Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Perceptual importance analysis for H.264/AVC bit allocation*

Gui-xu LIN†[1,2], Shi-bao ZHENG[1,2]

([1]*Institute of Image Comm. & Info. Processing, Dept. of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

([2]*Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiao Tong University, Shanghai 200240, China*)

†E-mail: guixu_lin@sjtu.edu.cn

Received July 1, 2007; revision accepted Sept. 26, 2007; published online Dec. 20, 2007

Abstract: The existing H.264/AVC rate control schemes rarely include the perceptual considerations. As a result, the improvements in visual quality are hardly comparable to those in peak signal-to-noise ratio (PSNR). In this paper, we propose a perceptual importance analysis scheme to accurately abstract the spatial and temporal perceptual characteristics of video contents. Then we perform bit allocation at macroblock (MB) level by adopting a perceptual mode decision scheme, which adaptively updates the Lagrangian multiplier for mode decision according to the perceptual importance of each MB. Simulation results show that the proposed scheme can efficiently reduce bit rates without visual quality degradation.

Key words: Visual quality, Human visual system, Rate control, H.264/AVC
doi:10.1631/jzus.A071355 **Document code:** A **CLC number:** TN919.8

INTRODUCTION

H.264/AVC is gradually applied in the broad areas of video applications for its high encoding efficiency, including broadcast television, video telephony and video streaming over Internet, etc. Since the channel bandwidth and buffer size are limited in most of the applications, rate control is employed to achieve the best visual quality under certain rate constraints. Traditional rate control schemes mainly focus on peak signal-to-noise ratio (PSNR) performance improvement (Sun *et al*., 1997; Sullivan *et al*., 2003; Sarshar and Wu, 2005; Cherniavsky *et al*., 2007). However, PSNR is not a good visual metric for image. It is well known that large gain in PSNR may not always result in comparably large improvement in visual quality (Wong *et al*., 2003). Since the ultimate video quality is judged by human, a well-designed video encoder should be optimized in terms of human visual system (HVS) characteristics.

The improvements of visual quality mainly depend on two factors: the accuracy of the HVS models for video sequences and the way to incorporate the models into video coding frameworks. In this paper, we develop a fast but effective block-based perceptual importance analysis algorithm to abstract the HVS characteristics of video contents. The outcome is used to allocate bits in a frame by adopting a perceptual mode selection scheme, which adaptively updates the Lagrangian multiplier in rate-distortion optimization (RDO) process according to the perceptual importance of each MB. The experimental results show that our method can efficiently reduce bit rates without visual quality degradation.

In the following sections, we will first review the RDO technique in H.264, and then describe our perceptual importance analysis algorithm and its application to mode decision. Finally, the experimental results are discussed.

RATE-DISTORTION OPTIMIZATION IN H.264/AVC

RDO is an important technique to achieve the best quality under certain rate constraints (Wiegand *et al*., 2003). In H.264, RDO is employed for mode

* Project supported by the Shanghai Key Laboratory of Digital Media Processing and Transmissions, China

decision. To select the best coding mode for an MB, which means an optimal trade-off between bit rate and coding distortion, the Lagrangian optimization method is adopted. It is equivalent to selecting the mode with the minimal RD cost

$$J_{\mathrm{MODE}}=D_{\mathrm{REC}}+\lambda_{\mathrm{MODE}}\cdot R_{\mathrm{REC}}, \quad (1)$$

where D_{REC} is the distortion between the original MB and its reconstruction, R_{REC} is the resulting bits, λ_{MODE} is the Lagrange multiplier. RDO is also adopted in motion estimation for motion vector selection:

$$J_{\mathrm{MOTION}}=D_{\mathrm{DFD}}+\lambda_{\mathrm{MOTION}}\cdot R_{\mathrm{MOTION}}, \quad (2)$$

where D_{DFD} is the motion compensation prediction error, R_{MOTION} is the encoding bits for motion vector, $\lambda_{\mathrm{MOTION}}$ is the Lagrange multiplier. In (Sullivan *et al*., 2003), λ_{MODE} and $\lambda_{\mathrm{MOTION}}$ are calculated using a pre-decided quantization parameter (QP):

$$\lambda_{\mathrm{MODE}}=0.85\times 2^{(QP-12)/3}, \quad (3)$$

$$\lambda_{\mathrm{MOTION}}=\sqrt{\lambda_{\mathrm{MODE}}}\ . \quad (4)$$

Different λ_{MODE} values can result in different optimum modes. The small λ_{MODE} values correspond to the modes with high bit rates and low distortions, and vice versa (Wiegand *et al*., 2003). $\lambda_{\mathrm{MOTION}}$ has similar effect for motion estimation. Hence the choice of Lagrange multiplier can influence the rate-distortion performance. There are some literature related to this issue. In (Jiang and Ling, 2006), λ_{MODE} is adjusted according to actual MB coding bits to meet the target frame bits. In (Zhang *et al*., 2006), context adaptive Lagrange multiplier (CALM) for motion estimation is presented, which adjusts $\lambda_{\mathrm{MOTION}}$ for each block based on the Lagrange cost of its neighboring or upper layer blocks. It has been included in the new version of JM (JVT, http://iphome.hhi.de/suehring/tml/download/jm12.2.zip). These methods can achieve considerable PSNR gain. However, the perceptual considerations are ignored and the choice of λ_{MODE} is not perceptually optimal. In (Tsai *et al*., 2004; Yu *et al*., 2005), λ_{MODE} is adjusted based on the perceptual characteristics of video contents. The schemes can reduce bit rates to some extent with virtually the same visual quality. The coding performance depends largely on the good estimation of HVS features. However, in these methods, only one of the spatiotemporal masking effects of HVS is exploited. For video coding, both the spatial and temporal masking effects of HVS should be considered. Thus a more accurate HVS model is needed.

PERCEPTUAL IMPORTANCE ANALYSIS SCHEME

In this section, we present a perceptual importance analysis scheme for video contents. A region with high perceptual importance can heavily influence our perception of overall picture quality. Distortions in these areas are more noticeable. It should be noted that the perceptual importance is not totally equivalent to visual attention in a video sequence. A region with high visual attention is not always perceptually important. For example, moving objects are likely to attract visual attention. However, if motions occur in random texture regions, HVS is hardly noticed in the distortion. The moving objects are perceptually less important in this case.

The perceptual importance analysis scheme includes a moving region detector and a texture analysis model.

Moving region detector

As mentioned above, HVS is more sensitive to moving objects. Moving objects in video sequences tend to arrest attention and to be of interest (Park and Kim, 2006). For some video applications such as visual surveillance and video telephony, the moving regions often correspond to humans or vehicles, which are always the observing focus. Hence the perceptual qualities of moving regions are very important to the overall picture quality. For an MB-based video encoder, we simply classify the MBs in a frame as moving MB or static MB.

We first pass the target frames through a low-pass filter to remove the high frequency noise, which is a simple 3×3 averaging filter with uniform weights of 1/9 in our experiments. The error image of two successive frames is then given by

$$D_n(i,j)=I_n(i,j)-I_{n-1}(i,j), \quad (5)$$

where $I_n(i, j)$ and $I_{n-1}(i, j)$ respectively represent the luminance pixels at location (i, j) in current and previous frames. Let f_n denote frame n, $B_n(X, Y)$ denote the MB at location (X, Y) in f_n. The difference value of $B_n(X, Y)$ is calculated by

$$MD_n(X,Y)=\sum_{(i,j)\in B_n(X,Y)} D_n(i,j). \tag{6}$$

$B_n(X, Y)$ is classified as a moving MB if $MD_n(X, Y)$ is larger than the threshold T_n. Otherwise, $B_n(X, Y)$ is a static MB. T_n is defined as

$$T_n = t\cdot\frac{1}{N}\sum_{B_n(X,Y)\in f_n} MD_n(X,Y). \tag{7}$$

Here N is the number of MBs in a frame, t is a weighting factor with typical value 1.2 in our experiments. Finally, for the integrity of segmented regions, the isolated MBs are merged. An isolated moving (static) MB is the MB whose neighboring MBs are all static (moving) MBs. An isolated moving (static) MB $B_n(X, Y)$ is classified as a static (moving) MB if its difference value $MD_n(X, Y)$ is in the last (first) 30% of the list sorted for all moving (static) MBs in f_n.

Texture analysis model

The texture analysis model divides a picture into three types of regions: smooth, random texture and structure texture. According to HVS, human observers are more likely to be attracted by texture regions containing high spatial contrasts (Ma and Zhang, 2003). Meanwhile, distortions are less noticeable in the texture regions that contain a high number of edges, which are usually small and directionally random. We call it random textured region. The other is called structure texture region, which usually contains the borders of objects.

First, the texture regions are abstracted. Both the average edge intensity and the edge pixel density in an MB are considered. We adopt the Sobel operator to obtain the edge image. The edge intensity of $B_n(X, Y)$ is then calculated by

$$MI_n(X,Y)=\sum_{(i,j)\in B_n(X,Y)} EI_n(i,j), \tag{8}$$

where $EI_n(i, j)$ represents the intensity value of the pixel at location (i, j) in edge image n. The edge pixel density of $B_n(X, Y)$ is calculated by

$$MED_n(X,Y)=\sum_{(i,j)\in B_n(X,Y)} EP_n(i,j), \tag{9}$$

$$EP_n(i,j)=\begin{cases}1, & EI_n(i,j)>\alpha,\\ 0, & \text{otherwise.}\end{cases} \tag{10}$$

Here α is a threshold for edge pixel selection and its typical value is 50 experimentally. When $EI_n(i, j)$ is larger than α, it indicates an edge pixel. The threshold for edge intensity and the threshold for edge pixel density are set by

$$T_{\mathrm{EI}} = \mu\cdot\sum_{B_n(X,Y)\in f_n} MI_n(X,Y)/N, \tag{11}$$

$$T_{\mathrm{MED}} = \rho\cdot\sum_{B_n(X,Y)\in f_n} MED_n(X,Y)/N, \tag{12}$$

where μ and ρ are weighting factors and their typical values are 0.6 and 1.0 respectively in our experiments. An MB with edge intensity larger than T_{EI} or edge pixel density larger than T_{MED} is classified as a texture MB. Otherwise, it belongs to smooth regions.

Since random texture regions usually contain a large number of small edges with random directions, it is reasonable to think the edge intensities in these regions distribute more evenly than those in structure texture regions. We divide the 16×16 MB into small 4×4 blocks and calculate the normalized deviation of each block as

$$Md_n(X,Y)=\sum_{b_n(x,y)\in B_n(X,Y)}\frac{\left|Eb_n(x,y)-mEb_n(X,Y)\right|}{mEb_n(X,Y)}, \tag{13}$$

$$Eb_n(x,y)=\sum_{(i,j)\in b_n(x,y)} EI_n(i,j), \tag{14}$$

$$mEb_n(X,Y)=MI_n(X,Y)/N_{\mathrm{block}}, \tag{15}$$

where $b_n(x, y)$ denotes the block at location (x, y) of f_n, $Eb_n(x, y)$ the edge intensity of $b_n(x, y)$, $mEb_n(X, Y)$ the mean block edge intensity of $B_n(X, Y)$, N_{block} the number of 4×4 blocks in an MB. Let $N_n(X, Y)$ denote the s×s blocks neighboring $B_n(X, Y)$. We calculate the normalized deviation of $N_n(X, Y)$ by

$$Nd_n(X,Y)=\sum_{b_n(x,y)\in N_n(X,Y)}\frac{\left|Eb_n(x,y)-mNb_n(X,Y)\right|}{mNb_n(X,Y)}, \quad (16)$$

$$mNb_n(X,Y)=\frac{1}{4s^2+16s}\sum_{b_n(x,y)\in N_n(X,Y)}Eb_n(x,y). \quad (17)$$

For the MB $B_n(X, Y)$, if both $Md_n(X, Y)$ and $Nd_n(X, Y)$ are under the mean level of the texture MBs in f_n, it is classified as a random texture MB. Otherwise, it belongs to structure texture regions.

MB perceptual importance map

Based on the moving region detector and texture analysis model, the perceptual importance of MBs is assigned as shown in Table 1. "4" is the highest level of perceptual importance and "1" is the lowest level. For Moving MBs, on the one hand, the visual distortion sensitivities decrease with the increase of object motion velocities; on the other hand, the loss of visual sensitivities may be compensated by eye movement, which reduces object motion velocities from the image plane velocities to the retinal velocities (Jia *et al*., 2006; Tang, 2007). In this paper, we focus on video contents containing slow motions and assume that the slowly moving regions of smooth/structure texture gain more visual importance than that of static regions, since the moving objects are more likely to arrest attention and to be of interest. And, the slowly moving regions with random texture are assigned to relatively low level, since distortions in these regions are less noticeable than those in other moving regions.

Table 1 The perceptual importance levels

Temporal characteristics	Spatial characteristics	Perceptual importance level
Static	Random texture	1
	Structure texture	3
	Smooth	2
Moving	Random texture	2
	Structure texture / Smooth	4

We also take temporal consistency into consideration. To avoid large unbalance of bit allocation to the same object in successive frames, which can probably cause visible quality difference and thus flickering artifacts, the assigned perceptual importance of an MB is further adjusted. Let $\Delta\rho$ denote the maximal difference of perceptual importance between current MB and its reference MB. We set $\Delta\rho$ to 1 in our experiments.

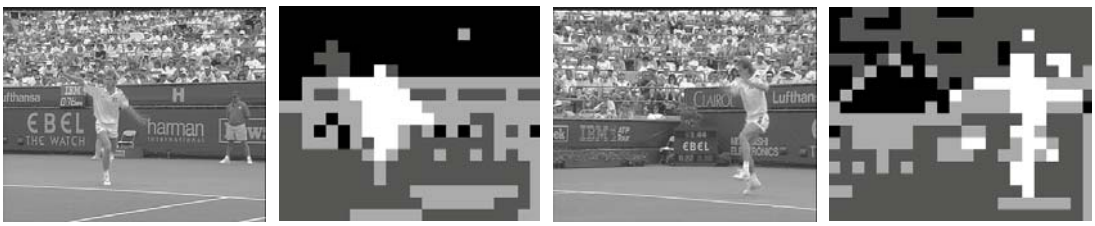
Fig.1 gives an example of MB perceptual importance map for sequence "Stefan". The lighter regions represent areas of higher perceptual importance. It is observed that the moving structure texture regions (the player) can be successfully abstracted from the random texture regions (the audience), which are static in the 28th frame and moving in the 107th frame.

PERCEPTUAL MODE DECISION SCHEME

In our method, the Lagrangian multiplier for mode decision, which is calculated by Eq.(3), is adaptively adjusted based on the perceptual importance level of MBs. It can be expressed as

$$\lambda_{\text{MODE}}=\lambda_{\text{MODE}}\cdot\eta_t, \quad (18)$$

where η_t is a weighting factor related to perceptual importance level t (t=1, 2, 3, 4). The value of η_t is decreased with the increase of perceptual importance level. In our experiments, η_1, η_2, η_3 and η_4 were 4.0, 2.0, 1.0 and 0.7, respectively. That is, high perceptual importance level corresponds to small η_t and consequently to small λ_{MODE}, so that the mode emphasizing



(a) (b) (c) (d)

Fig.1 Perceptual importance maps for the Stefan sequence. The 28th frame: (a) Original picture; (b) The perceptual importance map. The 107th frame: (c) Original picture; (d) The perceptual importance map

low distortion and high bits can be selected, and vice versa. Hence, we can allocate more bits to high perceptual importance regions for visual quality improvement and fewer bits to low perceptual importance regions for bit rate saving.

EXPERIMENTAL RESULTS

The proposed scheme has been implemented on JM 12.2. Its performance was compared with those of JM and the method proposed in (Yu *et al*., 2005), which is presented as Yu in the rest of this paper. To demonstrate the effects of perceptual importance analysis model for coding efficiency, we encoded the test sequences with constant QPs. All sequences were 150 frames with CIF 4:2:0 format. No B frame was inserted. CALM was enabled to select the method in (Zhang *et al*., 2006). RDO was on (high complexity mode). Hadamard and CABAC were also used. The number of reference frames for motion estimation was 2 and the search range was 16.

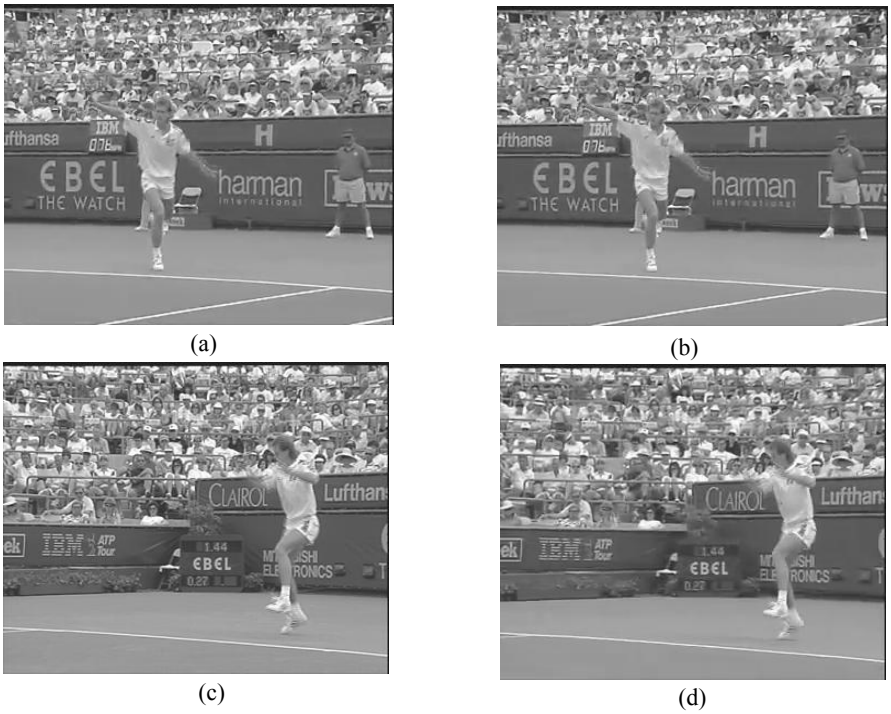
For subjective quality valuation, we adopted the Double Stimulus Continuous Quality Scale (DSCQS) method (ITU-R, 2000). Ten observers (mainly students with image processing knowledge) were asked to view the video under normalized conditions, and vote using a scale with 5 grades: excellent (5), good (4), fair (3), poor (2) and bad (1). Mean opinion scores (MOS) were calculated from raw scores of all observers. The results can be found in Table 2. It is observed that the subjective scores of our method are almost the same as those of Yu and JM 12.2 with constant QPs. In the case of low bit rates, our method even gained slightly higher scores.

The objective quality metrics is PSNR. From Table 2, we can see that both Yu and our scheme result in PSNR degradations. As compared to JM, the PSNR degradation of our scheme is more than 1 dB. This is understandable because fewer bits are allocated to the less perceptual importance regions, e.g. the audience of "Stefan". It leads to no visual quality degradation since distortions in those regions are less noticeable. As an example, Fig.2 shows some reconstructed frames of our scheme and JM. It can be found that the reconstructed frames are hardly distinguished in terms of visual quality, even though there is a PSNR difference of 1.82 dB for the 28th frame and 1.74 dB for the 107th frame. Meanwhile, our scheme achieves a bit rate reduction of 24.21% and 21.89% respectively. The overall bit rate is decreased by about 10%, while there is only about 2% bit rate reduction with Yu.

We also compared the coding performance with rate control enabled. It was observed that our scheme still caused some PSNR degradations. However, the visual quality was slightly better, especially in the case of low bit rates. Table 3 shows the coding results of three test sequences under low target bit rates. The frame rate is 30 fbps for all the sequences. It can be found the MOS scores of our scheme are obviously higher than those of JM and Yu. Fig.3 gives the reconstructions of the 107th frame of "Stefan". The fully reconstructed frames of our scheme and JM12.2 are shown in Fig.3a and Fig.3b, respectively. To demonstrate the subjective improvement more clearly, Fig.3c and Fig.3d give the 200% enlarged part of the player in the 107th frame, which is of high perceptual importance. It is found that, in our reconstructed frame, the legs and the shoes are clearer than those in

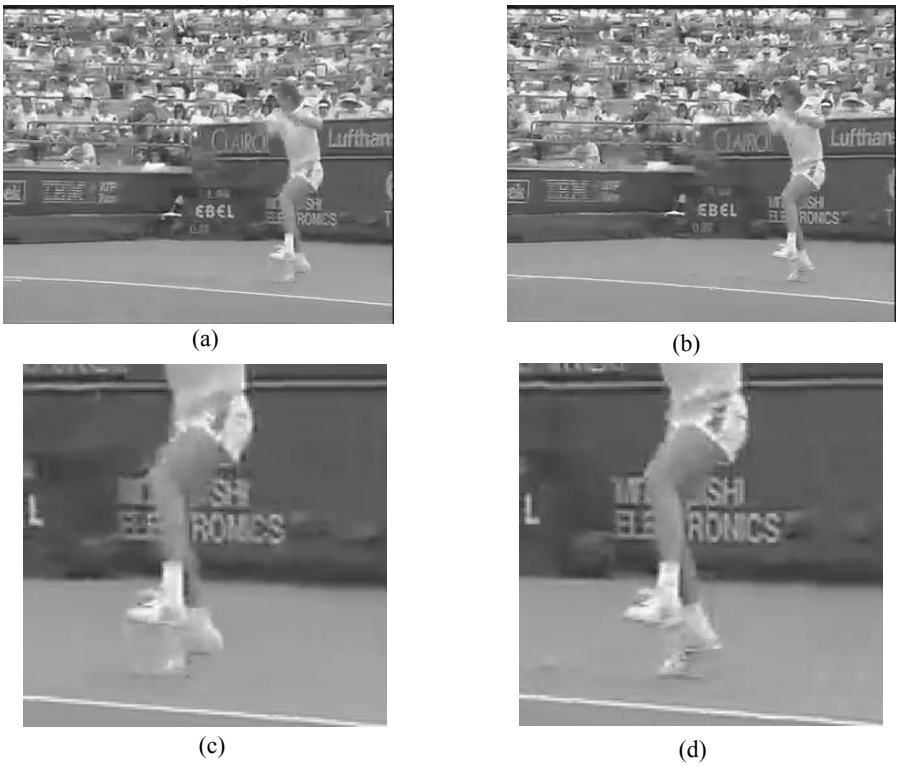
Table 2 Coding performance comparisons under constant quantization parameters (QPs)

Sequence	QP	MOS			PSNR (dB)			Rate (kbps)			
		JM	Yu	Ours	JM	Yu	Ours	JM	Yu	Ours	Saving (%)
Coastguard	28	4.1	4.1	4.1	35.49	35.01	33.87	1300	1274	1156	11.1
	32	3.8	3.8	3.9	32.13	31.67	30.87	622	601	535	14.0
	36	3.4	3.5	3.5	29.23	28.92	28.38	262	253	224	14.5
Mobile	28	4.2	4.3	4.3	35.27	34.98	33.95	2097	2052	1921	8.4
	32	3.9	3.9	3.9	31.35	31.03	30.29	998	975	904	9.4
	36	3.5	3.5	3.6	27.97	27.58	27.16	412	403	381	7.5
Stefan	28	4.0	4.0	4.0	36.15	35.65	34.53	1390	1362	1221	12.2
	32	3.7	3.7	3.8	32.50	31.99	31.16	696	675	607	12.8
	36	3.3	3.4	3.5	29.27	28.93	28.27	323	318	288	10.8



(a) (b) (c) (d)

Fig.2 Experimental results of sequence "Stefan" with *QP*=32. The 28th frame: (a) Reconstructed frame of JM 12.2: *PSNR*=31.59 dB, 16488 bits/pic; (b) Reconstructed frame of the proposed scheme: *PSNR*=29.77 dB, 12496 bits/pic. The 107th frame: (c) Reconstructed frame of JM 12.2: *PSNR*=32.01 dB, 28432 bits/pic; (d) Reconstructed frame of the proposed scheme: *PSNR*=30.27 dB, 22208 bits/pic



(a) (b) (c) (d)

Fig.3 Reconstructions of the 107th frame of "Stefan" with rate control enabled: bit rate=323 kbps, initial *QP*=36, frame rate=30 fbps. (a) Reconstructed frame of JM 12.2, *PSNR*=25.63 dB; (b) Reconstructed frame of our scheme, *PSNR*=24.80 dB; (c) 200% enlarged local area of (a); (d) 200% enlarged local area of (b)

Table 3 Coding performance comparisons with rate control enabled

Sequence	Rate (kbps)	PSNR (dB)			MOS		
		JM	Yu	Ours	JM	Yu	Ours
Coast-guard	262	29.19	29.10	29.01	3.1	3.2	3.4
Mobile	412	27.89	27.72	27.36	3.2	3.3	3.5
Stefan	323	29.15	28.95	28.51	3.1	3.2	3.3

JM12.2 reconstruction. This indicates our scheme can enhance the visual quality with rate control enabled. In addition, besides perceptual mode decision, the perceptual importance analysis scheme can be combined with other MB and frame layer rate control algorithms to further improve the performance.

CONCLUSION

In this paper, we present an MB-based perceptual importance analysis scheme. MB level perceptual bit allocation is then realized by adaptively updating the Lagrangian multiplier for mode decision based on the perceptual importance of MBs. Simulation results show our scheme can efficiently reduce bit rates without visual quality degradation.

References

Cherniavsky, N., Shavit, G., Ringenburg, M.F., Ladner, R.E., Riskin, E.A., 2007. MultiStage: a MINMAX bit allocation algorithm for video coders. *IEEE Trans. on Circuits Syst. Video Technol.*, **17**(1):59-67. [doi:10.1109/TCSVT.2006.887135]

ITU-R, 2000. Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU-R Recommendation BT.500-10.

Jia, Y., Lin, W., Kassim, A.A., 2006. Estimating just-noticeable distortion for video. *IEEE Trans. on Circuits Syst. Video Technol.*, **16**(7):820-829. [doi:10.1109/TCSVT.2006.877397]

Jiang, M., Ling, N., 2006. On Lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control. *IEEE Trans. on Circuits Syst. Video Technol.*, **16**(5):663-669. [doi:10.1109/TCSVT.2006.873159]

Ma, Y.F., Zhang, H.J., 2003. Contrast-based Image Attention Analysis by Using Fuzzy Growing. Proc. ACM Int. Multimedia Conf. and Exhibition, p.374-381. [doi:10.1145/957013.957094]

Park, S., Kim, M., 2006. Extracting moving/static objects of interest in video. *LNCS*, **4261**:722-729. [doi:10.1007/11922162_83]

Sarshar, N., Wu, X.L., 2005. Optimal Channel Rate Allocation for Multimedia Communication Over Fading Wireless Channels. Proc. 1st ACM Workshop on Wireless Multimedia Networking and Performance Modeling. Montreal, p.8-15. [doi:10.1145/1089737.1089740]

Sullivan, G., Wiegand, T., Lim, K.P., 2003. Joint Model Reference Encoding Methods and Decoding Concealment Methods. Section 2.6: Rate Control. JVT-I049, San Diego.

Sun, H.F., Kwok, W., Chien, M., Ju, C.H.J., 1997. MPEG coding performance improvement by jointly optimizing coding mode decisions and rate control. *IEEE Trans. on Circuits Syst. Video Technol.*, **7**(3):449-458. [doi:10.1109/76.585924]

Tang, C.W., 2007. Spatiotemporal visual considerations for video coding. *IEEE Trans. on Multimedia*, **9**(2):231-238. [doi:10.1109/TMM.2006.886328]

Tsai, C.J., Tang, C.W., Chen, C.H., Yu, Y.H., 2004. Adaptive Rate-Distortion Optimization using Perceptual Hints. IEEE Int. Conf. on Multimedia and Expo. Taipei, p.667-670.

Wiegand, T., Schwarz, H., Joch, A., Kossentini, F., Sullivan, G.J., 2003. Rate-constrained coder control and comparison of video coding standards. *IEEE Trans. on Circuits Syst. Video Technol.*, **13**(7):688-703. [doi:10.1109/TCSVT.2003.815168]

Wong, C.W., Au, O.C., Meng, B., Lam, H.K., 2003. Perceptual Rate Control for Low-delay Video Communications. IEEE Int. Conf. on Multimedia and Expo, p.361-364.

Yu, H.T., Pan, F., Lin, Z.P., Sun, Y., 2005. A Perceptual Bit Allocation Scheme for H.264. IEEE Int. Conf. on Multimedia and Expo. Amsterdam, p.4-7. [doi:10.1109/ICME.2005.1521423]

Zhang, J., Yi, X., Ling, N., Shang, W., 2006. Context Adaptive Lagrange Multiplier (CALM) of Motion Estimation in JM - Improvement. JVT-T046, Klagenfurt.