International Journal of Information Technology, Vol. 11, No. 7, 2005

A Novel Rate-Distortion Optimization Based on Structural Similarity in Color Image Encoder

Zhi-Yi Mai[1], Chun-Ling Yang[2], and Sheng-Li Xie[3]

School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, Guangdong, 510640, China
[1] kathymaizy@yahoo.com.cn,
[2] eeclyang@scut.edu.cn,
[3] adshlxie@scut.edu.cn

Abstract

In H.264 I-frame encoder, the best intra prediction modes are chosen by utilizing the rate-distortion (R-D) optimization whose distortion is the sum of squared differences (SSD), which means the same as MSE, between the reconstructed and the original block. Recently a new image quality measurement called Structural Similarity (SSIM) based on the degradation of structural information was brought forward. It is proved that the SSIM can provide a better approximation to the perceived image distortion than the currently used PSNR (or MSE). In this paper, a new rate-distortion optimization for H.264 I-frame encoder using SSIM as the distortion metric is proposed. Experiment results show that the proposed algorithm can reduce 2~4.2% bit rate while maintaining the perceptual quality, but the computation complexity increases a little.

Keyword: Structural similarity (SSIM), intra prediction, rate-distortion optimization.

1 Introduction

As the rapid development of digital techniques and increasing use of internet, image/video compression plays a more and more important role in our life. The newest international video coding standard H.264 adopts many advanced techniques, such as directional spatial prediction in I-frame encoder, variable and Hierarchical block transform, arithmetic entropy coding, multiple reference frame motion compensation, deblocking etc. All these novel and advanced techniques make it provide approximately a 50% bit rate savings for equivalent perceptual quality relative to the performance of prior standards [1]. Except for the new innovations, the rate-distortion tradeoff of H.264 is still controlled by the Lagrangian optimization techniques, just like the prior standards, MPEG-2, H.263 and MPEG-4. In the R-D optimization function for H.264 intra prediction, distortion is measured as the sum of squared differences (SSD) between the reconstructed and the original block, which has the same meaning with MSE. Although Peak Signal-to-Noise Ratio (PSNR) and MSE are currently the most widely used objective metrics due to their low complexity and clear physical meaning, they were also widely criticized for not correlating well with Human Visual System (HVS) for a long time [2]. During past several decades a great deal of effort has been made to develop new image quality assessment based on error sensitivity theory of HVS, but only limit success has been achieved by the reason that the HVS is much more complex and has not been well comprehended.

Recently a new philosophy for image quality measurement was proposed, based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. It follows that a measure of structural information change can provide a good approximation to perceived image distortion [3], [4]. In this new theory, an item called Structural Similarity (SSIM) index including three comparisons is introduced to measure the structural information change. Experiments showed that the SSIM index method is easy to implement and can better correspond with human perceived measurement than PSNR (or MSE). Thus, in this paper we propose to employ SSIM in the rate-distortion optimizations of H.264 I-frame encoder to choose the best prediction mode(s).

The remainder of this paper is organized as follows. In section 2, the I-frame encoding of H.264 and the idea of SSIM is summarized. Detail of our proposed method is given in section 3. Section 4 presents the experimental results to demonstrate the advantage of the SSIM index method. Finally, section 5 draws the conclusion.

2 H.264 I-frame Encoder and SSIM

2.1 H.264 I-Frame Encoder

In H.264 I-frame encoder, each picture is partitioned into fixed-size and non-overlapped macroblocks (MB) each of which covers a rectangular area of 16×16 samples of the luma component and 8×8 samples of each chroma component. Then each macroblock is spatially predicted by using its neighbouring samples of the previously coded blocks which are to the left and/or above the block, and the prediction residual is integrally transformed, quantized and entropy encoded. The JVT reference software version JM92 of H.264 [5] provides three classes of intra prediction types denoted as Intra_16x16, Intra_8x8 and Intra_4x4 for the luma components and an Intra_Chroma type for the chroma components. The Intra_16x16 which supports 4 prediction modes performs prediction of the whole macroblock and is suited for smooth area, while both Intra_8x8 and Intra_4x4 which performs the prediction of 8×8 and 4×4 block respectively support 9 prediction modes and are suited for detailed part of the picture. The Intra_Chroma predicdtion is performed for the whole 8×8 chroma block and supports 4 prediction modes which is similar to the Intra_16x16 prediction. The best prediction modes are chosen by utilizing the R-D optimization formula [6] which is described as:

$$J(s,c,MODE\mid QP)=D(s,c,MODE\mid QP)+\lambda_{MODE}R(s,c,MODE\mid QP)\ . \qquad \textbf{(1)}$$

In the above formula, the distortion D(***s,c***,MODE|QP) is measured as SSD between the original block ***s*** and the reconstructed block ***c***. Herein is the SSD defined as:

$$SSD=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[s(i,j)-c(i,j)]^2 \quad . \tag{2}$$

where M and N are the dimensions of the image in width and height respectively, and s(i,j) and c(i,j) are the original and reconstructed pixel values at position (i,j).

In formula (1), QP is the quantization parameter. MODE is the prediction mode. R(***s,c***,MODE|QP) is the bit number after encoding the block. The mode(s) with the minimum J(***s,c***,MODE|QP) are chosen as the prediction mode(s) of the macroblock.

2.2 Structural Similarity (SSIM)

Different from the popular used MSE which simply quantifies the strength of error signal, the new idea of SSIM index is to introduce the measure of structural information degradation including three comparisons: luminance, contrast and structure [4]. It's defined as

$$\mathrm{SSIM}(\boldsymbol{x},\mathrm{y})=\mathrm{l}(\boldsymbol{x},\mathrm{y})\cdot\mathrm{c}(\boldsymbol{x},\boldsymbol{y})\cdot\mathrm{s}(\boldsymbol{x},\boldsymbol{y}) \quad . \tag{3}$$

where $\mathrm{l}(\boldsymbol{x},\boldsymbol{y})$ is Luminance comparison, $\mathrm{c}(\boldsymbol{x},\boldsymbol{y})$ is Contrast comparison and $\mathrm{s}(\boldsymbol{x},\boldsymbol{y})$ is Structure comparison. They are defined as:

$$\mathrm{l}(\boldsymbol{x},\boldsymbol{y})=\frac{2\mu_{\mathrm{x}}\mu_{\mathrm{y}}+\mathrm{C}_1}{\mu_{\mathrm{x}}^2+\mu_{\mathrm{y}}^2+\mathrm{C}_1} \quad . \tag{4}$$

$$\mathrm{c}(\boldsymbol{x},\boldsymbol{y})=\frac{2\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}+\mathrm{C}_2}{\sigma_{\mathrm{x}}^2+\sigma_{\mathrm{y}}^2+\mathrm{C}_2} \quad . \tag{5}$$

$$s(\mathbf{x},\mathbf{y})=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \quad . \tag{6}$$

where ***x*** and ***y*** are two nonnegative image signals to be compared, μ_{x} and μ_{y} are the mean intensity of image ***x*** and ***y*** respectively (as formula (7)), σ_{x} and σ_{y} are the standard deviation of image ***x*** and ***y*** respectively (as formula (8)), σ_{xy} is the covariance of image x and y (as formula (9)).

$$\mu_x=\frac{1}{N}\sum_{i=1}^{N}x_i \quad . \tag{7}$$

$$\sigma_x = \left(\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} . \tag{8}$$

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) . \tag{9}$$

In fact, without C_3, the equation (6) is the correlation coefficient of image ***x*** and ***y***. C_1, C_2 and C_3 are small constants to avoid the denominator being zero. It's recommended by [4]:

$$C_1 = (K_1 L)^2 ,\ C_3 = C_2 / 2 ,\ C_2 = (K_2 L)^2 . \tag{10}$$

where K_1,K_2<<1 and L is the dynamic range of the pixel values (255 for 8-bit gray-scale images). In addition, the higher the value of SSIM(x,y) is, the more similar the image ***x*** and ***y*** are.

3 The R-D Optimization Using Structural Similarity in H.264

As the SSIM index method performs better as the image quality measurement than MSE (SSD), we propose to replace the SSD with the SSIM index in the R-D optimization of H.264 I-frame encoder.

According to the theory of SSIM, the quality of the reconstructed picture is better when its SSIM index is higher while the SSD performs the other way. Therefore the distortion in our method is measured as:

$$D(s,c,MODE \mid QP) = 1 - SSIM(s,c) . \tag{11}$$

where ***s*** and ***c*** are the original and reconstructed image block respectively.

Due to the change of distortion measure, the Lagrangian multiplier should be modified correspondingly. In conformity to the relation between SSIM(***s***,***c***) and R(***s***,***c***,MODE|QP) and motivated by the theory in [7] and [8], the new Lagrangian multiplier in our algorithm becomes

$$\lambda_{MODE} = 1.11 * 2^{(QP-60)/5} . \tag{12}$$

where QP denotes the quantization parameter. Consequently, the new R-D cost function can be written as:

$$J(s,c,MODE \mid QP) = 1 - SSIM(s,c) + 1.11 * 2^{(QP-60)/5} * R(s,c,MODE \mid QP) \tag{13}$$

The major steps for each macroblock selecting the best prediction mode(s) in our method can be summarized as follows:

Step 1: Choose one Intra_Chroma prediction mode and generate the intra prediction blocks for U and V component respectively.

Step 2: Find the best Intra_16x16 prediction mode.

a. Generate the four prediction blocks respectively for the Luma component according to the four Intra_16x16 prediction modes.

b. Perform Hadamard transform for the residual blocks and then sum up the absolute values of all the Hadamard transform coefficients as the cost.

c. The mode that has the lowest cost is chosen as the best Intra_16x16 prediction mode.

Note: This step is the same as the H.264 algorithm.

Step 3: Find the best Intra_4x4 prediction modes

Divide the Luma component of that macroblock into sixteen 4×4 non-overlapped blocks. For each 4×4 block do the following sub-steps:

a. Generate nine prediction blocks based on the nine Intra_4x4 prediction modes.

b. Compute the SSIM of the 4×4 reconstructed block and the original one.

c. Calculate the cost of the 4×4 block according to formula (13).

d. The mode that has the minimum cost is chosen as the best mode.

Step 4: Find the best Intra_8x8 prediction modes

Divide the Luma component of that macroblock into four 8×8 non-overlapped blocks. For each 8×8 block do the following sub-steps:

a. Generate nine prediction blocks based on the nine Intra_8×8 prediction modes.

b. Compute the SSIM of the 8×8 reconstructed block and the original one.

c. Calculate the cost of the 8×8 block according to formula (13).

d. The mode that has the minimum cost is chosen as the best mode.

Step 5: Find the best prediction mode for the whole macroblock

a. Figure out the SSIM of the reconstructed and the original macroblock in best Intra_16x16 mode, the best Intra_4x4 modes and the best Intra_8x8 modes respectively.

As each macroblock includes 16x16 pixels of Y component, 8x8 pixels of U component and 8x8 pixels of V component, we first count the SSIM of the reconstructed and the original block for each component and then combine them to a weighted averaged SSIM. Following weighted summation is used to generate the quality index for each macroblock by the reason that HVS is more sensitive to luma than chroma component.

$$SSIM_{MB} = 0.5 * SSIM_{Y} + 0.25 * SSIM_{U} + 0.25 * SSIM_{V} \ . \qquad \mathbf{(14)}$$

b. Calculate the rate-distortion cost using formula (13) for the best Intra_16x16 mode, the best Intra_4x4 modes and the best Intra_8x8 modes of the whole macroblock respectively.

c. The mode having the minimum cost is chosen as the best prediction mode of that macroblock.

Step 6: Repeat step 1 to step 5 until all the Intra_Chroma prediction modes are used.

4 Experimental Results

Experiments are carried out using several color video pictures of various sizes (as Table 1) in YUV format (4:2:0). All the experiments are based on the JVT reference software JM92 program [5] and conducted on a P4/2.0GHz personal computer with 256MB RAM and Microsoft Windows 2000 as the operation system.

An item called MSSIM is introduced to indicate the quality of the entire image. First we calculate the local SSIMs for the Y component by using 16×16 slide window, which moves rightwards and downwards pixel by pixel. Then these local SSIMs are averaged into $\mathrm{MSSIM_Y}$. $\mathrm{MSSIM_U}$ for U component and $\mathrm{MSSIM_V}$ for V component are generated in the similar way while a 8×8 slide window is used instead. Finally, the $\mathrm{MSSIM_Y}$, $\mathrm{MSSIM_U}$ and $\mathrm{MSSIM_V}$ are combined into an overall image quality measurement MSSIM as formula (14).

Table. 1 Test pictures

Size	Picture	
176×144	Apple	Coastguard
256×256	House	Tiffany
512×512	Baboon	Lena

Results in terms of total bits of the compressed image, MSSIM (a weighted average of Y, U, V component as formula (13)) of the whole reconstructed image and the comparison between the two methods are listed in Table 2~4 with the Quantization Parameter (QP) equal to 10, 20 and 30 respectively.

Results in Table 2 to 4 show that the proposed algorithm can achieve about 2~4.2% bits savings while maintaining almost the same MSSIM index comparing with the original H.264 algorithm. In order to illustrate the perceptual quality of the reconstructed image, here we show the original and reconstructed images with the largest MSSIM decrement in our experiments in Fig. 1 to 3, from which it’s clear that the visual difference between the reconstructed images using H.264 JM92 (Fig.1-3 (b)) and our proposed algorithm (Fig.1-3 (c)) can hardly be found. That means the new R-D optimization algorithm can achieve about 2~4.2% bit saving while maintaining almost the same perceptual quality. Results also show that our method can retain the computation complexity as H.264 for small QP (QP=10), but it costs 2.3~3.8% more coding time than H.264 when QP is large (QP=30).

Table 2. Results of comparison between H.264 and our method with QP=10

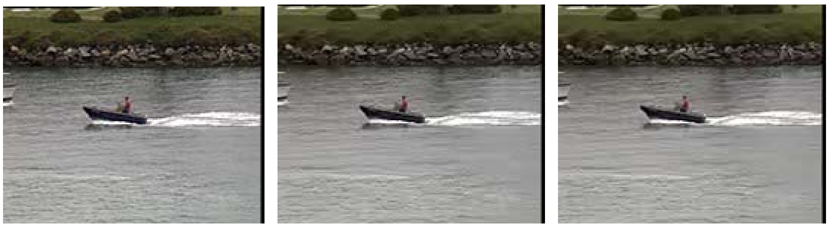
Image	H.264_JM92			Our method			Comparison (%)		
	Bits	MSSIM	Time (ms)	Bits	MSSIM	Time (ms)	Bit Inc.	MSSIM Inc.	Time Inc.
Apple	81512	0.9969	2567	78072	0.9965	2599	-4.22	-0.04	1.21
Coast-guard	111888	0.9971	3405	107520	0.9968	3433	-3.90	-0.03	0.83
House	254536	0.9967	7962	245320	0.9963	8034	-3.62	-0.05	0.91
Tiffany	331816	0.9975	8927	321464	0.9970	8953	-3.12	-0.04	0.30
Baboon	1860424	0.9985	43658	1822272	0.9984	44039	-2.05	-0.02	0.87
Lena	1147472	0.9967	33102	1104368	0.9962	33572	-3.76	-0.06	1.42

Table 3. Results of comparison between H.264 and our method with QP=20

Image	H.264_JM92			Our method			Comparison (%)		
	Bits	MSSIM	Time (ms)	Bits	MSSIM	Time (ms)	Bit Inc.	MSSIM Inc.	Time Inc.
Apple	23864	0.9814	1788	22944	0.9802	1828	-3.86	-0.11	2.27
Coast-guard	55848	0.9889	2494	53928	0.9882	2530	-3.44	-0.07	1.44
House	95272	0.9798	5539	92272	0.9787	5675	-3.15	-0.10	2.46
Tiffany	146560	0.9785	6275	142584	0.9774	6366	-2.71	-0.11	1.44
Baboon	1106184	0.9860	32659	1075896	0.9850	33070	-2.74	-0.10	1.26
Lena	414832	0.9735	22465	400112	0.9722	23069	-3.55	-0.14	2.69

Table 4. Results of comparison between H.264 and our method with QP=30

Image	H.264_JM92			Our method			Comparison (%)		
	Bits	MSSIM	Time (ms)	Bits	MSSIM	Time (ms)	Bit Inc.	MSSIM Inc.	Time Inc.
Apple	7528	0.9622	1472	7328	0.9597	1525	-2.66	-0.25	3.63
Coast-guard	19616	0.9549	1811	19064	0.9514	1851	-2.81	-0.37	2.23
House	27296	0.9375	4195	26304	0.9340	4328	-3.63	-0.37	3.16
Tiffany	43376	0.9193	4489	41824	0.9156	4622	-3.58	-0.40	2.96
Baboon	445888	0.9006	22969	429056	0.8958	23497	-3.77	-0.53	2.30
Lena	108984	0.9324	16652	105808	0.9287	17288	-2.91	-0.40	3.82

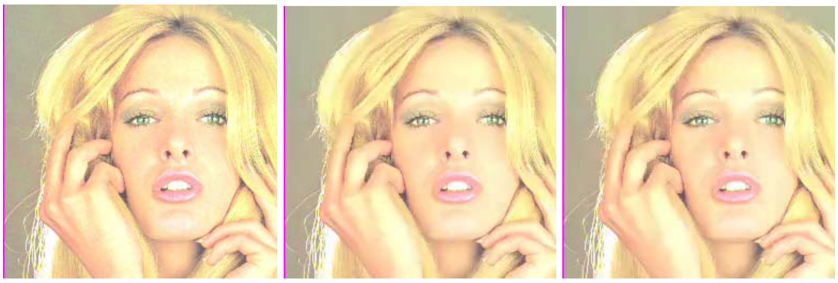


(a) Coastguard (original)

(b) Encoded by H.264 I-frame encoder with QP=30

(c) Encoded by our method with QP=30

Fig.1. The reconstructed image produced by the two methods respectively for Coastguard

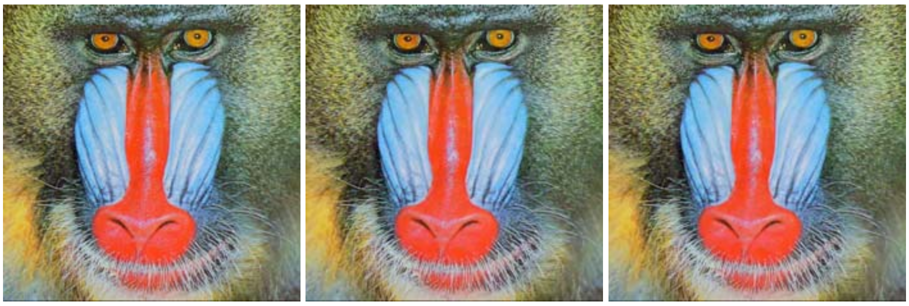


(a) Tiffany (original)

(b) Encoded by H.264 I-frame encoder with QP=30

(c) Encoded by our method with QP=30

Fig.2. The reconstructed image produced by the two methods respectively for Tiffany



(a) Baboon (original)

(b) Encoded by H.264 I-frame encoder with QP=30

(c) Encoded by our method with QP=30

Fig.3. The reconstructed image by the two methods respectively for Baboon

5 Conclusion

In this paper, we propose a new R-D optimization using the structural similarity (SSIM) instead of SSD for quality assessment in H.264 I-frame encoder. Experiments show that it can reduce approximately 2~4.2% bit rate while maintaining the same perceptual quality and costing almost the same encoding time for small QP, but a little more for large QP. The improvement of coding efficiency is not very large, but the new idea and the beginning results are inspiring. Thus, it's possible to obtain better results through further study. In addition, the proposed R-D optimization can be transplanted easily into motion estimation of inter frame encoding.

Acknowledges

The work described in this paper was substantially supported by research projects from National Natural Science Foundation of China. [Project No. 60402015, No.60325310]

References

[1] Wiegand, T., Sullivan, G.J., Gisle, B., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. IEEE Trans. CAS for Video Technology. Vol. 13 (7) (2003), pp. 560-576.

[2] Mannos, J.L., Sakrison, J.D.: The effects of a visual fidelity criterion on the encoding of images. In IEEE Trans. Information Theory. Vol. 20 (4) (1974), pp. 525-536.

[3] Wang, Z., Bovik, A.C., Lu, L.: Why is image quality assessment so difficult. In Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol. 4, Orlando, FL (2002), pp. 3313-3316.

[4] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing. Vol. 13 (4) (2004), pp. 600–612.

[5] http://bs.hhi.de/~suehring/tml/download.

[6] Siwei, M., Wen, G., Peng, G., Yan, L.: Rate control for advance video coding (AVC) standard. In Proc. IEEE International Symposium on Circuits and Systems, Vol. 2, Bangkok, Thailand (2003), II pp. 892-895.

[7] Wiegand, T., Girod, B.: Lagrangian multiplier selection in hybrid video coder control. In Proc. IEEE Int. Conf. Image Processing, Thessaloniki, Greece (2001), pp. 542-545.

[8] Sullivan, G.J., Wiegand, T.: Rate-Distortion Optimization for Video Compression, IEEE Signal Processing Magazine. Vol. 15 (6) (1998), pp. 74-90.

Zhi-Yi Mai is currently a graduate student of South China University of Technology, Guangzhou, China, majoring in Electronic and Information Engineering.

Her research interests are video coding and image quality assessment.

Chun-Ling Yang received the Ph.D degree from Department of Electronic Engineering, Nanjing University of Science and Technology, Nanjing, China, in 1999.

Now, She is with the School of Electronic and Information Engineering, South China University of Technology, as an associate professor. Her research interests are image analysis, image and video coding.

Sheng-Li Xie is a professor at the Institute of Radio and Automation, South China University of Technology. He is also a senior member of IEEE.

Now his research interests include nonlinear learning control, robotic systems, adaptive acoustic echo cancellation, blind signal processing and image processing.