VISUAL PERCEPTION BASED LAGRANGIAN RATE DISTORTION OPTIMIZATION FOR VIDEO CODING

Xi Wang[1,2], Li Su[1], Qingming Huang[1,2], Chunxi Liu[1]

[1]Graduate University of Chinese Academy of Sciences, Beijing, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing, China
{xwang,lsu,qmhuang,cxliu}@jdl.ac.cn

ABSTRACT

In the conventional rate distortion optimization (RDO) video coding, the measure of distortion is mainly from the perspective of signal processing, while dose not fully take into account the characteristics of visual perception. People have concerns about not only the information of independent pixels, but also the temporal and spatial correlations between them. For different video content, human visual perception has different sensitivity. In this paper, in order to establish a RDO model which is more consistent with the human visual perception, we introduce the structural similarity and the content saliency information into the distortion metric. An adaptive Lagrange multiplier selection scheme is presented to allocate the bit resources more rationally by keeping the balance of the bit-rate and the visual quality. Experimental results show that the proposed method averagely reduces 10.14% bit-rate under the similar visual quality.

Index Terms— Structural similarity, Selective attention, Rate-distortion optimization

1. INTRODUCTION

The hybrid video coding framework provides various modes for selection, such as {Intra16x16, Intra8x8, Intra4x4, inter16x16, inter16x8, inter8x16, inter8x8, inter8x4, inter4x8, inter4x4, SKIP, DIRECT}. How to choose the best mode is crucial for the encoding process. Actually the mode selection is an optimization problem to minimize the distortion D for a given rate R_c, which can be defined as:

$$min\{D\}\ subject\ to\ R \leq R_c \quad (1)$$

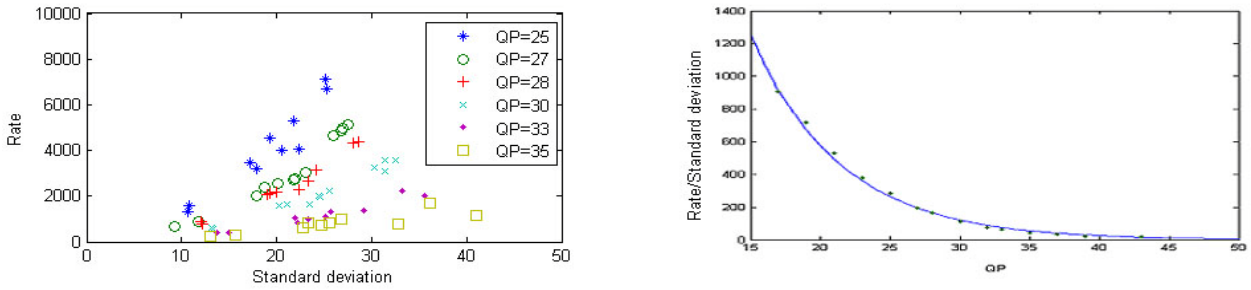
where R and D indicate rate and distortion, respectively. The above constrained optimization problem can be converted to an unconstrained form (2) by the Lagrange multiplier method, that is:

$$min\{J\}, where\ J = D + \lambda R \quad (2)$$

here J is the Lagrange cost function and λ is the Lagrange multiplier. In general, λ is determined by experimental results and typical rate-distortion models [1].

In the current rate distortion optimization (RDO) scheme, the sum of absolute difference (SAD) or sum of square difference (SSD) is used as the distortion metric to measure the video quality. However, these statistical error based metrics do not work very well as human visual distortion metrics. The defects in these distortion metrics make the coding algorithm very difficult to get the most optimized encoding effectiveness. The structural similarity index (SSIM) proposed by Wang [2] is a new theory to measure the distortion of signals from the perspective of visual perception. The examples in [3] show that under the same MSE distortion, the quality of the demo images differs greatly according to human perception. By incorporating the characteristics of human visual system such as human selective attention, video compression algorithms may achieve the equal visual quality with less bit-rate consumption. It has great significance in improving the efficiency of video compression. Especially, in the low bit-rate applications, in order to improve the subjective visual quality, algorithms could allocate the main bit-rate resources to the regions which having a higher distortion sensitivity. Therefore, establishing a perception based system is meaningful.

Many methods have been proposed to develop the rate-distortion (R-D) model [4]-[8]. The method in [4] describes an adaptive RDO model by using Laplace distribution of transformed residuals, but the use of human visual perception is insufficient. An experimental relationship between the SSIM index and the bit-rate is developed in [5]. However, the value range of the SSIM index is in [0,1] and a static Lagrange multiplier can't satisfy all the sequences with huge different bit-rate. Therefore, an adaptive Lagrange multiplier is very important to the RDO scheme which uses the SSIM index as the distortion metric. Methods in [6][7] both utilize the perception based adaptive rate-distortion optimization scheme. The algorithm in [6] obtains the dynamic Lagrange multiplier by approximating the gradient descent of encoded frames' R-D points. The method in [7] classifies the blocks of a frame into smooth ones and non-smooth ones and calculates the percentage of skipped mode macroblocks. However, neither of them takes the human attention parameters into account.



(a) (b)

Fig. 1 Experimental results of R-D model. (a) P frames in baseline profile: first 50 frames (I+49P) in IPPP structure. Each point indicates an average of 49 P frames for one test sequence. 10 sequences are tested. (b) The relationship between QP and R/σ_{sd}

In this paper, in order to make the measure of visual quality more consistent with the human visual system, the structural similarity and the content saliency information and human attention are introduced into the distortion metric. In order to cope with the different input sequences, the standard deviation of transformed residuals is used to establish an adaptive Lagrange multiplier selection scheme. In this way, we obtain the adaptive visual perceptual based RDO model.

The rest of this paper is organized as follow. Firstly, a plain relationship between the SSIM index and bit-rate is developed in Section 2. Secondly, the application of visual attention is described in Section 3. Subsequently, the performance of the proposed method is shown in Section 4. Finally, conclusions are presented in Section 5.

2. ADAPTIVE SSIM-RATE MODELS

Although the transformed residuals are unstable, it should be noted that the standard deviation of the transformed residuals is a stable parameter. It actually indicates an inherent property of the input sequence [4].

In this paper, we present the SSIM-Rate model by the standard deviation and the quantization step size. The standard deviation of the transformed residuals for one frame is defined as:

$$\sigma_{sd} = \sqrt{E(x^2) - [E(x)]^2} \tag{3}$$

where x is the DCT coefficient of the frame, E is the expectation. The relationship between σ_{sd} and bit-rate is shown in Fig. 1(a). An approximately linear relationship can be observed for sequences under a setting QP.

$$R = a \cdot (\sigma_{sd} + b) \tag{4}$$

where a and b are both constants. From Fig. 1(a), we can see that when the rate goes close to zero, the σ_{sd} gets close to 10. In order to obtain the best coding effect for all values of QP, we set $b = -11.5$ in our experiment. From Fig. 1(b), the relationship between a and QP can be approximated by:

$$a = 0.47 \times e^{(51-QP)/6.43} \tag{5}$$

Consequently, by incorporating (5) into (4), the rate model R can be obtained, that is:

$$R = 0.47 \times e^{(51-QP)/6.43} \times (\sigma_{sd} + b) \tag{6}$$

The SSIM-QP model in [5] is given as:

$$D_{ssim} = 10^{-4} e^{(QP+11.80)/6.87} \tag{7}$$

here D_{ssim} is the distortion by using the SSIM index as the distortion metric, and the Lagrange multiplier is calculated as:

$$\lambda = -\frac{dDssim}{dR} = -\frac{\partial Dssim/\partial Q}{\partial R/\partial Q} \tag{8}$$

The final Lagrange multiplier can be figured out by deriving the expression (6) and (7).

$$\lambda = \frac{10^{-7} \times 4.04}{\sigma_{sd} - 11.50} \times e^{0.30QP} \tag{9}$$

The probability of skipped blocks is much higher when σ_{sd} value is every small. That will cause poor encoding results. In order to avoid this case, the constraint condition is given that the $min\{\sigma_{sd} - 11.5\} = 10$.

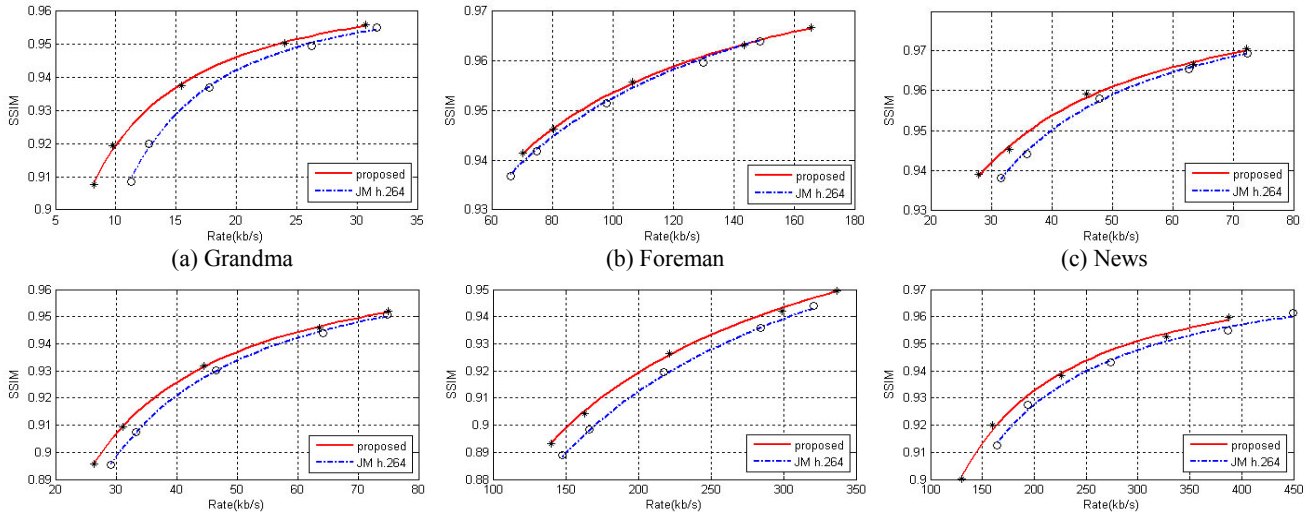
To perform the RDO, Lagrange multiplier should be determined before coding the current frame. However, σ_{sd} for the frame is unavailable at that time. Thus, they have to be predicted. Fortunately, as the inherent property of input sequence, σ_{sd} is a stable parameter which can be estimated by the average value of the five previously coded frames with the same type.

In this way, we model an adaptive Lagrange multiplier selection scheme based on the standard deviation of transformed residuals which indicates the property of the input sequence.

3. HUMAN ATTENTION BASED RDO MODEL

Human sensitivity of the distortion varies from region to region. Visual perception of distortion is more intensive in the positions with higher attention. We use the attention analysis method in [9] to get attention maps for frames. Each pixel in a frame has its corresponding attention value in the map. The SSIM index between two blocks x and y is defined as [4]:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$



(a) Grandma (b) Foreman (c) News

(d) Salesman (e) Soccer (f) Stefan

Fig. 2 Performance comparison

where μ_x is estimated by the mean intensity, μ_y is similar with μ_x.

$$\mu_x = \sum_{i=0}^{N} \omega_i x_i \tag{11}$$

and σ_x is an unbiased estimation of the standard deviation as the estimation of the signal contrast:

$$\sigma_x = \left(\sum_{i=1}^{N} \omega_i (x_i - \mu_x)^2\right)^{\frac{1}{2}} \tag{12}$$

where σ_{xy} can be estimated by:

$$\sigma_{xy} = \sum_{i=1}^{N} \omega_i (x_i - \mu_x)(y_i - \mu_y) \tag{13}$$

where the parameter ω is the weight of a pixel. The constraint condition of ω is $\sum_{i=1}^{N} \omega_i = 1$. Therefore, we use the attention value of a pixel to calculate ω:

$$\omega_i = \frac{a_i}{\sum_{j=1}^{N} a_j} \tag{14}$$

where a_i , a_j are the attention values of a pixel calculated by [9], and N is the local window size. The parameters C_1 and C_2 are the constants which are determined by experimental results. By incorporating (11)(12)(13)(14) into (10), we can obtain the SSIM index for a block. We can see that a pixel of higher attention is more influential to the SSIM index.

Generally speaking, people usually pay more attention to a region of an image instead of a pixel. Therefore, we make use of the attention parameter in the macroblock level. Firstly, the attention value of a macroblock can be figured out by the attention map:

$$A_i = \sum_{j \in MB_i} a_j \tag{15}$$

where A_i is the attention value of the macroblock i, and a_j is a pixel which belongs to the macroblock i. Secondly, we normalize the value range of A_i to be (0, 1]. A_i is used to RDO mode selection scheme as follow:

$$J = D_{ssim} \times (\alpha + \beta \times A_i) + \lambda R \tag{16}$$

where the expression$(\alpha + \beta \times A_i)$ means that small value of A_i correspond to low fidelities and bit-rate, while large value of A_i correspond to high fidelities and bit-rate. The parameters α and β is used to adapt the weight of attention. In our experiments, we set $\alpha = 0.5$, $\beta = 1$.

In this way, the final human visual perception based RDO model is obtained.

4. EXPERIMENTAL RESULTS

In this section, the proposed RDO algorithm is verified by the most recent H.264/AVC reference software JM17.1. All the test sequences in QCIF format are encoded under baseline profile. Evaluations are performed to compare the proposed method with JM17.1.

4.1. Objective evaluation

All the test sequences are encoded with fixed QPs (QP = 27,28,30,32,33), high complexity RDO and one I frame followed by 49 P frames. The SSIM index of each frame is obtained by averaging all of the 4x4 blocks. And the D_{SSIM} is calculated by $D_{SSIM} = 1 - SSIM$.

The average ΔD_{SSIM} and ΔRate of all test sequences (grandma, foreman, news, salesman, soccer, stefan) are calculated for various QPs. The results are shown in Table.1. It can be observed that the rate reduction and the D_{SSIM} reduction for low bit-rate are much higher than the ones for high bit-rate. The performance curves of the proposed methods are shown in Fig. 2. From the results we can see that for all test sequences, the coding effectiveness has been improved. When the encoded bit-rate is high, the visual quality is good for all methods. However, when the encoded bit-rate is low, the visual assessment of these algorithms is rather different. The simulation results are shown in Table. 2. On average, 7.28% D_{SSIM} reduction or 10.14% rate reduction is achieved. For the peak gain, 13.77% D_{SSIM} reduction or 21.29% rate reduction is achieved for grandma. Under the condition of high bit-rate, JM and the proposed method have the similarly good result. Under the condition

Table. 1 Results by averaging six sequences

QP	ΔD_{SSIM}	ΔRate
27	-1.69%	-3.08%
28	-4.37%	-7.89%
30	-4.79%	-8.83%
32	-7.89%	-10.24%
33	-9.23%	-12.31%

Table. 2 Simulation results of proposed method

Sequence	ΔD_{SSIM}	ΔRate
grandma	-13.77%	-21.29%
foreman	-3.33%	-4.95%
news	-7.94%	-7.64%
salesman	-7.07%	-8.70%
soccer	-4.41%	-9.76%
stefan	-7.17%	-8.52%
Average	*-7.28%*	*-10.14%*

of low bit-rate, an average 10.14% rate reduction is achieved by the proposed method. In the proposed method, the additional complexity is mainly brought by getting attention map. The average cost of attention analysis method in [9] is less than 3.1% coding time. The computation complexity on σ_{sd} and SSIM can be ignorable compared to other parts used in video coding such as motion estimation.

4.2. Subjective quality evaluation experiment

As shown in Fig. 3, the numbers attract much more human attention than other regions do. Therefore, the player's numbers are the most important regions. We can observe that the content in the red rectangles have a better perceptual quality by using our method. The numbers are more clearly in Fig. 3(a). Although there is some PSNR loss by the proposed algorithm, the visual quality has been improved. From Fig. 4, we can see that the visual quality of the proposed method and JM are almost the same. However, our method obtains average 21.29% bit-rate reduction.

5. CONCLUSIONS

In this paper, a human visual perception based distortion rate optimization method for video coding is presented. Firstly, the SSIM index is used as the distortion metric which is more consistent with the human perceptual distortion. Secondly, the standard deviation of transformed residuals which indicates the inherent property of input sequence is used to establish an adaptive Lagrange multiplier selection scheme. Thirdly, because the visual sensitivity of distortion varies in different regions of a sequence, the visual attention parameter is used to adapt the weight of distortion metric in RDO coding. Finally, we obtain the RDO model based on visual perception. Experimental results show that average 10.14% rate reduction is achieved. At the same time, the regions with the higher visual attention show better visual quality.

(a) Proposed (b) JM

Fig. 3 Quality comparison of the 39th reconstructed frame for football using bit-rate = 200 kbps. (a) Bits/pix = 0.248, PSNR= 31.02, SSIM = 0.8514; (b) Bits/pix = 0.251, PSNR= 31.72, SSIM = 0.8457.

(a) Proposed (b) JM

Fig. 4 The 44th reconstructed frame for grandma. (a)Bits/pix = 0.029, PSNR=34.61, SSIM=0.9309; (b) Bits/pix = 0.045, PSNR=35.25, SSIM = 0.9299.

6. ACKNOWLEDGEMENTS

This work was supported in part by National Natural Science Foundation of China: 61025011, 60833006, and 61001177, and in part by National Basic Research Program of China (973 Program): 2009CB320906.

7. REFERENCES

[1] G.J. Sullivan, T. Wiegand, "Rate-distortion optimization for video compression," *MSP*, vol.15, no.6, pp. 74-90, 1998.

[2] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, vol.13,no.4, pp. 600-612, 2004.

[3] Z. Wang, A.C. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures,"*MSP*, vol.26, no.1, pp. 98-117, 2009.

[4] X. Li, N. Oertel, A. Hutter and A. Kaup, "Laplace distribution based lagrangian rate distortion optimization for hybrid video coding," *CSVT*, vol.19, pp. 193-205, 2009.

[5] C. Yang, R. Leung, L. Po and Z. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," *ICIS*, vol.4, 2009.

[6] Y. Huang, T. Ou, P. Su and H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *CSVT*, vol.20, pp. 1051-8215, 2010.

[7] S. Wang, S. Ma, W. Gao, "SSIM based perceptual distortion rate optimization coding," *VCIP*, 2010.

[8] Z. Chen, W. Lin, K. Ngan, "Perceptual video coding: challenges and approaches," *ICME*, pp.784-789, 2010.

[9] H. Liu, S. Jiang, Q. Huang and C. Xu, "A generic virtual content insertion system based on visual attention analysis", *ACM Multimedia*, pp. 379-388, 2008.