CONTENT ADAPTIVE UPDATE STEPS FOR LIFTING-BASED MOTION COMPENSATED TEMPORAL FILTERING

Li Song[1], JiZhen Xu[2], Hongkai Xiong[1], Feng Wu[2]

[1] Institute of Image Communication, Shanghai JiaoTong University, 200030, Shanghai
[2] Microsoft Research Asia, 100080, Beijing

ABSTRACT

A fundamental difference in the MCTF coding scheme from the conventional compensated DCT schemes is that the predicted residue is further used to update the temporal low-pass frames. However, it may cause the annoying ghost artifact if the predicted residues are generated by inaccurate motion prediction and several temporal high-pass frames are dropped. This paper proposes a content adaptive update scheme, where the HVS (Human Vision System) model is used to evaluate the impact of the update steps in terms of visual quality at the low-pass frames. The potential ghost artifacts detected by the model can be alleviated by adaptively removing visible part of the predicted residues. Experimental results show that the proposed algorithm not only significantly improves subjective visual quality of the temporal low-pass frames but also maintains the PSNR performance compared with the normal full update.

1. INTRODUCTION

Scalable video coding is an efficient and flexible to deliver compressed video over heterogeneous networks and/or in error prone environment. In various scalable video coding schemes, motion compensated/aligned temporal filtering (MCTF) schemes have attracted many attentions because of their inherent properties to support temporal, SNR and even spatial scalability [1][2].

The lifting structure of wavelet transform is widely employed in the temporal decomposition of the 3D sub-band video coding because it facilitates various efficient motion alignment techniques like fractional pixel, variable block size and overlapped block motion alignment [1][2]. With such a power tool, the coding performance of MCTF is improved significantly. The recent results have shown that the MCTF scheme with scalability support can achieve comparable coding performance with the state-of-the-art non-scalable H.264 standard, and even sometimes it outperforms H.264 [3].

The temporal scalability in the MCTF scheme is usually achieved by keeping the temporal low-pass frames while discarding all or some of the high-pass frames in the transform domain. However, it may suffer from the visual artifact. Firstly, block-based motion models that are extensively used in the MCTF scheme cannot represent a non-translator motion perfectly, lots of high frequency and visible residues from the prediction steps will be introduced into the low-pass frame, which will result in ghosting artifact in the low-pass frames and possibly decrease coding efficiency of the subsequent global spatial transform. Secondly, the artifact will be much notable at the decoder if the high-pass frames are partially or completely discarded at lower bit rates or in the case of temporal scaling.

Many techniques have been investigated on the predict steps to reduce the appearance of the ghosting artifact in the low-pass frames by adopting more accurate motion models, for example, fractional pixel motion estimation and compensation [1], overlapped block motion compensation [3], deformable mesh motion model [4]. However, these techniques can reduce the artifact in some cases but still fail in other cases (e.g. scene changes or occluded regions). To further reduce the ghosting artifact, Luo et al [5] argued that the update steps degrade the rate-distortion performance and proposed a truncated lifting scheme, where the update steps are omitted. Turaga et al [6] also proposed to omit the update steps based on quality of motion estimation and the nature of the motion in the sequence. However, Nagita et al [7] showed that skipping the update steps not only impact negatively the compression performance but also increase the fluctuation in PSNR. They proposed a method which adaptively weights the update steps only according to the energy in the high-pass frames that can not reflect the perception of human vision system.

This work has been done while the author is with Microsoft Research Asia.

In this paper, we first propose a content adaptive update scheme, which uses the HVS model to evaluate the signal from the predict step in terms of visual quality. The metric of *JND* (Just Noticeable Difference) detects the signal available from the predict steps. Any part of perception sensitivity will be removed by an adaptive threshold before the update step, which is selected adaptively based on human perception. It is worthy to mention that there is no any overhead to be coded and delivered to the decoder. The encoder and the decoder have the same *JND* metric.

The rest of this paper is organized as follows. Section 2 overviews the motion compensated lifting implementation with the 5/3 filter. In section 3, we investigate the visual impact of the signal from the predict step and propose a new updating scheme with adaptive threshold. Experimental results are given in section 4. Finally, Section 5 concludes this paper.

2. LIFTING BASED MOTION COMPENSATED TEMPORAL DECOMPOSTION

This section first overviews the lifting-based motion compensated temporal decomposition with the 5/3 filter. Assume that a video sequence, $\boldsymbol{I_0, I_1,\ldots, I_{2n-1}}$ are to be processed with temporal transform. *Figure 1* shows how the lifting based temporal transform is performed with the bi-orthogonal 5/3 wavelet filter.

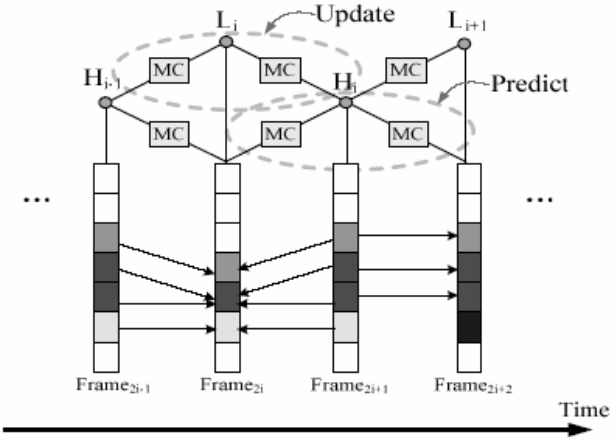


Figure 1: One level temporal transform in lifting based motion compensated video coding

The first step is prediction to calculate the high-pass frame, which predicts the odd frame from consecutive even frames as follows,

$$H_i = I_{2i+1} - P(I_{2i+1}), \tag{1}$$

where

$$P(I_{2i+1}) = \frac{1}{2}(MC(I_{2i,}MV_{2i+1->2i}) + MC(I_{2i+2}, MV_{2i+1->2i+2}))$$

H_i is the high-pass frame generated in the predict step. $MV_{2i+1->2i}$ mean motion vectors from the frame *2i+1* to the frame *2i*. So do $MV_{2i+1->2i+2}$. And *MC* () means motion compensation process that generates the current frame's prediction from its consecutive frame.

The update step follows the predict step to complete one level 5/3 sub-band transform which generates the low-pass frame

$$L_i = I_{2i} + U(I_{2i}) \tag{2}$$

where

$$U(I_{2i}) = \frac{1}{4}((MC(H_{i-1}, MV_{2i->2i-1}) + MC(H_i, MV_{2i->2i+1}))$$

Since the predict step (seeing equation (1)) attempts to minimize the bit-rate required to encode the high-pass frame along with motion vectors used for prediction, H_i is essentially the residue from bi-directional motion compensated prediction of the relevant odd indexed input video frames I_{2i+1}. Then the "original" even indexed frame I_{2i} is updated with the predicted residues as the low-pass frame (see equation (2)).

If the motion prediction is accurate, the predicted residue usually has small magnitude. When it is used to update, it facilitates to remove the high frequent part in the low-pass frame, thus improving coding efficiency. However, if the motion prediction is inaccurate, it would introduce ghost in edges and contours and increase the energy of high frequency. Obviously, it has a negative effect on the coding performance. In particular, the situation at the decoder is more serious when some high-pass temporal frames are partially and completely dropped due to limited channel bandwidth and device capability. The ghost at the low-pass frames can never be removed even more bits are allocated to these frames.

3. CONTENT ADPATIVE UPDATE STEPS BASED ON NOISE VISIBILITY FUNCTION

As discussed in Section 2, the update step may have a negative effect in terms of PSNR and visual quality when motion prediction is inaccurate. However, it will definitely hurt the coding efficiency if the update step is removed totally. To solve the dilemma, a better update scheme should achieve the good trade-off between the two contradicting goals of adding original update information as much as possible and at the same time, decreasing the visual artifacts introduced by the update steps. Specifically, the proposed update step is generalized as follows:

$$L_i = I_{2i} + f(U_{2i}) \tag{3}$$

$f()$ means the proposed adaptation function. The function introduced here takes advantage of the results and developments of human visual model in computer vision.

Extensive researches have been conducted to develop computing models based on the human visual system and it is found that there is inconsistency in sensitivity of the HVS to stimuli of varying levels of contrast and luminance changes in the spatial and temporal domain. Among numerous computing models of the HVS, the *JND* (Just Noticeable Difference) is widely used in perceptual coding or image watermarking. It is referred as visibility threshold that are defined as functions of the amplitude of luminance edge in which perturbation is increased until it becomes just discernible [8]. Indeed, the *JND* is closely related with texture masking property of HVS: the noise is more visible in flat or texture-less areas and less visible in region with edges and textures. The *JND* thresholds are image dependent, and as long as the update information remains below these thresholds, we achieve "update residual" transparency. Therefore, the *JND* matches very well with the subject of update steps addressed before.

In this intuitive implementation of the adaptive update scheme, we define the following *JND* models:

$$JND_x(i,j) = 1 - \frac{1}{1+\theta\sigma_x^2(i,j)} \cdot \qquad (4)$$

Where $\sigma_x^2(i,j)$ denotes the local variance of the image $\boldsymbol{x}$ in a window centred on the pixel with coordinates (i, j), θ is a tuning parameter that can be chosen for particular image. The above *JND* definition is similar to the threshold model proposed by Voloshynovskiy used for image watermarking [9], and the second item of equation (4) is same as the *NVF* (noise visibility function) value supposed that image is non-stationary Gaussian process.

The θ plays the role of contrast adjustment in *JND*, to make θ image-dependent, we can compute it as follows:

$$\theta = \frac{D}{\sigma_{x_{\max}}^2} \qquad (5)$$

$\sigma_{x_{\max}}^2$ is the maximum local variance for a given image, and $D \in [50,100]$ is an experimentally determined parameter. It can be seen that the *JND* value is small in flat areas since $\sigma_x^2(i,j)$ is small and vice versa.

Based on the above analysis, we propose the following content adaptive update (CAU) steps:

$$L_i = I_{2i} + f(I_{2i}, U_{2i}) \qquad (6)$$

where

$$f(I_{2i}, U_{2i}) = \begin{cases} U_{2i} & |U_{2i}| < JND_{I_{2i}} \cdot \mathrm{S} \\ JND_{I_{2i}} \cdot \mathrm{S} & U_{2i} \geq JND_{I_{2i}} \cdot \mathrm{S} \\ -JND_{I_{2i}} \cdot \mathrm{S} & U_{2i} \leq -JND_{I_{2i}} \cdot \mathrm{S} \end{cases}$$

(7)

$JND_{I_{2i}}$ is defined as equation (4), and S denotes the strength factor. Since *JND* function is adaptive with local frame characteristics associated with visual masking and it is computed from the odd index frame to be updated, the proposed update steps can effectively alleviate ghosting artifacts from the high-pass frame and improve the coding performance for temporal scalability.

There is no any overhead to be coded and delivered to the decoder for adaptive update. The encoder and the decoder use the same *JND* metric. Although they operate on different image (the original at the encoder and the reconstructed at the decoder), experimental results have shown that the resulting update mask at the decoder is a very close approximation to that at the encoder.

4. EXPERIMENTAL RESULTS

We have conducted extensive experiments to test the performance of our proposed update steps. The motion threading (5,3) lifting-based MCTF scheme [2] is selected as the test benchmark in this paper, each sequence is temporally de-composed into four-layer and each temporal frame is further spatially decomposed by spatial wavelet transform. The resulted wavelet coefficients are coded and truncated to the target bit rate. The parameters of equation (7) are same for all sequences: the size of the window that we compute local variance is 3x3, the strength factor S=12.5; and D=100.

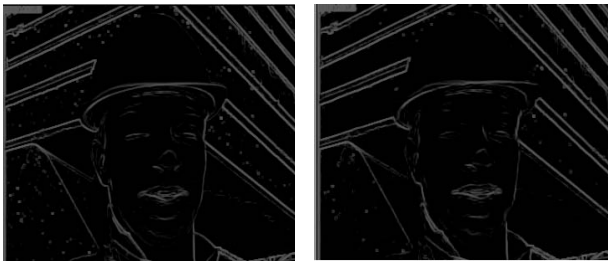
(a) Normal full update

(b) Content adaptive update

Figure 2: visual quality comparison for foreman.

In order to demonstrate the improvement of our techniques, comparison is done with the normal full update scheme. *Figure 2* shows the improved quality of the reconstructed video frame #1 at 192kbps and at half frame rate (15 frames) for the *foreman* sequence. Obviously the update steps in the proposed CAU scheme work best in removing the artefacts from low pass frames.

Since the *JND* map does not include in bit stream, the decoder estimates this *JND* map from the received low pass frame. In theory, there are mismatch between the encoder and the decoder; however, our experiments show we can get a very close approximation in the decoder as shown in *Figure 3*.



(a)The encoder *JND* map (b) the decoder *JND* map

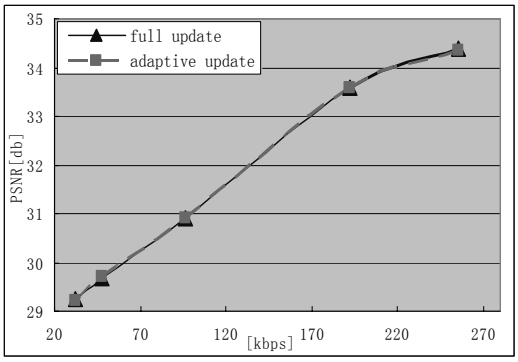
Figure 3: The JND map comparison for frame in figure 2 in the encoder and the decoder.

(a) Classical full update

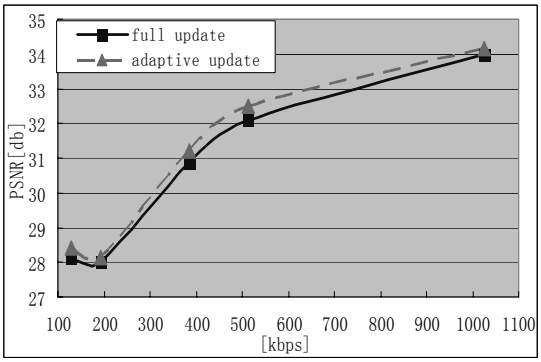
(b) Content adaptive update

Figure 4: visual quality comparison for football.

Figure 4 show the results of the *football* sequence, which is the reconstructed video frame #9 at 512kbps and at 15 fps. As expected, many artefacts in normal full update steps are removed. The coding performance comparison of full update steps and the proposed CAU scheme is depicted in *Figure 5* (*foreman and football*) at different bit rates and at different spatial-temporal scalability. We can see there is no loss for *foreman* sequence and PSNR has gain for *football* sequence.



(a) Code performance comparison for *foreman*



(b) Code performance comparison for *football*

Figure 5: The performance evolutions of the proposed technique.

5. CONCLUSIONS

A content adaptive update step is proposed in the lifting based MCTF, which uses the HVS property to adaptively threshold the update information. The experimental results validate the effectiveness of the proposed update scheme. The CAU scheme can be used with any wavelets-based scalable video coding architecture, for example, MCTF based t+2D techniques and In-band motion compensation based 2D+t techniques since the *JND* model can be applied to wavelet domain as well. In this preliminary implementation, we compute *JND* only using texture masking property of HVS; we will investigate more complicated and effective visual models to further reduce the ghosting artifacts in lifting based MCTF.

6. REFERENCES

[1]. P. Chen, K. Hanke, T. Rusert, and J. W. Woods, "Improvements to the MC-EZBC scalable video coder", Proceedings of the IEEE Int. Conf. on Image Processing (ICIP2003), Barcelona, vol.2, pp.14-17, September 2003.

[2]. L. Luo, F. Wu, S. Li, and Z. Zhuang, "Advanced lifting-based Motion Threading (MTh) techniques for 3D wavelet video coding", Proceedings of the SPIE/IEEE Visual Communications and Image Processing (VCIP2003), Luhano, Switzerland, Vol.5150, pp.707-718, Jul.2003.

[3]. R. Xiong, F. Wu, S. Li, Z. Xiong and Y.-Q. Zhang, "Exploiting temporal correlation with adaptive block-size motion alignment for 3D wavelet coding", SPIE/IEEE Visual Communications and Image Processing (VCIP2004), San Jose, California, USA, Jan.2004.

[4]. A. Secker, and D. Taubman, "Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation", Proceedings of the IEEE Int. Conf. on Image Processing (ICIP2002), Rochester,Vol.3 ,pp.24-28, June 2002

[5]. L. Luo, S. Li, Z. Zhuang, and Y.-Q. Zhang, "Motion compensated lifting wavelet and its application in video coding", Proceedings of the IEEE Int. Conf. on Multimedia and Expo (ICME2001), Tokyo, pp. 365–368, August 2001.

[6]. D. S. Turaga, M. van der Schaar, "Content-adaptive filtering in the UMCTF framework", Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2003), Hong Kong, Vol. 3, pp.6-10, April 2003.

[7]. N. Mehrseresht, and D. Taubman, "Adaptively weighted update steps in motion compensated lifting based on scalable video compression" , Proceedings of the IEEE Int. Conf. on Image Processing (ICIP2003), Barcelona, vol.2, pp.771-774, September 2003.

[8]. A. N. Netravali, and B. Prasada, "Adaptive quantization of picture signals using spatial masking", Proceedings of the IEEE, vol.65, pp.536-548, Apr.1977.

[9]. S. Voloshynovskiy, A. Herrigel, N. Baumgärtner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking, In International Workshop on Information Hiding", Vol. LNCS1768 of Lecture Notes in Computer Science, Dresden, pp. 212-236, October 1999.