Perceptually-adaptive in-band preprocessing for 3-D wavelet video coding

Dongdong Zhang
Wenjun Zhang
Xiaokang Yang
Shanghai Jiao Tong University
Institute of Image Communication and
Information Processing
Shanghai 200240, China

Abstract. We present a perceptually-adaptive in-band preprocessing scheme for 3-D wavelet video coding. In our scheme, after the original video is decomposed by 2-D spatial wavelet transform, a preprocessor is incorporated to remove some visually insignificant wavelet coefficients (noise-like) before the motion compensated temporal filtering of each spatial subband. The preprocessing process is guided by a wavelet domain just-noticeable-distortion profile, which locally adapts to spatial wavelet transform coefficients. Experimental results show that the proposed scheme can efficiently enhance the visual quality of coded video with the same objective quality at different bitrates.
[DOI: 10.1117/1.2180771]

Subject terms: perceptually-adaptive in-band preprocessing; wavelet video coding.

Paper 050828LR received Oct. 16, 2005; revised manuscript received Dec. 29, 2005; accepted for publication Jan. 3, 2006; published online Mar. 13, 2006.

Three-dimensional wavelet video coding has been investigated by many scholars because its multiresolution nature can support spatial and temporal scalabilities simultaneously. Of the various wavelet video coding schemes, most can be classified into two categories: "T+2D" and "2D+T".[1] The major difference between them is whether the temporal transform is implemented before spatial decomposition or not. Since motion compensated temporal filtering (MCTF) is usually used for the temporal transform, "T+2D" is also called a spatial domain MCTF (SD-MCTF) scheme and "2D+T" is called an in-band MCTF (IBMCTF) scheme. The IBMCTF scheme is particularly attractive because of its inherent spatial scalability and flexible coding framework.

In the IBMCTF scheme, the coefficients of each spatial band obtained by 2-D spatial wavelet decomposition have some perceptual redundancy. At a given bitrate, if such visually redundant coefficients are completely coded, it will lead to the decrease of coding bits for comparative important coefficients in the spatial band, thus the overall perceptual quality of the coded video will be deteriorated. In fact, some redundant coefficients below the just-noticeable-distortion (JND) value can be removed safely since human eyes cannot sense any changes below the JND threshold around a coefficient due to their underlying spatial/temporal sensitivity and masking properties.[2] From the signal compression viewpoint, the removal of the visually redundant coefficients will increase the coding bits of the visually important coefficients, improving visual quality.

In this paper, we propose a perceptually-adaptive preprocessing method for in-band MCTF-based 3-D wavelet video coding. A locally adaptive wavelet domain JND profile is first proposed, which is then incorporated into a preprocessor of the in-band MCTF to remove the visually redundant coefficients before performing the MCTF of each spatial band.

Figure 1 shows the framework of the proposed perceptually-adaptive in-band preprocessing scheme for 3-D wavelet video coding. The spatial wavelet transform is first applied to the original video sequence, which generates multiple spatial bands. Then each spatial band is preprocessed to remove the visually insignificant coefficients guided by a wavelet domain JND profile, which is built according to both the local property of each wavelet coefficient and the quantization noise visibility of each spatial band. After preprocessing, MCTF is performed to exploit the temporal correlation within each spatial band. For each temporal band of a certain spatial band, the spatial transform can be further employed to exploit the spatial correlation. Finally, the residual coefficients, motion vectors and modes of each spatiotemporal band are coded independently so that the server can simply drop the unnecessary spatiotemporal bands according to the resolution requested by the client.

Since human eyes have underlying spatial/temporal sensitivity and masking properties, an appropriate JND model can significantly help to improve the performance of video coding algorithms. Several methods for finding JND have been proposed based upon intensive research in subbands as well as some work in the image domain.[3–5] Watson et al.[3] constructed the model of discrete wavelet transform (DWT) noise visibility thresholds as a function of scale, orientation, and display visual resolution. Their threshold model is based on the psychovisual detection of noise injected to wavelet bands. In their model, the local property of each wavelet coefficient was not considered, so each coefficient in a spatial band shares the same threshold.

Based on Watson's threshold model, we formulate a locally adaptive wavelet domain JND profile as given in Eq. (1), in which the Watson's band-wise thresholds are modulated by the local activity factor of each wavelet coefficient:

$$JND_{Th}(l,\theta,i,j)=S_t(l,\theta,i,j)T(\theta,f) \tag{1}$$

where $T(\theta,f)$ is the threshold of the quantization noise visibility of each spatial band, $S_t(l,\theta,i,j)$ is a local activity factor, l denotes the scale of spatial wavelet transform, θ is the different spatial band after each spatial wavelet transform, and its possible values of θ are $\{1,2,3,4\}$, corresponding to the spatial low-low-pass band (LL), high-low-pass band (HL), high-high-pass band (HH) and low-high-pass band (LH), and i and j denote the coordinates of the coefficient of each spatial band. The threshold $T(\theta,f)$ can be computed as follows[3]:

$$\log[T(\theta,f)]=\log(a)+k[\log(f)-\log(g_\theta f_0)]^2 \tag{2}$$

where f denotes the spatial frequency, which is determined by the viewing condition (maximum display resolution and

0091-3286/2006/$22.00 © 2006 SPIE

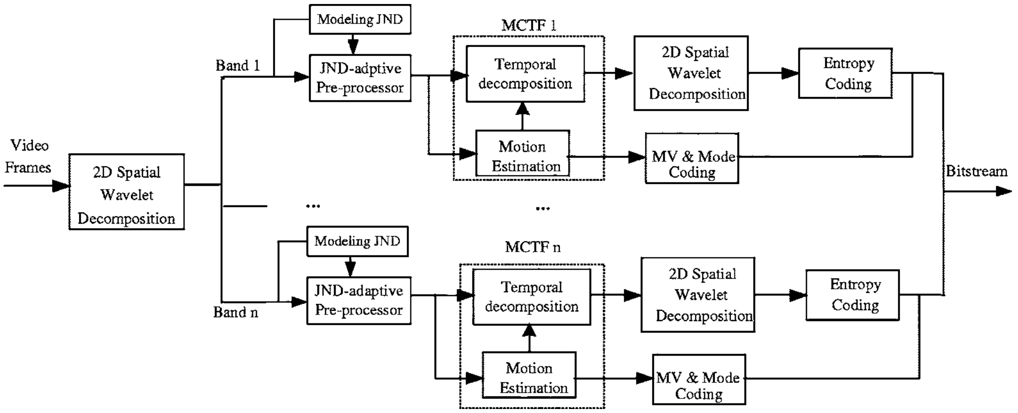


Fig. 1 Framework of perceptually-adaptive in-band preprocessing scheme.

viewing distance). In our implementation, the value of f is 32 cycles/deg for the Y component and 16 cycles/deg for the C_r and C_b components. Here $\log[T(\theta,f)]$ is a parabola with a minimum at $g_\theta f_0$ and a width of k^{-2}. The optimized parameters a, k, g_θ, and f_0 can follow the corresponding values in Ref. 3.

Considering that variance is a good indication of the local activity, we define the local activity factor of each wavelet coefficient $S_t(l,\theta,i,j)$ as follows:

$$S_t(l,\theta,i,j) = 1 - \frac{1}{1 + \lambda\sigma_R^2(l,\theta,i,j)} \tag{3}$$

where $\sigma_R^2(l,\theta,i,j)$ is the local variance in a $R \times R$ window centered at (i,j) in the spatial band (l,θ). The second item in the expression is similar to the most known form of the empirical noise visibility function (NVF) in image restoration applications.[6] It is the basic prototype for many adaptive regularization algorithms in the image domain.[7,8] Since the wavelet coefficients still have strong local activity even in the spatial high-frequency band, we can apply this prototype to the wavelet domain. Here λ is a subband-dependent contrast adjustment parameter computed as in Eq. (4), assuming that the noise can be modeled by a nonstationary Gaussian process.[7]

$$\lambda = \frac{D}{\sigma_{\max}^2(l,\theta)} \tag{4}$$

where $\sigma_{\max}^2(l,\theta)$ is the maximum local variance for the spatial band (l,θ), and $D \in [50,100]$ is an empirical parameter.

The above adjustment factor shows that the JND values in the highly textured and edged areas are stronger than those in the flat regions in the same subband. With the above wavelet domain JND, we can define the following perceptually adaptive in- band preprocessor:

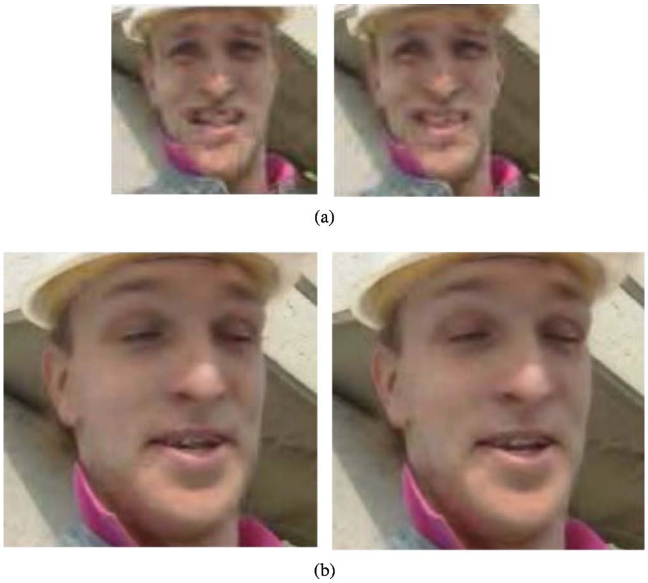
$$Co(l,\theta,i,j) = \begin{cases} 0 & \text{if } |Co(l,\theta,i,j)| \le JND_{TH}(l,\theta,i,j) \\ Co(l,\theta,i,j) & \text{else} \end{cases}, \tag{5}$$

where $Co(l,\theta,i,j)$ is the coefficient value at the coordinate (i,j) in the spatial band (l,θ).

In the above preprocessor, if a coefficient is below the wavelet domain JND value, it will be viewed as insignificant and set to be zero. Since the JND profile is locally adaptive, after this processor the visually insignificant coefficients are removed while the visually significant coefficients can remain. It will benefit the following processing process of each spatial band since the corresponding coding bits for the visually important coefficients will be increased. Thus the overall visual quality of coded video will be improved.

We validated the perceptually-adaptive in-band preprocessing scheme in MPEG scalable video coding (SVC) reference software for a wavelet ad-hoc group.[9] In the experiments, the video is first decomposed into four spatial bands with a 9/7 filter. The coefficients of each spatial band are then perceptually preprocessed with the proposed scheme, in which the window size is 5×5 for computing local variance and the contrast adjustment factor D is set to be 100. After the preprocessing step, a four-level MCTF with a 5/3 filter is performed in each spatial band.

Figure 2 shows the visual quality comparison of the dif-


(a)

(b)

Fig. 2 Decoded pictures for Foreman sequences with different schemes. (a) Foreman_QCIF_15Hz_48k (frame No. 23): without preprocessing (left) and with preprocessing (right). (b) Foreman_CIF_30Hz_256k (frame No. 2): without preprocessing (left) and with preprocessing (right).

ferent decoded Foreman sequences with preprocessing and without preprocessing, respectively. In the figure, the decoded sequence named "Foreman_QCIF_15Hz_48k" means that the bit-stream of the "Foreman" sequence is decoded with image size of QCIF at a frame rate of 15 frames/s and a bitrate of 48 kbits/s. We can see that the visual quality is consistently better for the decoded video with the proposed preprocessing method at different resolution, different frame rate, and different bitrate. As shown in the figure, some artifacts and noise are removed. It makes that the flat areas, such as Foreman's face and neck, look more smooth and comfortable. In addition, some important detail texture becomes clearer, such as Foreman's mouth, teeth, and ears.

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 the double stimulus continuous quality scale method in Rec. ITU-R BT.500.[10] 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-in. SONY G520 professional color monitor with a resolution of 1600 × 1200. The viewing distance is approximately six times that 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 1 shows the averaged DMOSs over all five subjects for the Foreman decoded sequences, where scheme I and II denote the IBMCTF without preprocessing and with preprocessing, respectively. 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 6.71 measured in DMOS is achieved by the proposed scheme.

The PSNR results for the Foreman decoded sequences are listed in Table 1. From the table, we can find that the IBMCTF scheme with the proposed preprocessing has almost the same PSNR performance as the IBMCTF scheme without preprocessing. Interestingly, the objective coding performance does not increase. The underlying reason may be that signal distortion of a conventional IBMCTF is introduced by the embedded quantization for wavelet coefficients, while additional distortion from the JND-adaptive preprocessing needs to be considered in the proposed scheme. Therefore, although the removal of the visually insignificant coefficients can save some bits for coding the visually significant coefficients, it cannot guarantee the improvement of the overall objective quality measured by PSNR due to the additional signal distortion from preprocessing. In the motion-compensated residues preprocessor for the close-loop predictive coding paradigm,[5] a method for determining the optimum parameter has been devised for improvement of PSNR at a given bitrate for nonscalable video coding. But such an optimization idea is inapplicable for the open-loop MCTF coding paradigm, which has to adapt to a wide range of bitrate and spatiotemporal resolutions. Therefore, the proposed preprocessing scheme ensures the improvement of the overall subjective quality instead of the objective quality.

Table 1 Average objective and subjective performance for Foreman (300 frames) sequence without preprocessing (scheme I) and with preprocessing (scheme II).

Decoded sequence	Scheme	PSNR(Y)	PSNR(U)	PSNR(V)	DMOS
QCIF_7.5Hz_32k	I	29.2284	36.1330	34.9841	33.70
	II	29.2434	36.2042	35.0537	29.42
QCIF_15Hz_48k	I	29.9829	36.7561	36.9186	35.67
	II	30.0137	36.8749	37.0029	28.34
CIF_15Hz_96k	I	30.8515	37.2755	38.0539	37.28
	II	30.8696	37.2846	38.0390	31.59
CIF_15Hz_192k	I	33.3847	39.1865	40.2487	28.96
	II	33.3322	39.1912	40.3064	20.81
CIF_30Hz_256k	I	33.7564	39.5507	40.7925	30.15
	II	33.7156	39.5924	40.8259	22.07

Acknowledgment

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

References

1. J.-R. Ohm, "Advances in scalable video coding," *Proc. IEEE* **93**(1),42–56 (2005).
2. N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception," *Proc. IEEE* **81**, 1385–1422 (1993).
3. A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.* **6**(8), 1164–1175 (1997).
4. I. S. Hontsch and L. J. Karam, "Adaptive image coding with perceptual distortion control," *IEEE Trans. Image Process.* **11**(3), 213–222 (2002).
5. 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 Syst. Video Technol.* **15**(6), 742–752 (2005).
6. S. Efstratiadis and A. Katsaggelos, "Adaptive iterative image restoration with reduced computational load," *Opt. Eng.* **29**(12) 1458–1468 (1990).
7. 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 (1999).
8. L. Song, J. Z. Xu, H. K. Xiong, and F. Wu, "Content adaptive update steps for lifting-based motion compensated temporal filtering," *Proc. Picture Coding Symp.*, pp. 589–593 (2004).
9. R. Q. Xiong, X. Y. Ji, D. D. Zhang, J. Z. Xu, G. Pau, M. Trocan, and V. Bottreau, "Vidwav wavelet video coding specifications," Intl. Standards Org./Intl. Electrotech. Comm., ISO/IEC JTC1/SC29/WG11 Document M12339 (2005).
10. ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R Rec. BT. 500-9, Std. (1999).