RATE-SSIM OPTIMIZATION FOR VIDEO CODING

Shiqi Wang [1,2], *Abdul Rehman*[2], *Zhou Wang*[2], *Siwei Ma*[1], *Wen Gao*[1]

[1]Institute of Digital Media, Peking University, Beijing 100871, China
[2]Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

The structural similarity (SSIM) index has been found to be a good indicator of perceived image quality. In this paper, we propose a rate-SSIM optimization scheme for mode selection in H.264/AVC video coding. To derive the Lagrange multiplier based on the properties of input sequences, a novel reduced-reference statistical SSIM model and a source-side information combined rate model are established. The proposed method is fully standard-compatible. Experimental results demonstrate that, compared with conventional rate distortion optimization coding schemes, the proposed scheme can achieve better rate-SSIM performance and provide better visual quality.

Index Terms— SSIM index, rate distortion optimization, reduced-reference image quality assessment, Lagrange multiplier, rate-SSIM optimization

1. INTRODUCTION

Video codecs are primarily characterized in terms of the throughput of the channel and the perceived distortion of the reconstructed video. The fundamental issue in video coding is to obtain the best trade-off between the rate and perceived distortion. The process used to achieve this objective is commonly known as Rate Distortion Optimization (RDO). Mathematically, the RDO problem can be written as follows

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c \tag{1}$$

where D is the perceived distortion for a given rate budget R_c. This is a typical constrained optimization problem which can be converted to an unconstrained optimization problem by

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R \tag{2}$$

where J is called the Rate Distortion (RD) cost and the rate R is measured in number of bits per pixel. λ is known as the Lagrange multiplier which controls the trade-off between R and the perceived distortion D.

In practice, imperfect distortion models such as Sum of Absolute Difference (SAD) and mean squared error (MSE) are used in most actual implementations. Recently, a lot of work has been done to develop objective quality assessment measures which more accurately reflect perceived image distortion. Among these measures, the structural similarity (SSIM) index [1], which is proved to more effectively quantify the suprathreshold compression artifacts, has been preferred due to its simplicity and efficiency. It was incorporated into motion estimation [2], mode selection [3] and rate control [4] in hybrid video coding. However, the Lagrange multiplier was derived experimentally in [2] and [3] so that the properties of input sequences were ignored in the RDO scheme. In [4], an SSIM motivated rate control scheme was proposed based on an approximation R-D curve, while the properties of the SSIM index were not fully exploited.

In this paper, we use SSIM to define distortion model and propose a perceptual RDO scheme for mode selection. In particular, we incorporate a novel statistical reduced-reference SSIM model and a source-side information combined rate model in the RDO process to derive the Lagrange multiplier adaptively. Consequently, the mode for H.264 coding is selected by the SSIM index and the Lagrange multiplier.

2. SSIM BASED RATE DISTORTION OPTIMIZATION

To incorporate the SSIM index into the RDO process, the RD cost in (2) is redefined as

$$J = (1 - SSIM) + \lambda \cdot R \tag{3}$$

The spatial domain SSIM index [1] is based on similarities of local luminance, contrast and structure between a reference image and a distorted image. Given two local image patches **x** and **y**, the local SSIM index is defined as

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

where μ_x, σ_x and σ_{xy} are the mean, standard deviation and cross correlation between the two patches, respectively. C_1 and C_2 are used to avoid instability when the means and variances are close to zero. The SSIM index of the whole image is obtained by averaging the local SSIM indices calculated using a sliding window.

It is important to note that in the conventional RDO scheme, the final coding mode is determined based on the

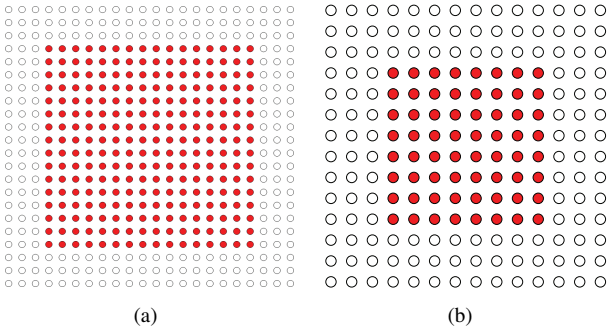

Fig. 1. Illustration of using surrounding pixels to calculate the SSIM index. Solid pixels: to be encoded; Hollow pixels: surrounding pixels from the input frame. (a) Y component; (b) Cb, Cr components.

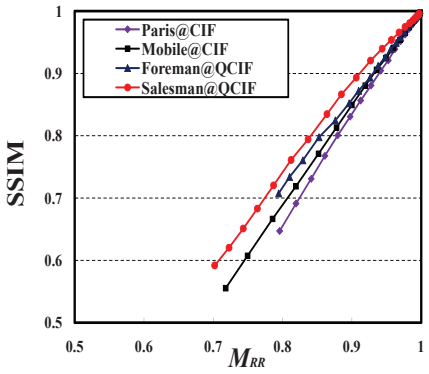

Fig. 2. Relationship between SSIM and M_{RR} for different sequences.

residual information only, while the properties of the reference image are ignored. Unlike MSE, the SSIM index is totally adaptive according to the reference signal [1].

The SSIM index is evaluated from overlapping blocks obtained using a sliding window, while in the video coding framework we encode individual non-overlapped blocks separately. To bridge this gap, we calculate the SSIM index between the reconstructed macroblock (MB) and the original MB with a larger window, as illustrated in Fig. 1. In case of Y component, the SSIM index of the current 16×16 MB to be encoded is calculated within a 22×22 block by a sliding window. For 8×8 Cb and Cr components, 14×14 blocks are used. In this way, the problem of discontinuities at the MB boundaries can also be alleviated.

Finally, the SSIM indices of Y, Cb and Cr components are weighted averaged to obtain a single measure of structural similarity. The weights of Y, Cb and Cr components are defined as $W_Y = 0.8$ and $W_{Cb} = W_{Cr} = 0.1$ [5].

3. STATISTICAL SSIM AND RATE MODELS

From (3), the Lagrange parameter is obtained by calculating the derivative of J with regarding to R, then setting it to zero and finally solving for λ,

$$\lambda = \frac{dSSIM}{dR} = \frac{\frac{dSSIM}{dQ}}{\frac{dR}{dQ}} \tag{5}$$

where Q is the quantization step. The equation above implies that in order to calculate λ for the current frame, without actually encoding it first, we need to establish both SSIM and the rate models.

SSIM is a full-reference (FR) measure that requires both the reference and distorted images to compute, and thus can not be directly used in this framework. Therefore, the SSIM model is deduced by a reduced-reference (RR) quality assessment algorithm which requires a set of RR features extracted from the reference frame and quantization process for quality evaluation.

Motivated by the DCT domain SSIM index [6], to calculate an RR-SSIM measure, we first divide the frame into 4x4 non-overlapping blocks and calculate the DCT transform of each block. Furthermore, we divide the coefficients into subbands. This is achieved by grouping the coefficients having the same frequency from each 4x4 DCT window, which results in 16 subbands in the 4x4 DCT transform. Let σ_i be the standard deviation of the DCT coefficients from the i^{th} subband of the original frames. The new RR distortion measure is defined as

$$M_{RR} = (1 - \frac{D_0}{2\sigma_0^2 + C_1})(1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{D_i}{2\sigma_i^2 + C_2}) \tag{6}$$

where N is the block size. D_i is the MSE between the original and distorted frames in the i^{th} subband and can be modeled by Laplace distribution of the residuals [7].

Because the design principles between M_{RR} and SSIM index are similar, M_{RR} exhibits a nearly perfect linear relationship with SSIM, as shown in Fig. 2. More specifically, the RR-SSIM estimator can be written as

$$\hat{S} = \alpha + \beta \cdot M_{RR} \tag{7}$$

Before coding the current frame, the parameters α and β can be estimated by two points on the line, which are (1,1) and the estimated values of $\hat{S}$ and M_{RR} from the previous frames.

The rate model is derived based on three observations. First, the blocks for which skip mode is selected should not be included in the rate model because the skipped blocks will not be entropy coded [7]. Second, in H.264 the side information (or header bits) may take a large portion of the total bits, especially in low bit rate video coding scenario [8]. Third, the dependent entropy coding would make the estimated entropy value larger than the true source rate [7]. In this work, all above mentioned observations are taken into consideration in the development of the rate model.

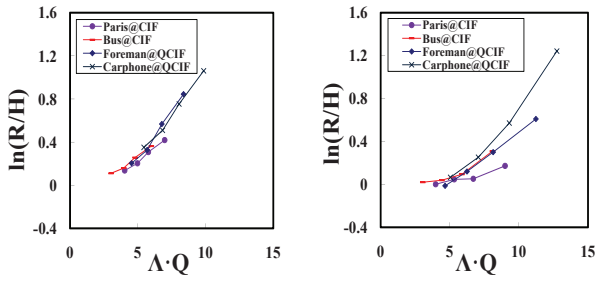

Fig. 3. The relationship between $ln(R/H)$ and $\Lambda \cdot Q$ for different sequences. (a) P frame, GOP structure: IPP; (b) B frame, GOP structure: IBP.

We adopt an entropy model that excludes the bit rate of the skipped blocks [7]:

$$H = (1-P_s)\cdot[-\frac{P_0-P_s}{1-P_s}\cdot\log_2\frac{P_0-P_s}{1-P_s} - 2\sum_{n=1}^{\infty}\frac{P_n}{1-P_s}\cdot\log_2\frac{P_n}{1-P_s}] \quad (8)$$

where P_s is the probability of the skipped blocks, P_0 and P_n are the probabilities of transformed residuals quantized to the zero-th and n-th quantization levels, respectively.

Supposing Λ to be the Laplace parameter of the transformed residuals, a linear relationship between $ln(R^*/H)$ and $\Lambda \cdot Q$ is observed in [7], where R^* represents the source rate. Notice that for the same quantization step, a larger Λ indicates smaller residuals, which corresponds to a larger proportion of the side information. Interestingly, for total rate R, there is also an approximately linear relationship between $ln(R/H)$ and $\Lambda \cdot Q$, as can be seen in Fig. 3. Also, the relationship is totally consistent with the effect of dependent entropy coding and side information. Considering the dependent entropy coding, the estimated entropy value should be larger than the source rate and thus in high bit rate video coding $ln(R/H)$ approaches zero. However, in the low bit rate coding scenario, the side information plays a greater role and the entropy should be smaller than the total rate, which leads to a larger $ln(R/H)$ because of the dominating effect of side information, as illustrated in Fig. 3. Consequently, the final rate model R can be approximated by

$$R = H \cdot e^{\xi\Lambda Q+\psi} \quad (9)$$

where the parameters ξ and ψ are not very sensitive to the video content. For CAVLC and CABAC entropy coding methods, ξ and ψ are set to be

$$\xi = \begin{cases} 0.03 & B\ \ frame \\ 0.07 & Otherwise \end{cases} \quad \psi = \begin{cases} -0.07 & B\ \ frame \\ -0.1 & Otherwise \end{cases} \quad (10)$$

4. EXPERIMENTAL RESULTS

To validate the accuracy and efficiency of the proposed perceptual RDO scheme, we integrate our mode selection scheme into the H.264/AVC reference software JM15.1. All test video sequences are in YCbCr 4:2:0 format.

Table 1. MAE and PLCC between FR-SSIM and RR-SSIM for different sequences.

Sequences	GOP Structure	PLCC	MAE
Foreman(CIF)	IPP	0.999	0.002
News(CIF)	IPP	0.999	0.002
Mobile(CIF)	IBP	0.999	0.004
Paris(CIF)	IBP	0.999	0.003
Highway(QCIF)	IPP	0.998	0.003
Suize(QCIF)	IPP	0.998	0.004
Carphone(QCIF)	IBP	0.997	0.006
Akiyo(QCIF)	IBP	0.998	0.005
All		0.998	0.004

To verify the validity of our proposed SSIM model, we compare the estimated (RR) and actual (FR) values of the SSIM index for eight sequences with a set of various QP values. The Pearson Linear Correlation Coefficient (PLCC) and Mean Absolute Error (MAE) between FR-SSIM and RR-SSIM are given in Table 1. The values suggest that the proposed RR-SSIM model achieves high accuracy for different sequences.

We compare the R-D performance of our proposed perceptual RDO algorithm and the conventional RDO with distortion measured in terms of SSIM and PSNR. The common coding configurations are set as follows: all available inter and intra modes are enabled; five reference frames; one I frames followed by 99 inter frames; high complexity RDO and the fixed quantization parameters are set from 28 to 40. The results of the experiments are shown in Table 2.

For IPP GOP structure, 14% rate reduction on average is achieved in terms of SSIM index. When the GOP structure is IBP, the rate reduction is about 8% on average. For the peak gain, 17.23% rate reduction is achieved for $Container$. The lower gain of IBP coding scheme could be explained by two reasons. First, with the conventional RDO, a large percentage of MBs in B frames have already been coded with the best mode. Second, the parameter estimation scheme as proposed in [7] is not very accurate for this GOP structure because the frames of the same coding types are not adjacent to each other. We have also compared the performance in terms of PSNR, which is also obtained by weighted averaging the respective values of Y, Cb and Cr components [7]. Because of the adaptivity of our proposed scheme, for some sequences such as $Salesman$ and $Container$, PSNR increases. However, on average PSNR decreases because our optimization objective is SSIM rather than PSNR.

Table 2. Performance of the proposed scheme (compared with the original RDO technique).

Sequence		$\Delta SSIM$	ΔR^*	$\Delta PSNR$
Silent (CIF)	IPP..	0.0115	-14.62%	-0.14dB
	IBP..	0.0064	-8.07%	-0.25dB
Flower (CIF)	IPP..	0.0076	-14.34%	-0.66dB
	IBP..	0.0034	-6.73%	-0.55dB
Bus (CIF)	IPP..	0.0136	-14.71%	-0.51dB
	IBP..	0.0081	-8.95%	-0.62dB
Salesman (QCIF)	IPP..	0.0185	-17.09%	0.08dB
	IBP..	0.0096	-8.45%	-0.15dB
Carphone (QCIF)	IPP..	0.0038	-6.89%	-0.47dB
	IBP..	0.0008	-2.11%	-0.67dB
Container (QCIF)	IPP..	0.0087	-17.23%	0.06dB
	IBP..	0.0049	-12.41%	-0.26dB
Average	IPP..	0.0106	-14.15%	-0.27dB
	IBP..	0.0055	-7.79%	-0.42dB

* ΔR in terms of SSIM.

Fig. 4 shows the original frame, H.264 coded frame with the conventional RDO and H.264 coded frame with the proposed RDO method. Since our proposed RDO scheme is based on the optimization of SSIM, PSNR of our scheme is lower while higher SSIM value is achieved. As can be seen from Fig. 4, the quality of the reconstructed frame has been obviously improved and more details have been preserved by the proposed scheme.

5. CONCLUSION

We propose an RDO scheme for H.264/AVC video coding, aiming for achieving the best rate-SSIM performance. The novelty of our approaches lies in the adaptive Lagrange multiplier selection methods at the frame level, where we incorporated a new RR-SSIM estimation algorithm and a source-side information combined rate model. Our experiments show that the proposed scheme offers significant rate reduction while keeping the same level of SSIM quality value. Visual quality improvement is also observed when compared with conventional RDO scheme.

6. ACKNOWLEDGMENT

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada in the form of Discovery, Strategic and CRD Grants, Ontario Early Researcher Award program, National Basic Research Program of China (973 Program, 2009CB320903), National Science Foundation (60833013, 60803068) of China and the State Scholarship Fund from the China Scholarship Council, which are gratefully acknowledged.

(a) (b) (c)

Fig. 4. Visual quality comparison between conventional and proposed RDO schemes (the crop region is from the fortieth frame of $Flower$). (a) Original; (b) H.264 coded with conventional RDO; Bit rate: 203.5 kbit/s, SSIM: 0.8710, PSNR: 27.70dB; (c) H.264 coded with proposed RDO; Bit rate: 194.25 kbit/s, SSIM: 0.8777, PSNR: 27.08dB.

7. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[2] Z. Mai, C. Yang, K. Kuang, and L. Po, "A novel motion estimation method based on structural similarity for H.264 inter prediction," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 913–916, 2006.

[3] C. Yang, H. Wang, and L. Po, "Improved inter prediction based on structural similarity in H.264," *IEEE International Conference on Signal Processing and Communications*, vol. 2, pp. 340–343, 2007.

[4] T. Ou, Y. Huang, and H. Chen, "A perceptual-based approach to bit allocation for H.264 encoder," *SPIE Visual Communications and Image Processing*, Jul. 2010.

[5] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, pp. 121–132, Feb. 2004.

[6] S. Channappayya, A. C. Bovik, and Jr. R. W. Heathh, "Rate bounds on SSIM index of quantized images," *IEEE Trans. on Image Processing*, vol. 17, pp. 1624–1639, Sep. 2008.

[7] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, pp. 193–205, Feb. 2009.

[8] D. Kwon, M. Shen, and C. Kuo, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, pp. 517–529, May 2007.