A NOVEL VISUAL DISTORTION SENSITIVITY ANALYSIS FOR VIDEO ENCODER BIT ALLOCATION

Chih-Wei Tang[2], Ching-Ho Chen[1], and Ya-Hui Yu[1], Chun-Jen Tsai[1+]*
[1] *Department of Computer Science and Information Engineering,*
[2] *Department of Electronics Engineering,*
National Chiao Tung University, Hsinchu, Taiwan, R.O.C
[+]cjtsai@csie.nctu.edu.tw , [*]chihwei.ee88g@nctu.edu.tw

ABSTRACT

A novel video bit allocation technique adopting a visual distortion sensitivity model for better rate-visual distortion coding control is proposed in this paper. Instead of applying complicated semantic understanding, the proposed automatic distortion sensitivity analysis process analyzes both the motion and the texture structures in the video sequences in order to achieve better bit allocation for rate-constrained video coding. This analysis evaluates the tolerable perceptual distortions on a macroblock basis, and allocates fewer bits to regions permitting large perceptual distortions for rate reduction. The proposed algorithm can be incorporated into any existing video coding rate control schemes to achieve same visual quality at greatly reduced bitrate. Experiments based on H.264 show that this technique achieves bit-rate saving of up to 40% with no perceptual quality degradations. The experiments also demonstrate the inadequacy of using PSNR as a distortion measure in a video coding framework.

1. INTRODUCTION

Rate control plays a key role in a high quality video encoder. The goal is to achieve the best perceptual picture quality at a given bit rate through a proper bit allocation process. Existing rate control practices (such as MPEG-4 Annex-L) analyze motion activity predictability, for example, the magnitude of mean absolute differences (MAD), for bit-allocation. However, from visual perception point of view, a hard-to-predict area does not necessarily catches as much human attention as an easily predictable area. In order to achieve constant visual quality across different area with optimal bit-allocation, psychophysical model must be taken into account in the bit allocation process.

Several human attention-based rate control techniques have been developed in the literature. In [1], the human visual system was taken into account by imposing constraints on the PSNR value of the face regions and the temporal delay time. Other researchers propose that, instead of the pixel-wise mean square error (MSE) measure, the perceptual distortion weighted measures should be used ([2], [3]). The work in [4] adopts an object tracking technique and a temporal filter to reduce the bits consumption of highly moving background without visual quality loss for scenes with a static face region and high movement background. However, such face-focused coding techniques cannot be applied in a broader sense for general video sequences. In [4], more bits are to the foreground satisfying some target visual quality while allowing the background quality gracefully to degrade as a function of the distance from the foreground.

One key concept of the proposed psychovisual model is that for video rate control, visual attention is not the most important cue for bit-allocation. Visual distortion sensitivity (VDS), namely the capability for human vision to detect distortion in moving scenes, is what a high quality video coder should be taking advantage of during the bit-allocation process. VDS is influenced by the motion structure as well as the texture structure of the scene. Moving objects with random textures in a video sequence, albeit attract human attentions in most cases, can tolerate high perceptual distortion introduced by the encoder. Therefore, they should consume much fewer bits than what a conventional MAD-based rate control algorithm will assign.

Without the complicated object segmentation and global motion (camera motion) estimation for visually significant object extraction, we propose an effective visual distortion sensitivity model to indicate the perceptually important regions. A motion attention model and a texture structure model are fused together for VDS analysis. Quantization parameters are then determined based on the analysis Bitrate saving is mainly achieved by allocating fewer bits to randomly textured moving regions in the video sequence. This bit-allocation mechanism can be incorporated into any rate control techniques of any existing video coding standards.

This paper is organized as follows. In section 2, the motion attention model presented in [4] is introduced. This model is adopted here due to its low complexity and reasonable performance. A new texture-structure model is developed in section 3. The proposed psychovisual mode combining the motion attention model and the texture-structure model is described in section 4. In this section, the bit-allocation mechanism is also proposed. Section 5 presents some experimental results based on JM7.6 of H.264 to show the

0-7803-8554-3/04/$20.00 ©2004 IEEE.

effectiveness of the proposed framework. Finally, the conclusions are given in section 6.

2. THE MOTION ATTENTION MODEL

Although the human visual model for still image has been well studied, the perceptual distortion metrics involving a more sophisticated psychophysical model are not fully understood yet. The visual model for video sequences is quite different from that for still images. People usually pay more attention to the foreground moving objects even the scene contains highly moving background due to camera motion.

The theory behind human attention has attracted great interests in the field of psychology, biology, neurophysiology and cognitive engineering in the past decades. William James, the father of American psychology, first came up with the idea of human attention theory [4]. The behavior of human attention consists of the top-down and the bottom-up processes. The top-down process is intentionally controlled by human brain to direct one's attention to objects in order to accomplish a task. One computational model simulating such process can be found in [4]. The bottom-up process is triggered unintentionally by certain objects in the surrounding environment and grabs our attentions. The bottom-up visual attention can be further classified into static attention and dynamic (motion) attention. In the proposed scheme, we employ the motion attention model developed in [4]. This model involves low computational complexity since it indicates the moving object without global motion estimation and object tracking.

This model is composed of the intensity inductor, spatial coherence inductor and the temporal coherence inductor. For a target frame with frame number *n*, the intensity inductor corresponding to the motion intensity for macroblock at location (i, j) is

$$I_{nij} = \sqrt{mvx_{nij}^2 + mvy_{nij}^2} / \max I, \qquad (1)$$

where (mvx_{nij}, mvy_{nij}) is the motion vector and $\max I$ is the maximal motion vector intensity in the target frame. Since camera motion could also cause large intensities, the other two inductors are developed to suppress such negative effect.

The spatial and temporal coherence inductors are based on the concept of the motion vector entropy. A large entropy value represents less coherence of the motion vectors. The spatial coherence inductor *Cs* is the spatial consistency of the directions of the motion vectors, and, it is

$$Cs_{nij} = -\sum_{b=1}^{n_s} ps_n(b) \log(ps_n(b)), \qquad (2)$$

where $ps_n(b)$ is the probability distribution function, and n_s is the number of histogram bins. For one macroblock, the histogram is generated from the motion vector directions within a spatial window of $w \times w$ macroblocks. Note that the regions belonging to the same moving object usually lead to small *Cs* value. However, for moving backgrounds, the *Cs* value is not always large while the intensity inductor is large.

The temporal consistency inductor is used to discriminate the camera motion from the object motion since the former is usually more stable than the latter during a longer period of time. This inductor is

$$Ct_{nij} = -\sum_{b=1}^{n_t} pt_n(b) \log(pt_n(b)), \qquad (3)$$

where $pt_n(b)$ is the probability distribution function, and n_t is the number of histogram bins. For each macroblock, the histogram is generated from the motion vector directions within a temporal window of *L* frames.

Finally, the motion attention index of macroblock at location (i, j) is

$$MI_{nij} = I_{nij} \times Ct_{nij} \times (1 - I_{nij} \times Cs_{nij}), \qquad (4)$$

All inductor values are between 0 and 1.

The motion attention model alone is not good enough to compute visual distortion sensitivity (VDS). Although a human typically pay more attention to the foreground objects, people are also sensitive to nicely-textured regions (static or moving globally) even in the background. While, on the other hand, more perceptual distortions are permitted in the randomly-textured moving regions (in the background). Therefore, in the next section, we further propose a texture structure model to aid the VDS analysis process.

3. THE TEXTURE STRUCTURE MODEL

For the regions with higher spatial contrast and weaker correlations of the intensities of the nearby image pixels, HVS tries to maximize the information transmitted to the early visual processing stages since such regions contain higher entropy. Itti et al. proposed a saliency-based model combining the topographic feature maps of local color contrast, intensity, and orientation as the computational visual attention model for still images [4]. In [4], a summarization scheme captures attended regions by tracking the color and orientations extracted from the first frame of each shot. However, for video coding purposes, we need a new model to discriminate the randomly-textured regions from the well-structured ones.

Intensity discontinuity (edge) is an important cue in the bottom-up early vision process. A randomly-textured region is typically composed of small edges in various orientations

while a nicely-textured region should be composed of consistent long edges. Although the randomly-textured regions carry more entropy than the nicely-textured regions, human visions are less sensitive to the distortions in the randomly-textured regions since these regions contain many random stimuli (small edges) that cover up coding noises. Thus, our proposed general texture structure algorithm is as follows. The mean and the edge density found by an edge detector are evaluated. The mean is computed by

$$M_{nij} = \sum_{u=0}^{BS-1}\sum_{v=0}^{BS-1} e_{nijuv} /(BS \times BS), \quad (5)$$

where *BS* is the macroblock size (16 in our experiments) and e_{nuvij} is the edge intensity value of pixel at location (u,v) on the macroblock at location (i,j). The maximal value of e_{nuvij} varies with different edge detectors. The edge density is computed by

$$D_{nij} = \sum_{u=0}^{BS-1}\sum_{v=0}^{BS-1} E_{nijuv} /(BS \times BS), \quad (6)$$

$$E_{nuvij} = \begin{cases} 1, & \text{if } e_{nuvij} > \alpha. \\ 0, & \text{otherwise.} \end{cases}, \quad (7)$$

where α is a threshold for judging the edge pixels, and E_{nuvij} indicates whether the pixel belongs to an edge or not. Finally, the texture structure index for the macroblock at location (i,j) is evaluated around $s \times s$ neighboring macroblocks, that is,

$$BI_{nij} = \sum_{k=i-(s-1)/2}^{i+(s-1)/2}\sum_{l=j-(s-1)/2}^{j+(s-1)/2} (M_{nkl} \times D_{nkl}) /(s \times s), \quad (8)$$

This index value is large in randomly-textured regions but small in nicely textured regions. For well-textured regions, the index values are much smaller than those in the randomly-textured regions.

One realization of the proposed algorithm employs both the Canny and Sobel edge detectors. The texture and smooth regions are classified based on the Canny maps by following (2) to (4) with scale *s*=1. For the textured regions indicated by the Canny maps, we further distinguish the randomly-textured regions from the nicely-textured regions with Sobel maps by following the same process (2) to (4) (*s*=1).

4. THE PROPOSED BIT ALLOCATION SCHEME

Our proposed bit allocation scheme applies the original bit allocation process to the motion attended regions while fewer bits are assigned to other regions for bit savings. First, the human visual attention analysis fuses the motion attention index with the texture structure index. Let the maximal value of visual distortion sensitivity index (VDSI) be $VDSI_{\max}$, which is scaled to 255 without loss of generality. We first map the motion attention index in (4) to a 2-level value by

$$MI'_{nij} = \begin{cases} VDSI_{\max}, & \text{if } MI_{nij} > \gamma. \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

where γ is a threshold for indicating visual attention regions and $0 \le \gamma \le 1$. Then, we map the texture masking index BI_{nij} to a modified index BI'_{nij} by

$$BI'_{nij} = \begin{cases} (V_2 \times \log_2 BI_{nij} / \log_2 \beta_2) + V_1, & \text{if } \beta_2 \le BI_{nij} \le \beta_1. \\ V_3 \times 2^{-(BI_{nij}-\beta_1)} + V_4, & \text{if } BI_{nij} > \beta_1. \\ V_1, & \text{otherwise.} \end{cases} \quad (10)$$

where β_1 and β_2 are thresholds for indicating textured regions. In the experiments, we set $V_1 = VDSI_{\max}/2$, $V_3 = VDSI_{\max}/8$, V_2 and V_4 all equal to $VDSI_{\max}/4$. Finally, the VDSI is

$$VDSI_{nij} = \begin{cases} MI'_{nij}, & \text{if } MI'_{nij} = VDSI_{\max}. \\ BI'_{nij}, & \text{otherwise.} \end{cases} \quad (11)$$

Accordingly, the larger VDSI value corresponds to the region permitting smaller perceptual distortions.

Examples of the VDSI maps in Figs 1(a) and (b) show that the players are more sensitive to perceptual distortions in STEFAN and FOOTBALL. On the other hand, the audience regions in STEFAN are assigned small index value since these areas can tolerate large perceptual distortions.

In a video encoder, bit budget is allocated to different regions based on VDSI. As a practical example, the quantization step size computed by any rate control algorithm can be adaptively adjusted by adopting the VDSI as follows,

$$QP'_{nij} = QP_{nij} + (1 - VDSI_{nij}/255) \times \Delta q, \quad (12)$$

where QP_{nij} is the initial quantization parameter assigned by the rate control algorithm, and Δq is a parameter for limiting the modification of QP_{nij}.

5. EXPERIMENTAL RESULTS

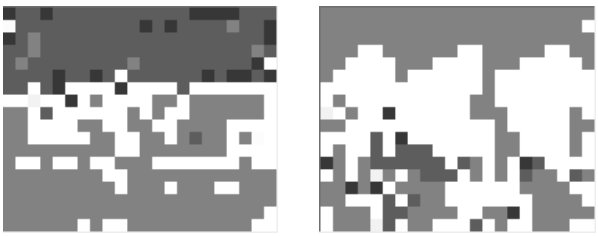
We use JM 7.6 of H.264 to conduct the experiments and the configuration is as follows. The Hadamard transform, CABAC, and reconstruction filter are enabled. No B frame is inserted. The encoded sequences are the CIF versions of STEFAN and FOOTBALL at 30 fps. The parameter settings of the proposed scheme are: $w = 5$, $L = 9$, $n_s = 16$, $n_t = 16$, $\alpha = 50$, $\beta_1 = 60$, $\beta_2 = 15$, and $\gamma = 0.4$. The motion vectors used in the attention model are generated with the full search motion estimation module in the JM 7.3 of H.264 with RDO turned on. Constant QP, i.e. no rate control, is used to demonstrate

the coding efficiency gain purely from the proposed psycho visual model.

Table 1 shows the bit reduction of the proposed mechanism (without losing visual quality) for STEFAN and FOOTBALL. For STEFAN with an initial QP=22 given by the JM 7.6, bit rate reduction is 41.08%. There is no visible difference observed even though the overall PSNR is decreased by 4.5dB. For example, Figs. 2(a) and (b) compare the visual qualities of the reconstructed frames between the constant-quality (PSNR-wise) and the proposed bit allocation scheme. The PSNR loss is around 5.66 dB. This is because fewer bits are allocated to the visually less sensitivity regions (e.g., the audience) while more bits are assigned to the distortion sensitive regions (e.g., the tennis player and the text on the fence). It turns out that there is no perceptual quality degradation even though the PSNR of the whole video sequence decreases a lot. These experiments also show strongly the inadequacy of PSNR as a distortion measure.

6. CONCLUSIONS

In this paper, we propose a novel video coder bit allocation technique based on a visual distortion sensitivity analysis. This analysis directs the video coder assigning fewer bits to regions less noticed, and accordingly, the bit-rate saving is achieved. Our future work will focus on extending and refining the proposed model.



(a) (b)

Figure 1. VDSI maps for (a) 26th frame in STEFAN and (b) 6th frame in FOOTBALL.

(a) (b)

Figure 2. The Comparison of the 66th frames of STEFAN with difference bit allocation techniques when QP = 22: (a) Constant quality (fixed QP) bit allocation (PSNR= 39.85 dB) and (b) Psychovisual model-based bit allocation (PSNR=34.19dB).

Table 1. The comparisons of coding performance with different bit allocation techniques for STEFAN (the first 3 rows) and FOOTBALL (the last 3 rows). Note that even though there is a large PSNR decrease in the proposed method, there is hardly any visible distortion increase compared to a conventional coder.

	H.264 Constant QP		H.264 with VDSI Analysis		
Initial QP	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	Coding Efficiency gain
22	3250	39.87	1915	35.36	41.08%
28	1408	34.99	928	31.95	34.09%
32	741	31.71	558	29.59	24.70%
22	2690	40.47	2339	39.06	13.05%
28	1356	36.08	1223	35.22	9.81%
32	830	33.32	768	32.76	7.47%

7. REFERENCES

[1] K. C. Lai, S. C. Wong and D. Lun, "A Rate Control Algorithm Using Human Visual System for Video Conferencing Systems," in *Proc. International Conference on Signal Processing*, vol. 1, pp. 656-659, August, 2002.

[2] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated Video Compression with Optimal Rate Control," *IEEE Trans. Image Processing*, vol. 10, no. 7, July 2001.

[3] C.-W. Wong, O. C. Au, B. Meng and H.-K. Lam, "Perceptual Rate Control for Low-Delay Video Communications," in *Proc. International Conference on Multimedia and Expo*, vol. 3, pp. 361-364, July, 2003.

[4] T. Adiono, T. Isshiki, K. Ito, T. Ohtsuka, D. Li, C. Honsawek and H. Kunieda, "Face Focus Coding under H.263+ Video Coding Standard," in *Proc. International Conference on Asia-Pacific Circuits and Systems,* pp. 461-464, December, 2000.

[5] S. Sengupta, S. K. Gupta and J. M. Hannah, "Perceptually Motivated Bit-Allocation for H.264 Encoded Video Sequences," *Proc. International Conference on Image Processing*, vol. 3, pp. 797-799, 2003.

[6] Y.-F. Ma, and H.-J. Zhang, "A Model of Motion Attention for Video Skimming," in *Proc. International Conference on Image Processing*, vol. 1, pp. I-129-132, September, 2002.

[7] W. James, *The Principles of Psychology*, Harvard University Press, Cambridge, Massachusetts, 1980.

[8] J. Han, M. Li, H. Zhang and L. Guo, "Automatic Attention Object Extraction from Images," *Proc. International Conference on Image Processing*, vol. 2, pp. 403-406, 2003.

[9] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, November, 1998.

[10] Y. Li, Y.-F. Ma and H.-J. Zhang, "Salient Region Detection and Tracking in Video," *Proc. International Conference on Multimedia and Expo*, vol. 2, pp. 269-272, 2003.