SSIM-BASED RATE-DISTORTION OPTIMIZATION IN H.264

Wei Dai[#], Oscar C. Au[#], Wenjing Zhu[#], Pengfei Wan[#], Wei Hu[#], Jiantao Zhou[$]

[#] Hong Kong University of Science and Technology, [$] University of Macau

ABSTRACT

In the current video coding standards, rate-distortion optimization (RDO) plays an important role in achieving best tradeoff between the perceived distortion and transmission rate. It is widely used in all kinds of encoder decisions, including block mode decision, motion vector selection and so on. Generally, the sum of absolute difference (SAD) or the sum of square difference (SSD) is used as the distortion measurement. However, it is well known that both of them cannot always reflect the perceptual quality of the encoded video. In this paper, an objective quality measurement structural similarity (SSIM) index is proposed as the distortion measurement in the RDO framework for video coding standards. By fully exploiting the relationship between SSIM and mean square error (MSE), the SSIM-based RDO framework can be approximated by the original SSD-based RDO framework with only a scaling of the Lagrange multiplier. Experimental results show that the proposed method outperforms the latest H.264 codec and also the state-of-the-art SSIM-based RDO video codec.

Index Terms— SSIM, video coding, rate-distortion optimization

1. INTRODUCTION

In the traditional hybrid video codecs, rate-distortion optimization (RDO) is introduced to make decisions which lead to the best performance. The goal of RDO is to minimize the perceived distortion with the number of encoded bits subject to a rate constraint [1, 2]:

$$\begin{aligned} \min_{\mathbf{\Omega}} \quad & D(\mathbf{\Omega}) \\ \text{s.t.} \quad & R(\mathbf{\Omega}) \leq R_c, \end{aligned}$$

where $\mathbf{\Omega}$ represents the set of encoder decisions for the block, $D(\mathbf{\Omega})$ and $R(\mathbf{\Omega})$ are the distortion and rate measurement using $\mathbf{\Omega}$ respectively. R_c is the rate constraint. In real applications, this constraint problem is reformulated into an unconstraint problem using Lagrange optimization method [1, 2], which can be expressed as:

$$\min_{\mathbf{\Omega}} J(\mathbf{\Omega}, \lambda) = D(\mathbf{\Omega}) + \lambda R(\mathbf{\Omega}), \tag{1}$$

where λ is the Lagrange multiplier which controls the tradeoff between rate and perceived distortion. Generally, the sum of absolute difference (SAD) or the sum of square difference (SSD) will be used as the distortion measurement of $D(\mathbf{\Omega})$.

However, SAD and SSD are widely criticized for not correlating well with perceived quality. Recently, a lot of works have been done to develop objective quality assessments which can accurately reflect the perceived distortion. Several promising algorithms including the structural similarity (SSIM) index [3], visual signal-to-noise ration [4] and visual information fidelity criterion [5] were proposed to deal with this problem. Among these algorithms, SSIM has been preferred due to its accuracy, simplicity and efficiency [6]. SSIM and its derivations have been applied to a broad range of applications, ranging from image restoration and compression, to visual communication and pattern recognition [6].

In order to improve the perceptual video coding performance, a lot of efforts have be incorporate to introduce SSIM index into the RDO framework to characterize the video distortion. Most of them used (1-SSIM) as the distortion measurement. Wang *el al.* proposed a SSIM-QP model and a model for rate as a function of residual coefficients statistics for the SSIM-based RDO of H.264 [7]. The authors in [8, 9] also used (1-SSIM) as the distortion measurement for H.264, they proposed an algorithm for computing an appropriate Lagrange multiplier for the SSIM-based RDO framework. Mai *et al.* proposed a SSIM-based RDO framework for Intra coding of H.264 [10] and then extended their work to fast Intra mode decision [11] and motion estimation [12]. Instead of using (1-SSIM) as the distortion measurement, 1/SSIM was also utilized as the distortion measurement in the RDO framework [13].

In this paper, we choose 1/SSIM as the distortion measurement of the SSIM-based RDO framework. By fully exploiting the relationship between SSIM and mean square error (MSE), the paper provides a convenient and efficient way of modifying the traditional SSD-based RDO framework into a SSIM-based RDO framework by just scaling the Lagrange multiplier properly.

This work has been supported in part by the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (GRF Project no. 610112), HKUST (FSGRF12EG01 and FSGRF14EG40), the Macau Science and Technology Development Fund under Grants FDCT/009/2013/A1, the Research Committee at University of Macau under Grants SRG023-FST13-ZJT and MRG021/ZJT/2013/FST.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

The rest of this paper is organized as follows: Section 2 gives a briefly introduction on SSIM index and the detailed analysis of the relationship between SSIM and MSE is also provided. The proposed algorithm is described in Section 3 and experimental results are shown in Section 4. Finally, Section 5 concludes the paper.

2. SSIM AND ITS APPROXIMATION USING MSE

In this section, the basic idea of SSIM index is introduced in Section 2.1. Then, the relationship between SSIM and MSE is briefly investigated in Section 2.2.

2.1. SSIM Index

Based on the assumption that the Human Visual System model is highly adapted for extracting structural information, the SSIM index assesses three terms between the two image blocks x and y, which are luminance $l(x,y)$, contrast $c(x,y)$ and structure $s(x,y)$ [3]:

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \\ c(x,y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \\ s(x,y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \end{aligned}$$

where μ_x and μ_y are mean of x and y respectively, σ_x^2 and σ_y^2 are the variance of x and y respectively, σ_{xy} is the covariance between x and y. c_1, c_2 and c_3 are some constants which provide spatial masking properties and ensure stability with weak denominator. In general, $c_1 = \kappa_1 L$ and $c_2 = \kappa_2 L$, where L is the dynamic range of the pixel value, κ_1 and κ_2 are set to be 0.01 and 0.03 by default. For c_3, we simply set $c_3 = c_2/2$. Combing three terms together, the general form of SSIM is:

$$\begin{aligned} \text{SSIM} &= l(x,y)\cdot c(x,y)\cdot s(x,y) \\ &= \left(\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\right)\left(\frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}\right). \end{aligned} \quad (2)$$

The SSIM index of the whole image is obtained by averaging the local SSIM indices using a sliding window. If we denote x to be the original image block and y to be the reconstructed image block, SSIM can be regarded as the distortion quality measurement.

2.2. The relationship between SSIM and MSE

In [13], the author simply modeled the relationship between the reconstructed pixel y and the original pixel x by an additive distortion model, *i.e.* $y = x + e$, where e is the reconstruction error due to the lossy quantization. However, this additive model cannot fully describe the relationship between y and x in all the situations. Because the quantization process in the video mainly removes the high frequencies of the block to achieve compression, which makes y more likely to be a low-pass filtered version of x. In this paper, we use a five-tap low-pass filter to filter the original pixel x and its four neighboring pixels (up x_1, bottom x_2, left x_3 and right x_4) to approximate the reconstructed pixel y. The model between y and x can be represented as:

$$y = \mathbf{H}^T\mathbf{X} + e, \quad (3)$$

where $\mathbf{X} = [x, x_1, x_2, x_3, x_4]^T$ and $\mathbf{H} = [h_0, h_1, h_2, h_3, h_4]^T$ are the corresponding low-pass filter coefficients, which satisfies $\sum_{i=0}^{4} h_i = 1$, e is the zero mean noise which is independent to x and y.

On the other hand, MSE can be computed as

$$\begin{aligned} \text{MSE} &= E((y-x)^2) \\ &= E(y^2) + E(x^2) - 2E(xy). \end{aligned} \quad (4)$$

It can be easily verified from (3) that $\mu_y = \mu_x$. By considering equation (4), (2) can be rewritten as:

$$\begin{aligned} \text{SSIM} &= \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ &= \frac{2\sigma_{xy} + c_2}{E(x^2) - \mu_x^2 + E(y^2) - \mu_y^2 + c_2} \\ &= \frac{2\sigma_{xy} + c_2}{E(x^2) + E(y^2) - 2\mu_x\mu_y + c_2} \\ &= \frac{2\sigma_{xy} + c_2}{\text{MSE} + 2E(xy) - 2\mu_x\mu_y + c_2} \\ &= \frac{2\sigma_{xy} + c_2}{\text{MSE} + 2\sigma_{xy} + c_2}. \end{aligned} \quad (5)$$

In this paper, we define the SSIM-based distortion measurement as follows:

$$\begin{aligned} \text{dSSIM} &= \frac{1}{\text{SSIM}} \\ &= 1 + \frac{\text{MSE}}{2\sigma_{xy} + c_2}. \end{aligned} \quad (6)$$

Equation (6) gives a convenient and accurate relationship between dSSIM and MSE of the original block x and its reconstructed block y. However, the reconstructed block y is not available during the RDO process, we need to estimate σ_{xy}, with only the information of x. Since $\mu_y = \mu_x$, we can make x and y to be zero mean without affecting σ_{xy} for simplicity. If the correlation coefficient between x and x_i is ρ_i,

$i = 1, \ldots, 4$, then σ_{xy} can be rewritten using (3), which is:

$$\begin{aligned}\sigma_{xy} &= E(xy) \\ &= (\sum_{i=0}^{4} h_i\rho_i)E(x^2) \\ &= (\sum_{i=0}^{4} h_i\rho_i)\sigma_x^2 \\ &= \beta\sigma_x^2,\end{aligned}$$

where $\rho_0 = 1$ and $\beta = \sum_{i=0}^{4} h_i\rho_i$. So once we figure out the low-pass filter coefficients h_i and the correlation coefficients ρ_i, we can estimate the covariance σ_{xy} between the original block x and its reconstructed block y with only the information of x. In this paper, h_i and ρ_i are estimated on the frame by frame basis.

3. THE PROPOSED SSIM-BASED RDO FRAMEWORK

Based on the relationship between dSSIM and MSE, the new SSIM-based RDO framework is proposed in this section.

3.1. Objective Function

Recall that the original RDO framework is done by optimizing the Lagrangian cost in (1) (Here we use SSD as the distortion measurement):

$$J = \text{SSD} + \hat{\lambda}R = N \times \text{MSE} + \hat{\lambda}R,$$

for an appropriate $\hat{\lambda}$, where N is the number of the pixels in the block. If we use the dSSIM as the distortion measurement in the RDO framework, we want to minimize dSSIM under the rate constraint. As illustrated in the previous sections, the problem can be reformulated into an unconstrained minimization problem. For each block b, the objective function of SSIM-based RDO framework is formulated as follows:

$$\begin{aligned}J &= N\text{dSSIM} + \lambda R \\ &= N(1 + \frac{\text{MSE}}{2\beta\sigma_x^2(b) + c_2}) + \lambda R \\ &= N + \frac{\text{SSD}}{2\beta\sigma_x^2(b) + c_2} + \lambda R \\ &= N + \frac{1}{2\beta\sigma_x^2(b) + c_2}(\text{SSD} + (2\beta\sigma_x^2(b) + c_2)\lambda R).\end{aligned}$$

Equivalently, we can also optimize the following equation for each block b:

$$J = \text{SSD} + (2\beta\sigma_x^2(b) + c_2)\lambda R, \tag{7}$$

for an appropriate λ. Equation (7) offers a very convenient way to incorporate SSIM index into the RDO framework. With just a modification of the Lagrange multiplier, the original SSD-based RDO framework becomes the SSIM-based RDO framework. For different blocks within the same frame, we only need to scale the Lagrange multiplier according to the local characteristics of the block.

There exists an intuitive explanation for this process, compared to the smooth region, a texture region can tolerate a larger SSD with no significant perceptual quality loss. The term $\beta\sigma_x^2$ exactly measures the local texture property of the block.

3.2. λ Selection

Another important issue is how to choose the appropriate Lagrange multiplier λ. In this paper, the selection of λ is quite similar compared to [13], which wants to keep the overall rate of encoding one frame to be the same. In H.264, the RD model for each macroblock (MB) is:

$$\frac{R(D)}{N} = \alpha \log(\frac{\sigma^2}{D/N}), \tag{8}$$

where σ^2 is the variance of the difference in the MB, D is the SSD of the MB. To solve (1), we take the derivative of D for each block b:

$$\frac{\partial J(b)}{\partial D(b)} = 1 + \hat{\lambda}\frac{\partial R(b)}{\partial D(b)} = 0. \tag{9}$$

Using (8) in (9), we get:

$$\begin{aligned}D^*(b) &= N\alpha\hat{\lambda}, \\ R^*(b) &= N\alpha \log(\frac{\sigma^2(b)}{\alpha\hat{\lambda}}),\end{aligned}$$

where $D^*(b)$ and $R^*(b)$ are the optimal SSD and rate for the b-th MB respectively, $\sigma^2(b)$ is the variance of the SSD for the b-th MB. So the total rate of the frame is:

$$R_{\text{SSD}} = N\alpha \sum_{b=1}^{M} \log(\frac{\sigma^2(b)}{\alpha\hat{\lambda}}),$$

where M is the number of MBs in one frame.

Similar procedure is taken for the SSIM-based RDO framework (7), we can get:

$$\begin{aligned}D^*(b) &= (2\beta\sigma_x^2(b) + c_2)N\alpha\lambda, \\ R^*(b) &= N\alpha \log(\frac{\sigma^2(b)}{\alpha(2\beta\sigma_x^2(b) + c_2)\lambda}).\end{aligned}$$

So the total rate is:

$$R_{\text{SSIM}} = N\alpha \sum_{b=1}^{M} \log(\frac{\sigma^2(b)}{\alpha(2\beta\sigma_x^2(b) + c_2)\lambda}).$$

Since we want $R_{\text{SSD}} = R_{\text{SSIM}}$, we can derive the relationship between $\hat{\lambda}$ and λ as (assuming that the statistics $\sigma^2(b)$

Table 1. Performance Comparison (BD-Rate (SSIM)) of the proposed algorithm with JM 18.4 and the method in [13].

Resolution	Sequence Name	JM 18.4	Method in [13]
4K	Traffic	-21.3%	-0.2%
	PeopleOnStreet	-11.6%	0.3%
1080p	ParkScene	-12.0%	-2.5%
	Cactus	-2.0%	-0.7%
WVGA	PartyScene	-14.1%	-1.7%
	BQMall	-10.6%	-1.3%
WQVGA	RaceHorses	-18.4%	-2.3%
	BlowingBubbles	-13.7%	-0.7%
Average		**-13.0%**	**-1.1%**

remains the same whether SSE or SSIM is used as the distortion measurement):

$$\lambda = \hat{\lambda} \exp(-\frac{1}{M} \sum_{b=1}^{M} \log(2\beta\sigma_x^2(b) + c_2)).$$

This means for the b-th MB, the Lagrange multiplier is:

$$\lambda_b = \frac{2\beta\sigma_x^2(b) + c_2}{\exp(\frac{1}{M}\sum_{b=1}^{M} \log(2\beta\sigma_x^2(b) + c_2))} \hat{\lambda}.$$

4. EXPERIMENTAL RESULTS

The proposed algorithm is implemented in the latest H.264 reference software JM 18.4. IPPP structure is used and several QPs are tested. The proposed algorithm is tested using various sequences with different resolutions and properties. BD-Rate [14] is regarded as the performance measurement. A negative value of BD-Rate implies that the proposed approach brings coding gain while a positive value means coding loss. The BD-Rate value can be interpreted as the average rate decrease/increase with respect to the baseline while maintaining the same SSIM quality. The original SSD-based RDO framework and the state-of-the-art SSIM-based algorithm proposed in [13] are tested as the comparative algorithms. The simulation results are shown in Table 1. One example of the RD curve is also plotted in Fig. 1.

From Table 1, we can conclude that the proposed SSIM-based RDO framework outperforms the JM 18.4 reference software by 13.0% BD-Rate reduction on average, with only a scaling of the Lagrange multiplier. Moreover, the proposed algorithm outperforms the method in [13] by 1.1% BD-Rate reduction on average. This is because the proposed algorithm gives a more accurate model between the reconstructed block y and the original block x, especially for the low bit-rate region. As for high bit-rate region, since the high frequencies are mainly kept, the equation (3) becomes almost the same

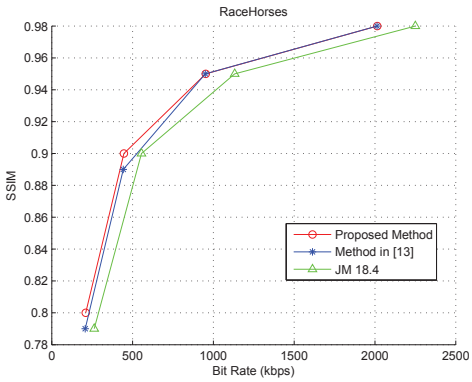

Fig. 1. RD curve comparison of the proposed algorithm with the algorithm in [13] and JM 18.4.

compared to [13]. It can also be verified from Fig. 1 that for high bit-rate region, the coding performance of the proposed algorithm is comparable with the algorithm in [13] but is better than [13] in low bit-rate region.

5. CONCLUSION

In this paper, a SSIM-based RDO framework is proposed. By fully investigating the relationship between SSIM and MSE, the proposed SSIM-based RDO framework can be modified by simply scaling the Lagrange multiplier of the original SSD-based RDO framework. Simulation results show that the proposed algorithm outperforms the original SSD-based RDO framework by 13.0% BD-Rate reduction and outperforms the state-of-the-art SSIM-based RDO framework by 1.1% BD-Rate reduction.

6. REFERENCES

[1] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *Signal Processing Magazine, IEEE*, vol. 15, no. 6, pp. 23–50, 1998.

[2] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *Signal Processing Magazine, IEEE*, vol. 15, no. 6, pp. 74–90, 1998.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[4] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, 2007.

[5] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[6] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98–117, 2009.

[7] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Rate-SSIM optimization for video coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 833–836.

[8] H. H. Chen, Y.-H. Huang, P.-Y. Su, and T.-S. Ou, "Improving video coding quality by perceptual rate-distortion optimization," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1287–1292.

[9] Y.-H. Huang, T.-S. Ou, and H. H. Chen, "Perceptual-based coding mode decision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE, 2010, pp. 393–396.

[10] Z.-Y. Mai, C.-L. Yang, L.-M. Po, and S.-L. Xie, "A new rate-distortion optimization using structural information in H.264 I-frame encoder," in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2005, pp. 435–441.

[11] Z.-Y. Mai, C.-L. Yang, and S.-L. Xie, "Improved best prediction mode(s) selection methods based on structural similarity in H.264 I-frame encoder," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. IEEE, 2005, vol. 3, pp. 2673–2678.

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

[13] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using SSIM," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 7, pp. 1170–1181, 2013.

[14] G. Bjøntegaard, "Improvements of the BD-PSNR model," *document VCEG-AI11, ITU-T SG16*, 2008.