An SSIM-Optimal H.264/AVC Inter Frame Encoder

Chun-Ling Yang[1], Rong-Kun Leung[1], Lai-Man Po[2], Zhi-Yi Mai[3]
[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong, 510641, China
[2] Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China
[3] Guangdong Nortel Telecommunications Equipment Co. Ltd. (GDNT)
eeclyang@scut.edu.cn, samuellrk@hotmail.com, eelmpo@cityu.edu.hk, kathymai@gdnt.com.cn

***Abstract*—Rate-distortion optimization (RDO), in which distortion metric plays a vital role, has been proved to be an effective way in hybrid video coding. This paper proposes an improved rate-distortion optimization method based on SSIM (IRDO-SSIM) in RDO mode selection process. And the derivation of the proper multiplier to fit for the IRDO-SSIM is mainly described in this paper. Simulation results show that the proposed algorithm has better rate-distortion performance, especially for image sequences with middle-motion complexity or low encoding bit-rate as comparing with H.264/AVC using conventional RDO.**

Keywords- H.264, Structural Similarity (SSIM), rate-distortion optimization (RDO), distortion metric.

I. INTRODUCTION

Modern video compression techniques offer the possibility to store or transmit the vast amount of the data necessary to present digital videos in an efficient and robust way. With increasing use of multimedia technologies, the enormous volume of video data is constantly fueling the demand for better and better compression performance. To address this need, more and more coding modes are developed to improve the coding efficiency. For example, in the H.264 standard-compliant coding environment, up to seven block types and 16 reference frames are allowed. Consequently, Rate-distortion Optimization (RDO) is developed to choose a best mode from many available candidate modes, and it is widely used in video compression applications.

RDO for video compression can be classified into two categories. The first category computes the theoretical RD function based on a given statistic model for video data, e.g.,[1][2]. The second category uses an operational RD function, which is computed based on the data to be compressed, such as ρ domain based RDO [3], context based RDO [4] and Laplace distribution based RDO [5]. Both of the two categories above are based on the PSNR-Rate framework. However the challenge for designing a method under this framework is that PSNR do not correlate well with HVS [6], which means that they cannot measure the images' perceptual distortions well. Thus, using advanced image quality metric in rate-distortion optimization (RDO) may achieve more efficient video compression methods.

Structure similarity (SSIM) [7] is a newly developed image quality measurement method, which extracts structure information from two corresponding image blocks. Because of its better performance in image quality assessment than PSNR, SSIM has been introduced in the newest H.264 as one of the video quality metrics [12]. In our previous works [8], SSIM has been adopted as a distortion metric in RDO and motion estimation. Experiment results show that the algorithm can save much more bit-rate at the expense of a little quality decline. However， what make the work challenging is that, when SSIM instead of SSD (Sum of Square Differences) is used as distortion metric in the RDO process, the multiplier λ , which is a very important parameter in RDO, has to be modified to fix the SSIM-rate optimization. In [8], the modified Lagrangian multipliers for three *QP* values (*QP*=10, 20 and 30) have been obtained by intensive experiment respectively. But it is far from enough, because the multipliers for the other *QP* values are still unknown. On the other hand, as SSIM is a little more complex than SSD and SAD (Sum of Absolute Differences), the computation load for encoder is another problem [8]. In this paper, we devote to improve the algorithm in the following two aspects: First, SAD is reused in motion estimation to solve the problem of computational complexity; And then, an improved RDO based on SSIM (IRDO-SSIM) is proposed, in which SSIM-rate multiplier is used to substitute the Lagrangian multiplier in IRDO-SSIM, and the calculation formula of SSIM-rate multiplier is derived.

The remainder of this paper is organized as follows. A brief introduction to the backgrounds about IRDO-SSIM, inter mode decision and Lagrangian multiplier is given in Section II. The proposed IRDO-SSIM is depicted in detail in Section III. Experiment results and analysis of the proposed algorithm are given in section IV. Finally the conclusion is given in Section V.

II. BACKGROUND

A. Structure similarity index (SSIM)

SSIM exhibits much more consistency with subjective measure compared with other image assessment methods,

978-1-4244-4738-1/09/$25.00 ©2009 IEEE

which include three comparisons: luminance, contrast and structure [7]. It is defined as follows:

$$SSIM = l(x,y) \times c(x,y) \times s(x,y) \tag{1}$$

The term $l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$ is the luminance comparison, $c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$ is the contrast comparison and $s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$ is the structure comparison. The quantities in above equations x, y are two nonnegative image signals, μ_x, μ_y are the means of x and y respectively, σ_x, σ_y are the corresponding standard deviations of x and y, and σ_{xy} is the sample cross-covariance between x and y. C_1, C_2 and C_3 are used to stabilize the distortion measure to avoid the denominator being zero or too small. As is recommended in [7]: $C_1 = (K_1 \times L)^2$, $C_2 = (K_2 \times L)^2$ and $C_3 = C_2/2$ (where $K_1 = 0.01$, $K_2 = 0.03$, $L = 255$)

B. Inter Mode Decision in H.264/AVC

As specified in H.264/AVC, there are 7 different block sizes (16×16, 16×8, 8×16, 8×8, 8×4, 4×8, 4×4) that can be used in interframe motion estimation and compensation. These different block sizes actually form a two-level hierarchy inside a macroblock. The first level includes block size of 16×16, 16×8, and 8×16. Each of the subdivided regions is a MB partition. In the second level, the MB is specified as P8×8 type, each 8×8 block can be one of the subtypes such as 8×8, 8×4, 4×8 or 4×4 (each of them is known as sub-MB partition). This method of partitioning MBs into motion compensated subblocks of varying size is known as tree structured motion compensation.

Currently, in the inter mode RDO of H.264/AVC, motion estimation is performed by choosing a mode with minimum Lagrangian RDcost as the best mode. The procedure can be defined as follows:

$$\begin{aligned} &J(s,c,MODE|QP) \\ &= D(s,c,MODE|QP) + \lambda_{Lagrangian} R(s,c,MODE|QP) \end{aligned} \tag{2}$$

where, *MODE* indicates a macroblock mode, which can be any one of 16×16, 16×8, 8×16 or P8×8. QP is the quantization parameter; $D(s,c,MODE|QP)$ is the SSD between original block s and reconstructed block c; $R(s,c,MODE|QP)$ is the bit number of the encoding MB which is associated with the chosen *MODE*, QP and the reconstruct macroblock (There is also a SKIP mode in P slice referring for the 16×16 mode, where no motion and residual information is encoded); $\lambda_{Lagrangian}$ is the Lagrangian multiplier, which is quite important in RDO and is described in the following subsection.

C. The Lagrangian multiplier $\lambda_{Lagrangian}$

Basically, the statistic model for inter mode RDO can be written as follows:

$$J_{\cos t} = D_{SSD} + \lambda_{Lagrangian} \cdot R_{MB} \tag{3}$$

where D_{SSD}, R_{MB} are represent the expectations of the distortion $D(s,c,MODE|QP)$ and the bit rate $R(s,c,MODE|QP)$ independently, and $J_{\cos t}$ is the RDcost.

Supposing R_{MB} and D_{SSD} are differentiable everywhere, the minimum of the Lagrangian cost $J_{\cos t}$ is given by setting its derivative to zero, i.e.

$$\lambda_{Lagrangian} = -\frac{\partial D_{SSD}}{\partial R_{MB}} \tag{4}$$

On the relationship formula between R_{MB} and D_{SSD}, a typical approximation curve for entropy-constrained scalar quantization can be written as the equation (5) [10].

$$R_{MB} = a \log_2 (\frac{b}{D_{SSD}}) \tag{5}$$

where a and b are two constants; For the D_{SSD} model, the distortion of one macroblock is defined as follows:

$$D_{SSD} = k \cdot E[D] = 128 \cdot 2^{(QP-12)/3} \tag{6}$$

where, k is a constant, it is equal to the number of pixels in the encoding macroblock (in this paper, the video source sampling in the experiment is supposed to be 4:2:0, and the total pixels in the encoding macroblock and the constant k are both equal to 384); $E[D]$ is the expectation of distortion of one pixel, the source probability distortion can be approximated as uniform within each quantization interval [10], and it is equal to $\frac{2^{(QP-12)/3}}{3}$.

Substitute (5) and (6) into (4), the $\lambda_{Lagrangian}$ can be derived as

$$\lambda_{Lagrangian} = c \cdot 2^{(QP-12)/3} \tag{7}$$

where, c is a constant which is experimentally suggested equal to be 0.85 [11].

III. IMPROVED RDO BASED ON SSIM

A. IRDO-SSIM

As described in the previous section, distortion measurement plays an important role in rate-distortion optimization. However, SAD and SSD both used in H.264/AVC are proved not correlating well with HVS. In the proposed method, SSIM rather than SSD is adopted as the distortion metric in RDO. Considering that using SSIM in motion estimation would greatly increase algorithm complexity, SAD is still used in motion estimation. Therefore, the major steps of selecting the best inter prediction mode and the best matching block(s) for each macroblock in the proposed IRDO-SSIM are summarized as follows:

Step 1. Choose the best matching block(s) for each inter prediction mode:

This part of the algorithm remains the same as H.264/AVC, and the cost function is shown as follows:

$$MCOST(s,c) = SA(T)D(s,c) + \lambda_{MOTION} Bit(\Delta MV) \quad (8)$$

Step 2. Choose the best prediction mode for the macroblock:

For each prediction mode and its best matching block(s), RDcost is calculated and the prediction mode with minimum RDcost is chosen as the best mode. The cost function is defined as follows:

$$J(s,c,MODE|QP) = \lambda(1 - SSIM(s,c)) + R(s,c,MODE|QP) \quad (9)$$

Comparing with (2), the multiplier is attached to the distortion term in (9) because the distortion calculated by $(1\text{-}SSIM(s,c))$ is much smaller than the bit number calculated by $R(s,c,MODE|QP)$. In equation (9), $SSIM(s,c)$ is the structure similarity between the original macroblock s and reconstructed macroblock c. And the SSIM-rate multiplier λ is most important in the cost function, and its detailed derivation process is described in the following subsection.

B. Derivation of SSIM-rate Multiplier λ

Basically, the equation (9) can be rewritten as follows:

$$J = \lambda \cdot D_{FSSIM} + R \quad (10)$$

where D_{FSSIM} represents the expectation of distortion $(1\text{-}SSIM(s,c))$, R represents the expectation of the bit number needed to encode one macroblock, and J is the RDcost.

Supposing R and D_{FSSIM} can be differentiable everywhere, the minimum of the RDcost J is given by setting its derivative to zero, i.e.

$$\lambda = -\frac{\partial R}{\partial D_{FSSIM}} \quad (11)$$

Equation (11) indicates that λ corresponds to the negative slope of the rate-distortion curve, which means that λ can be perfectly determined by the models of R and D_{FSSIM}. R and D_{FSSIM} are related to QP, the expression (11) can be turned to:

$$\lambda = -\frac{\partial R}{\partial D_{FSSIM}} = -\frac{\partial R / \partial QP}{\partial D_{FSSIM} / \partial QP} \quad (12)$$

In the following part of this subsection, the derivation of the expressions $\partial R / \partial QP$ and $\partial D_{FSSIM} / \partial QP$ will be described separately.

1) Derivation of $\partial R / \partial QP$

In fact, the encoded bit number for a macroblock R is only related to the chosen prediction mode, quantization step and the matched macroblock. It is independent of the quality metric used in RDO. This means that, the R model remains the same, while SSIM is used as distortion metric instead of SSD. So, the model of the bit-rate R is the same as R_{MB} which is given in the equation (5), and it is defined as follows:

$$R = R_{MB} = a \log_2 (\frac{b}{D_{SSD}}) \quad (13)$$

Substituting (6) into (13), the expression $\partial R / \partial QP$ can be derived as:

$$\frac{\partial R}{\partial QP} = -\frac{a}{3} \quad (14)$$

Equation (15) indicates that the value of $\partial R / \partial QP$ relate only to the constant a. Substitute (5), (6), and (7) into (4), the constant a can be derived， and it is approximately equal to 104.4. So:

$$\frac{\partial R}{\partial QP} = -\frac{a}{3} = -\frac{104.4}{3} = -34.8 \quad (15)$$

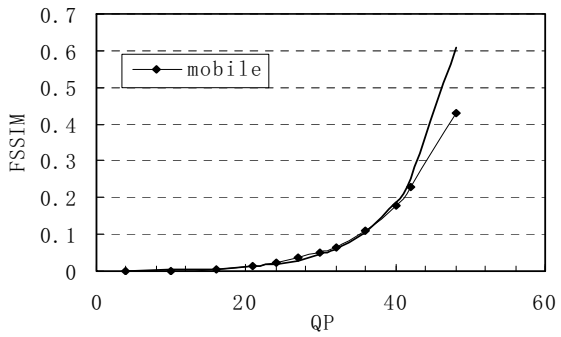

Figure 1. The distortion (*FSSIM*) of each frame in the sequence *mobile*, and the function (**17**)

Equation (15) indicates that the value of $\partial R/\partial QP$ is a constant. And the expression of $\partial D_{FSSIM}/\partial QP$ models can be derived by exponential approximation, which is described in the following part of this subsection.

2) Derivation of $\partial D_{FSSIM}/\partial QP$

In order to obtain the expression of $\partial D_{FSSIM}/\partial QP$, the $D_{FSSIM}(QP)$ model in RDO is necessary. However, it is difficult to derive the theoretical expression of $D_{FSSIM}(QP)$ model since SSIM expression is complex. In this paper, the $D_{FSSIM}(QP)$ model is derived by experiment.

Intensive experiments tell us that the sequences with high-detail regions and high motion complexity include plenty blocks types, so it is suitable to choose them as training sequences in deriving $D_{FSSIM}(QP)$ expression process. In this paper, the sequence of Mobile is used to derive $D_{FSSIM}(QP)$ expression. It is encoded by H.264/AVC reference software JM11.0 [12], and the average distortion of the reconstructed macroblocks in the sequence is measured by FSSIM, which is defined as follows:

$$FSSIM = 1 - SSIM \tag{16}$$

where, SSIM is calculated by equation (1).

Figure 1 shows the distortion (FSSIM) of each frame in the sequence， and the bold curve in Figure 1 depicts the function

$$D_{FSSIM} = 1 \times 10^{-4} \cdot e^{\frac{QP + 11.804}{6.8652}} \tag{17}$$

which is an approximation of the relationship between the macroblock quantizer value QP and the distortion D_{FSSIM}.

Substituting equations (15) and (17) into equation (12), the final λ can be determined as

$$\lambda = -\frac{\partial R/\partial QP}{\partial D_{FSSIM}/\partial QP} = 2.39 \cdot e^{\frac{QP + 11.804}{6.8652}} \tag{18}$$

IV. EXPERIMENT AND RESULTS ANALYSIS

A. Experimental environment

The proposed algorithm is implemented by modifying the H.264/AVC reference software JM11.0 [12]. For all tests, 5 reference frames and full search motion estimation are used for inter prediction, and the ME search window is set as 16×16. In order to compare the performance between the proposed algorithm IRDO-SSIM and the original H.264 in inter coding, intra mode coding is forbidden in inter frame coding in both algorithm.

B. Experimental Results

For that Mean SSIM (MSSIM) is better to assess Quantization Distortion than PSNR [7], MSSIM is proposed to applied to assess the reconstructed video quality. It is measured frame by frame, and then the average MSSIM of all frames is considered as score of the whole sequence quality.

In addition, MSSIM of each frame is obtained by averaging all the 8×8 sliding windows, and the SSIM of the sliding 8×8 widow is calculated as follows:

$$SSIM = 0.6 \times SSIM_y + 0.2 \times SSIM_u + 0.2 \times SSIM_v \tag{19}$$

where, $SSIM_y$, $SSIM_u$, and $SSIM_v$ represent the SSIM of the component y, u and v of the current block respectively, which is calculated by equations (1).

The coding performance is compared in terms of output rate-distortion curve of the reconstructed videos. Rate is measured by bit/pic which is the average bit number per picture, and is obtained by averaging all the P-frames' bit numbers. Distortion is represented by MSSIM which is the average of all the P-frames' SSIM. The comparison results are showed in Figure 2. In these figures, the label "IRDO-SSIM" represents the method proposed in this paper, and the label "H.264" represents the original method.

As shown in these figures, no matter which type of sequence (high, medium or low motion complexity), the proposed algorithm has a better rate-distortion performance than H.264/AVC. However, the gain varies from one sequence type to another. It is due to the fact that, SAD is still used as the distortion metric in motion estimation process in our proposed IRDO-SSIM. This means that the reconstructed macroblock obtained by each inter-prediction mode remains the same of the conventional H.264/AVC encoding. Thus, the rate-distortion gain is only obtained by better mode selection. For the low-motion complexity sequences, most of the macroblocks are encoded by SKIP or 16x16 mode, and for the high-motion complexity ones, most of the macroblocks are encoded by P8x8 mode, no matter which algorithm is used. It means that, the encoding mode of

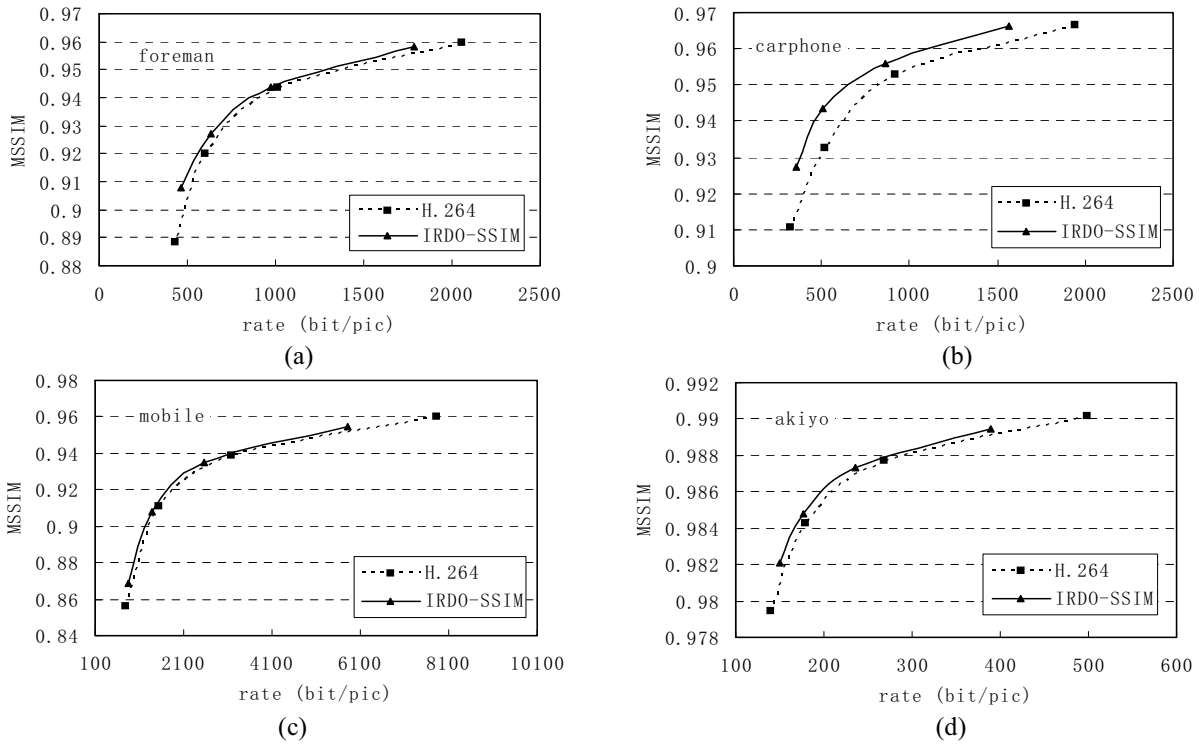

Figure 2. FSSIM vs. bit-rate in bit/pic with various *QP* for the video sequences foreman (a), carphone (b), mobile(c), akiyo(d)

a macroblock is relatively fixed in high or low-motion complexity sequence coding.

V. CONCLUSION

In this paper, an improved Rate-Distortion Optimization based on SSIM (IRDO-SSIM) is proposed, where a new rate-distortion cost function with SSIM as the distortion metric is proposed, and the important SSIM-rate multiplier parameter λ is also derived. According to simulation, the proposed IRDO-SSIM algorithm outperforms the current RDO in the reference software of H.264/AVC, and a gain up to 0.015 in SSIM score is observed. Since SAD is still used in motion estimation process, the computation load of the proposed IRDO-SSIM just increases a little. On the other hand, using SAD in motion estimation limits the gain of the rate-distortion performance for high- or low-motion complexity sequences, and it also limits the further gain in high-rate encoding. Improving the coding performance in the conditions of high-motion complexity sequences coding and high-rate coding will be studied in our following work.

ACKNOWLEDGMENT

The work described in this paper was substantially supported by research project from National Natural Science Foundation of China [Project No. 60402015], and research project from Guangdong Natural Science Foundation of China [project No. 06025642]

REFERENCES

[1] W. Ding and B. Liu, "Rate control of MPEG video coding and recording by rate quantization modeling," IEEE Trans. Circuits Syst. Video Technol., vol. 6, no. 2, pp. 12–20, Feb. 1996.

[2] H. M. Hang and J. J. Chen, "Source model for transform video coder and application-part I: Fundamental theory," IEEE Trans. Circuits Syst. Video Technol., vol. 7, no. 4, pp. 287–298, Apr. 1997.

[3] L. Chen and I. Garbacea, "Adaptive lambda estimation in Lagrangian rate-distortion optimization for video coding," presented at the Visual Commun. Image Process. (VCIP), San Jose, CA, Jan. 2006.

[4] J. Zhang, X. Yi, N. Ling, and W. Shang, "Chroma coding efficiencyimprovement with context adaptive Lagrange multiplier (CALM)," inProc. IEEE Int. Symp. Circuits Systems (ISCAS), New Oreleans, LA, pp. 293–296. May 2007.

[5] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace Distribution Based Lagrangian Rate Distortion Optimization for Hybrid Video Coding,"IEEE Trans. Circuits Syst. Video Technol., vol. 19, pp. 193–205, Feb. 2009.

[6] J.L.Mannos, J.D.Sakrison, "The effects of a visual fidelity criterion on the encoding of images," In IEEE Trans. Information Theory, no.4, pp. 525-536, 1974.

[7] Z. Wang, A.C.Bovik, H.R.Sheikh and E.P.Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Processing, vol. 13, No. 4, pp. 600-612, April 2004.

[8] Chun-Ling Yang, Hua-Xing Wang and Lai-Man Po, "Improved Inter Prediction based on Structural Similarity in H.264," 2007 IEEE International Conference on Signal Processing and Communications (ICSPC 2007) pp340-343, Nov. 2007.

[9] Wang, A.C. Bovik, AUniverse Image Quality Index. IEEE Signal Processing letters. 2002, 9(3): 81-84.

[10] H. Gish and J. pierce, "Asymptotically efficient quantizing", IEEE Trans. Inf. Theory, vol. 14, no.5, pp. 676-683.1968.

[11] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control", in IEEE Int.Conf. on Image Processing, pp.542-545, vol.3. 2001.

[12] http://iphome.hhi.de/suehring/tml/download