Perceptually-adaptive Motion Compensated Temporal Filtering

Dongdong Zhang, Xin Ma, Wenjun Zhang, Xiaokang Yang, Songyu Yu

Institute of Image Communication and Information Processing,
Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT

We propose a perceptually-adaptive motion compensated temporal filtering (MCTF) method to enhance the visual quality of 3D wavelet video coding schemes with spatial-domain MCTF. In our scheme, a spatio-temporal masking model in image domain is incorporated into the lifting structure of MCTF. The model is used to guide the motion search and the prediction step in MCTF so as to remove the visual redundancy in the video sequence. Experimental results show that the proposed scheme can significantly improve the visual quality of decoded video at different bitrates.

1. INTRODUCTION

In the emerging ubiquitous network environment, the scalability of a video codec has been becoming an important feature besides coding efficiency. 3D Wavelet video coding provides an elegant solution for scalable video coding due to its intrinsic multi-resolution nature. In existing 3D wavelet video coding schemes, spatial domain MCTF or wavelet domain MCTF is used to remove temporal redundancy.

In spatial domain MCTF (SDMCTF) schemes [1][2], the rate-constrained motion estimation based on sum of absolute difference (SAD) or sum of squared difference (SSD) is first used to search the optimum motion vector between the current frame and its adjacent frames. Then multi-level lifting-based temporal decomposition along the motion alignment is applied to remove the temporal correlation of original video. With the variable block-size sub-pixel motion search method like that in H.264/AVC, the MCTF scheme can achieve a good objective performance.

However, the blocking artifacts can be perceptually bothersome even at high bitrate since the conventional fidelity criteria, such as SAD, SSD, mean square error (MSE) and the related peak signal-to-noise ratio (PSNR), do not reflect perceptual distortion very well [3]. Since the ultimate receiver of most decompressed video signal is the human visual system (HVS), the goal of video compression and coding should be therefore to achieve the highest perceptual quality at a given bit-rate. It is imperative for us to design a coding algorithm that minimizes perceptual distortion between the original and the decoded visual signal [8].

In this paper, we propose perceptually-adaptive MCTF method (PAMCTF) to enhance the visual quality of SDMCTF video coding scheme. A spatio-temporal masking model in image domain is used to guide the motion search and the prediction step in the lifting structure of MCTF so as to remove the visual redundancy.

2. OVERVIEW OF MOTION COMPENSATED TEMPORAL FILTERING

In this subsection we overview the lifting structure of MCTF in order to describe our proposed PAMCTF method more clearly. Assume that *M*-level temporal filtering is used in the original video. Let $\{L_t^0\}$ denote the original video sequence. $\{L_t^m\}$ and $\{H_t^m\}$ denote the temporal low-pass sequence and high-pass sequence after *m*-level temporal filtering of $\{L_t^0\}$. Figure 1 shows the lifting structure of two-level MCTF with 5/3 filter. Prediction in (1) is first performed to obtain the high-pass frames using consecutive even frames to predict the odd frame, and then the update step in (2) follows the prediction step to complete one level 5/3 subband transform which generates low-pass frames.

$$\begin{aligned} &H_t^{m+1}(x,y) = L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m) \\ &P(L_{2t}^m, L_{2t+2}^m) = \frac{1}{2} MC(L_{2t}^m, MV_{2t+1\to 2t}^{(x,y)}) + \frac{1}{2} MC(L_{2t+2}^m, MV_{2t+1\to 2t+2}^{(x,y)}) \end{aligned} \quad (1)$$

$$\begin{aligned} &L_t^{m+1}(x,y) = L_{2t}^m(x,y) + U(H_{t-1}^{m+1}, H_{t+1}^{m+1}) \\ &U(H_{t-1}^{m+1}, H_{t+1}^{m+1}) = \frac{1}{4} MC(H_{t-1}^{m+1}, MV_{2t\to 2t-1}^{(x,y)}) + \frac{1}{4} MC(H_{t+1}^{m+1}, MV_{2t\to 2t+1}^{(x,y)}) \end{aligned} \quad (2)$$

where, $P()$ and $U()$ denote the prediction and the update operator, respectively. $MC()$ means motion compensation process that builds the temporal pixel correspondence between current frame and the adjacent frames. $MV_{2t+1\to 2t}^{(x,y)}$ and $MV_{2t+1\to 2t+2}^{(x,y)}$ are the sub-pixel motion vectors of the pixel (x,y)

This work was supported by National Natural Science Foundation of China under Grant No. 60332030, No. 60502034 and Shanghai Rising-Star Program under Grant No. 05QMX1435

from an odd frame to the forward and backward adjacent even one.

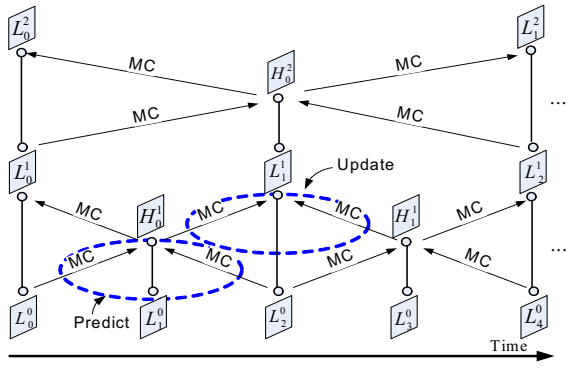

Figure 1. Lifting structure of two-level MCTF with 5/3 filter

3. SPATIO-TEMPORAL VISUAL MASKING

An accurate visual masking model is very important to exploit the perceptual redundancy of video sequences. In this section, a new spatio-temporal masking model is proposed, which can be effectively incorporated in MCTF lifting structure. Basically, it is extended from the nonlinear additivity model for masking (NAMM) for just-noticeable-distortion (JND) in image domain in our previous work [4][8]. In the NAMM, effects of luminance adaptation and texture masking are added with provision to deduct their overlapping. Except for the spatial masking effect of each frame, there is the temporal masking effect between the adjacent frames in a video sequence. Generally, the distortion is more easily perceived by human eyes at the lower frame rate than at the higher frame rate.

Considering these above factors, we model the spatio-temporal visual masking by the visibility threshold $T(x,y,t)$ of a particular pixel location (x,y) at time t as follow:

$$T(x,y,t) = T^s(x,y,t) \cdot T^f(x,y,t) \tag{3}$$

where,

$$T^s(x,y,t) = T^l(x,y,t) + T^t(x,y,t) - C^{lt} \cdot \min\{T^l(x,y,t), T^t(x,y,t)\},$$

$$T^f(x,y,t) = e^{f_s \cdot f_r} f(ild(x,y,t))$$

Here, $T^s(x,y,t)$ and $T^f(x,y,t)$ denote the visibility threshold due to spatial masking and temporal masking, respectively. The threshold $T^s(x,y,t)$ is primarily affected by luminance masking and texture masking. $T^l(x,y,t)$ and $T^t(x,y,t)$ denote the visibility threshold due to the luminance masking and texture masking, respectively. C^{lt} accounts for the overlapping effect in masking. The value of C^{lt} and the computation of $T^l(x,y,t)$ and $T^t(x,y,t)$ follow those in [4]. The threshold $T^f(x,y,t)$ is primarily affected by interframe luminance difference and frame rate f_r. $ild(x,y,t)$ is the average interframe luminance difference between the t-th and $(t-1)$-th frame. $f(ild(x,y,t))$ is a scale factor function of interframe luminance difference [5]. f_s is a constant factor, experientially set to 0.026.

4. PERCEPTUALLY-ADAPTIVE MOTION COMPENSATED TEMPORAL FILTERING

4.1. Perceptually-adaptive rate-constrained motion estimation

The rate-constrained motion searching criterion based on SAD or SSD is used for the variable block-size motion estimation in the existing SDMCTF schemes since it can efficiently balance the distortion and the coding bits of a macroblock/block [1][2]. However, SAD and SSD do not reflect perceptual distortion very well. The effect of SAD or SSD on perceptual quality depends upon not only its mathematical magnitude but also the local masking effect. If a pixel's difference is below the associated visibility threshold, it should be excluded in SAD evaluation because of the invisibility of such difference.

In this work, we propose a perceptually-adaptive rate-constrained motion searching criterion based on the sum of absolute perceptual difference (SAPD). For a given coding partition of a macroblock (MB), the best motion vector of each subblock S_k is found by minimizing the following Lagrange cost function:

$$J(mv,\lambda) = D_{SAPD}(S_k, mv, t) + \lambda \cdot R(mv - mvp) \tag{4}$$

For the given MB, the best mode is found by minimizing the following Lagrange cost function:

$$J(\text{mode},\lambda) = \sum_{S_k \in MB} (D_{SAPD}(S_k, mv, t) + \lambda \cdot R(mv - mvp)) + \lambda \cdot R(\text{mode}) \tag{5}$$

with the perceptual distortion term being given as

$$D_{SAPD}(S_k, mv, t) = \sum_{(i,j)} \left| S_k(i,j,t) - S_k(i+mv_x, j+mv_y, t) \right| - \eta \cdot T_{S_k}(i,j,t) \cdot \delta_{S_k}(i,j,t) \tag{6}$$

where mv, mvp denote the estimated motion vector and its corresponding motion vector predictor for the current block S_k; $R(mv - mvp)$ denotes the bits to transmit motion vector; and S_{ref} represents the referenced block of the current block. $R(\text{mode})$ represents the bits to code mode type. $T_{S_k}(i,j,t)$ is the visibility threshold at the pixel (i,j) of the current block; and

$$\delta_{S_k}(i,j,t) = \begin{cases} 1, & if \ \left| S_k(i,j,t) - S_{ref}(i+mv_x, j+mv_y, t) \right| \geq \eta \cdot T_{S_k}(i,j,t) \\ 0, & otherwise \end{cases} \tag{7}$$

where η is a scale factor. In (6), the absolute of a pixel is excluded in SAPD computation if it is below the associated visibility threshold, which will lead to improvement of perceptual coding quality since the visual property of human eyes is fully considered.

4.2. Perceptually-adaptive prediction

Since human eyes cannot sense any changes below the JND threshold around a coefficient due to their underlying spatial/temporal sensitivity and masking properties, some redundant coefficients below the JND value can be removed safely [6]. From the viewpoint of signal compression, the removal of the visually redundant coefficients will increase the coding bits of the visually important coefficients, thus improve the visual quality. According to the SAPD criterion, we can perceptually adjust the prediction step in the temporal lifting structure as (8) to remove the visual redundancy in the original prediction step.

$$H_t^{m+1}(x,y) = (L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m)) \cdot \alpha_t^{m+1}(x,y)$$
$$P(L_{2t}^m, L_{2t+2}^m) = \frac{1}{2} MC(L_{2t}^m, MV_{2t+1->2t}^{(x,y)}) + \frac{1}{2} MC(L_{2t+2}^m, MV_{2t+1->2t+2}^{(x,y)}) \quad (8)$$
$$m = 0, \ldots, M-1$$

with

$$\alpha_t^{m+1}(x,y) = \begin{cases} 1 - \dfrac{\eta \cdot T^m(x,y,2t+1)}{L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m)}, & if\ L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m) > \eta \cdot T^m(x,y,2t+1) \\ 0, & if\ \left| L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m) \right| \le \eta \cdot T^m(x,y,2t+1) \\ 1 + \dfrac{\eta \cdot T^m(x,y,2t+1)}{L_{2t+1}^m(x,y) - P(L_{2t}^m, L_{2t+2}^m)}, & otherwise \end{cases}$$

where $T^m(x,y,2t+1)$ denotes the spatio-temporal masking value of the pixel (x,y) in the frame L_{2t+1}^m.

5. EXPERIMENTAL RESULTS

We validated the proposed PAMCTF scheme in MPEG scalable video coding (SVC) reference software for wavelet ad-hoc group [7]. In the experiments, four-level motion compensated temporal filtering with 5/3 filter is first used in the original video. And then each temporal subband is decomposed by three-level spatial transform with 9/7 filter. The values of λ and η are set to 16 and 0.2, respectively. The frame rate f_r of the original sequence is 30frames/second.

5.1. Subjective performance

Comparing many decoded sequences between our PAMCTF scheme and the conventional SDMCTF scheme, we found that the visual quality is consistently better for the decoded video with PAMCTF. Figure 2 and Figure 3 show the visual quality comparison of the different decoded Foreman and Football pictures with PAMCTF and SDMCTF, respectively. Where the sequence named "Foreman_CIF_30Hz_256k" means that the bit-stream of Foreman sequence is decoded with image size of CIF at frame rate of 30frames/second and bitrate of 256 kbps. As shown in these figures, some artifacts and noise are removed. In addition, the important detailed texture becomes clearer, such as the Foreman's ears and eyes and the profile of Football's legs. The underlying reason is that when the visual redundancy is removed according to the proposed spatio-temporal masking model, the important coefficients comparatively get more coding bits. Therefore the visual quality of the decoded video can be improved.

In order to further confirm the visual quality improvement by the proposed scheme, we performed subjective quality evaluation. The subjective quality evaluation is performed according to Double Stimulus Continuous Quality Scale method in Rec. ITU-R BT.500 [9]. The Mean Opinion Score (MOS) scales for viewers to vote for the quality after viewing are: Excellent (100-80), Good (80-60), Fair (60-40), Poor (40-20) and Bad (20-0). Five observers were involved in the experiments. The subjective visual quality assessment was performed in a typical laboratory environment, using a 21" SONY G520 professional color monitor with resolution of 1600x1200. The viewing distance is approximately six times of the image height. Difference Mean Opinion Scores (DMOS) are calculated as the difference of MOSs between the original video and the decoded video. The smaller the DMOS is, the higher the perceptual quality of the decoded video is. Table I shows the averaged DMOSs over the all five subjects for Foreman and Football decoded sequences. From the table, we can see that the subjective rating is consistently better for the decoded sequences with the proposed scheme, and the average subjective quality gains of 7.96 and 5.84 measured in DMOS are achieved by the proposed scheme for Foreman and Football sequence, respectively.

5.1. Objective performance

The average PSNR comparison between PAMCTF scheme and SDMCTF scheme is listed in the Table I for the decoded Foreman and Football sequences at different bitrate and spatio-temporal resolution. We can see that our PAMCTF scheme has a better objective quality for the decoded sequences at low resolution and bitrate, and the average objective quality gains of about 0.12dB and 0.3dB are achieved by PAMCTF scheme for Foreman and Football sequences, respectively. For the decoded sequences at high resolution and bitrate, PAMCTF has almost the same objective quality to the SDMCTF.

6. CONCLUSION

We have proposed a perceptually-adaptive MCTF method to enhance the visual quality of 3D wavelet video coding schemes. In our scheme, a spatio-temporal masking model is proposed to guide the motion search and the prediction step in the lifting structure of MCTF so as to remove the visual redundancy in the video sequence. The proposed PAMCTF has been validated to significantly improve the visual quality of decoded video.

7. REFERENCES

[1] P. Chen, K. Hanke, T. Rusert, J. W. Woods, "Improvements to the MC-EZBC scalable video coder," *Proc. IEEE International Conf. on Image Processing, Vol.2,* pp. 14-17, 2003.

[2] R. Xiong, F. Wu, S. Li, Z. Xiong, Y. Zhang, "Exploiting temporal correlation with adaptive block-size motion alignment for 3D wavelet coding," *Proc. SPIE Visual Communications and Image Processing,,* pp. 144-155, 2004.

[3] B. Girod, "What's wrong with mean-squared error?," *Digital Images and Human Vision*, A. Watson, ed., MIT Press, 1993.

[4] X. Yang, W. Lin, Z. Lu, E. Ong and S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color image," *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing, Vol. 3,* pp. 609-612, 2003.

[5] C. Chou and C. Chen, "A perceptually optimized 3-D subband codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol., Vol. 6, no. 2,* pp.143-156, 1992.

[6] N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception", Proc. IEEE, vol. 81, pp. 1385-1422, 1993.

[7] R. Xiong, X. Ji, D. Zhang, J. Xu, G. Pau, M. Trocan and V. Bottreau, "Vidwav Wavelet Video Coding Specifications", Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) ISO/IEC JTC1/SC29/WG11 Document M12339, 2005.

[8] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong and S. S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile", IEEE Trans. Circuits and Systems for Video Technology, vol. 15, no. 6, pp. 742-752, 2005.

[9] ITU-R, Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R Rec. BT. 500-9, Std. 1999.

TABLE I. THE AVERAGE OBJECTIVE AND SUBJECTIVE PERFORMANCE FOR FOREMAN(300 FRAMES) AND FOOTBALL(260 FRAMES) SEQUENCES WITH SDMCTF SCHEME (I) AND PAMCTF SCHEME(II)

Sch-eme	Spatio-temporal resolution	Foreman			Football		
		Bit rate	PSNR (Y)	DM OS	Bit rate	PSNR (Y)	DM OS
I	QCIF@ 7.5Hz	32 kbits	29.273	32.7	128 kbits	30.128	30.2
II			29.383	28.4		30.416	24.6
I	QCIF@ 15Hz	48 kbits	29.784	34.4	192 kbits	29.229	33.7
II			29.908	21.6		29.534	26.8
I	CIF@ 15Hz	96 kbits	30.901	36.5	384 kbits	31.137	35.2
II			30.899	30.1		31.141	29.5
I	CIF@ 15Hz	192 kbits	33.619	27.8	512 kbits	32.489	30.4
II			33.613	19.9		32.497	24.8
I	CIF@ 30Hz	256 kbits	34.071	29.2	1024 kbits	33.912	26.3
II			34.055	20.8		33.901	20.9

(a) (b)

Figure 2. (a)The visual quality for the decoded sequence of Foreman_CIF_30Hz_256k(frame # 96):SDMCTF (left) and PAMCTF (right)
(b)The visual quality for the decoded sequence of Foreman_QCIF_15Hz_48k(frame # 65):SDMCTF (top) and PAMCTF (bottom)

(a) (b)

Figure 3. (a)The visual quality for the decoded sequence of Football_CIF_15Hz_384k(frame # 0):SDMCTF (left) and PAMCTF (right)
(b)The visual quality for the decoded sequence of Football_QCIF_15Hz_192k(frame # 111):SDMCTF (top) and PAMCTF (bottom)