alpha_i = \frac{m_{F_i} \cdot m_{P_i} + \rho_i \cdot \sigma_{F_i} \cdot \sigma_{P_i}}{m_{P_i}^2 + \sigma_{P_i}^2} \approx \frac{\rho_i \cdot \sigma_{F_i}}{\sigma_{P_i}} \approx \rho_i \quad (9)$$

It means the scaling parameter is mainly determined by the correlation parameter ρ_i. (2) When there exists reconstruction noise on P, the α_i is reduced by a factor of $1+\mathrm{E}[N_i^2]/\mathrm{E}[P_i^2]$. Here $\mathrm{E}[N_i^2]/\mathrm{E}[P_i^2]$ is the noise-to-signal ratio of P. Therefore, the principle of proposed adaptive technique is to apply strong filtering for subbands with strong correlation but reduce the strength of filtering for subbands with large noise.

When update steps are enabled in MCTF, the scaling parameter is applied to the update steps along the inverse path of motion compensated prediction, using the concept of EDU [11].

4. SYNTHESIS GAIN MODEL FOR QUANTIZATION

When the temporal filtering structure is adjusted adaptively for each subband, the error propagation in reconstruction process is changed accordingly. To achieve optimized quantization or rate allocation, the synthesis gain of each subband should be redressed based on this adaptive structure.

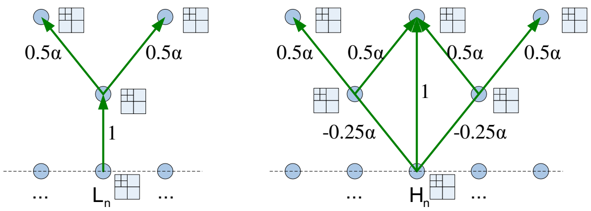


Figure 2: Lifting structure of inverse MCTF.

Fig. 2 shows the error propagation along the lifting steps of inverse MCTF. Since the temporal fluctuation of scaling matrix $\mathbf{A}$ at a MCTF level is small in a short time, we use its temporal average to simplify the model. From the Fig. 2, we can note the temporal synthesis filters for lowpass subband and highpass subband:

$$f_L(n) = \{\alpha/2,\ 1,\ \alpha/2\}$$
$$f_H(n) = \{-\alpha^2/8,\ -\alpha/4,\ 1-\alpha^2/4,\ -\alpha/4,\ -\alpha^2/8\}$$

Therefore, we get the synthesis gain of temporal subband in one MCTF level:

$$\omega_L = \sum\nolimits_n f_L(n)^2 = 1 + \alpha^2/2$$
$$\omega_H = \sum\nolimits_n f_H(n)^2 = 1 - 3\alpha^2/8 + 3\alpha^4/32$$

We can note the synthesis gain depends on $\mathbf{A}$ and varies with spatial subbands and temporal transform levels. When $\alpha = 1$, we have $\omega_L = 1.5$ and $\omega_H = 0.72$, which is the common case of conventional MCTF structure. The temporal lowpass subbands should be coded to a higher accuracy than highpass subbands. When α becomes very small, e.g., $\alpha \approx 0$, we have $\omega_L \approx 1$ and $\omega_H \approx 1$. In this case, both lowpass frame and highpass frame should be quantized equally.

5. EXPERIMENTAL RESULTS

To test the adaptive MCTF algorithm proposed in this paper, we have conduct experiments on many standard MPEG test sequences. The input sequences are of CIF size at 30 fps. Multi-level MCTF is performed on input video and the temporal subbands are transformed and bit-plane coded to generate embedded bitstreams.

For correlation noise characteristic analysis, discrete wavelet transform (DWT) with the structure illustrated in Fig. 1 is employed as the subband decomposition ***S.*** Each frame is divided into 16 subbands. For each subband, the correlation and noise characteristics are analyzed and the scaling matrix is calculated as defined in (8). For noise calculation, a middle reconstruction bit-rate within testing bit-rate range is used.

The performance with or without the proposed adaptive MCTF technique is compared, as illustrated in Fig. 3. The proposed algorithm is marked as "SA" (subband adaptive MCTF) and the baseline (without SA) is marked as "w/o SA". The algorithm with revised quantization is marked as "SA+AQ" (adaptive quantization). The results in Fig. 3 shows that the proposed adaptive MCTF scheme based on correlation and noise characteristics can improve coding performance by 0.1~0.7 dB.

6. CONCLUSIONS

In this paper, an adaptive MCTF scheme is proposed. It adjusts the strength of temporal filtering of each spatial subband at each MCTF level, according to the correlation and noise characteristics of that subband. Strong filtering is applied for subbands with strong correlation but the filtering is weakened for subbands with weak correlation or with large noise. The quantization of each spatial-temporal subband is also adjusted according to the subband synthesis gain determined by the MCTF filtering structure. The proposed adaptive MCTF technique can improve coding performance by up to 0.7dB.

REFERENCES

[1] J. R. Ohm, "Three-dimensional subband coding with motion compensation," IEEE Trans. Image Processing, vol. 3, no. 5, pp. 559-571, Sept. 1994

[2] S. J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," IEEE Trans. Image Processing, vol. 3, no. 2, pp. 155-167, Feb. 1999

[3] J. Xu, Z. Xiong, S. Li, Y.-Q. Zhang, Three-dimensional embedded subband coding with optimal truncation (3D ESCOT), Applied and Computational Harmonic Analysis, vol.10, pp. 290-315, 2001

[4] A. Secker, D. Taubman, Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression, IEEE Trans. on Image Processing, vol. 12, no. 12, pp. 1530-1542, 2002

[5] L. Luo, F. Wu, S. Li, Z. Xiong, Z. Zhuang, "Advanced motion threading for 3D wavelet video coding," Signal Processing: Image Communication, vol. 19, no. 7, pp. 601-616, 2004

[6] Heiko Schwarz, Detlev Marpe, Thomas Wiegand, Scalable extension of H.264/AVC, ISO/IEC JTC1/SC29/WG11 M10569/S03, Munich, 2004

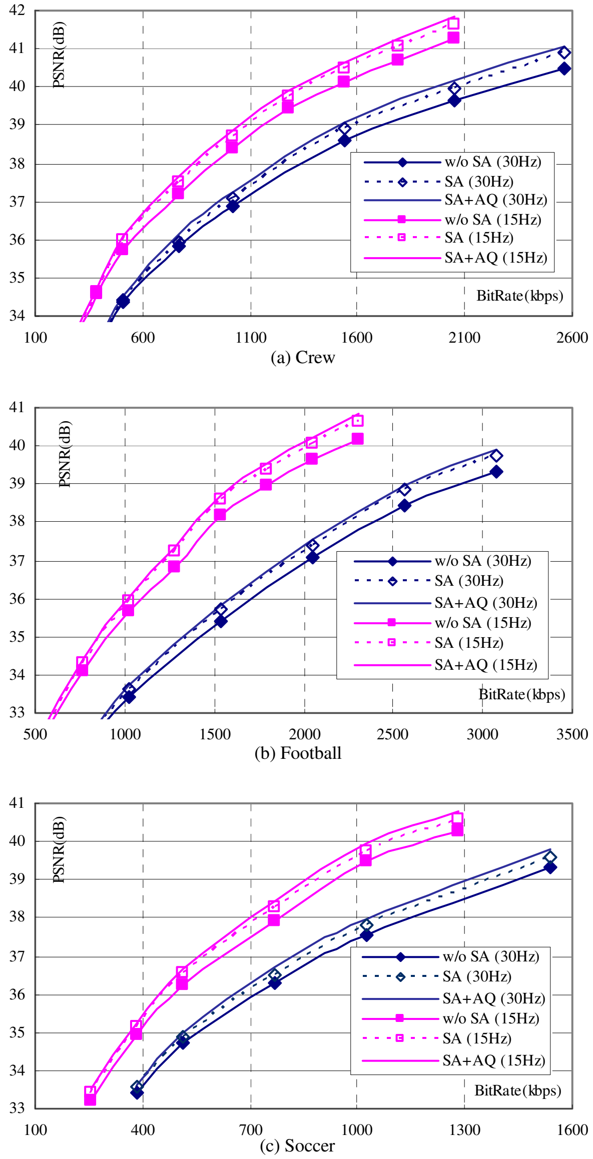
[7] Heiko Schwarz, Detlev Marpe and Thomas Wiegand, MCTF and Scalability Extension of H.264/AVC, Proc. of PCS 2004, San Francisco, CA, USA, Dec. 2004

[8] W. Li. Overview of Fine Granularity Scalability in MPEG-4 Video Standard. IEEE Trans. Circuits and Systems for Video Tech., 2001, 3(11):301-317

[9] A. Secker, D. Taubman, Highly scalable video compression with scalable motion coding, IEEE Trans. on Image Processing, vol. 13, no. 8, pp. 1029-1041, 2004

[10] Donghoon Yu and Jong Beom Ra, Fine Spatial Scalability in Wavelet Based Image Coding, Proc. ICIP 05, pp. 862-865, Genoa, Italy, Sep. 2005

[11] B. Feng, J. Xu, F. Wu, S. Yang, "Energy distributed update steps (EDU) in lifting based motion compensated video coding", ICIP, vol. 4, pp 2267-2270, 2004.



(a) Crew

(b) Football

(c) Soccer

Figure 3: Coding Performance of adaptive MCTF. (a) Crew (b) Football (c) Soccer