ON RATE DISTORTION OPTIMIZATION USING SSIM

Chuohao Yeo, Hui Li Tan, Yih Han Tan

Signal Processing Department
Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632

ABSTRACT

Rate-distortion optimization is widely used in modern video codecs to make various encoder decisions in order to optimize the rate-distortion trade-off. Typically, the distortion measure used is either sum-of-square error (SSE) or sum-of-absolute distance (SAD), both of which are convenient when used in the RDO framework but not always reflective of perceptual quality. In this paper, we show that by expressing SSIM in terms of SSE, SSIM can be used as the distortion metric in the RDO framework in an effective and efficient manner by simply scaling the Lagrange multiplier used in RDO based on the local variance in that region without further changes to the RDO engine. Experimental results show that compared to traditional RDO approaches, for the same SSIM score, the proposed approach can achieve an average rate decrease of 8% and 11% for random access and low-delay encoding configurations, with no significant change in encoding runtime.

Index Terms— SSIM, Perceptual based coding, video coding, rate distortion optimization

1. INTRODUCTION

Traditionally in hybrid video coding, rate distortion optimization (RDO) is used to choose the best operational point for each coding block. When this is done, the sum-of-square error (SSE) or sum-of-absolute distance (SAD) is usually used as the objective function, and so the optimization problem is of the form: minimize SSE/SAD while satisfying a given rate constraint.

However, it is well known that both SSE and SAD are not good measures of perceptual quality. While many other metrics, including the structural similarity index (SSIM) introduced by Wang et al. [1], have been proposed, a straightforward way to incorporate them within the RDO framework remains elusive despite several previous attempts. Typically, these involve directly replacing the SSE with (1-SSIM) as the distortion function, and using an empirical or model-based approach for estimating the Lagrange multiplier to be used within the RDO process [2–5]. Another approach involves maximizing the minimum SSIM using a bit-plane based image coder over multiple coding iterations [6]. Alternatively, instead of directly optimizing a perceptual metric, a locally varying perceptual-based lagrange multiplier is used for RDO in each local region [7]; however, the scaling is done using heuristics.

In this paper, we describe a simple approach that uses SSIM effectively within the RDO framework. Essentially, this can be achieved by scaling the Lagrange multiplier used in RDO for each local region by the local variance in that region.

2. ANALYSIS

2.1. Relationship between SSIM and MSE

The SSIM between two image regions is defined as [1]:

$$\text{SSIM} = \left(\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\right)\left(\frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_1}\right) \tag{1}$$

where x and y are the two image regions to be compared, and $c_1 = (\kappa_1 L)^2$ and $c_2 = (\kappa_2 L)^2$ are two constants used for numerical stability. $\kappa_1 = 0.01$ and $\kappa_2 = 0.03$, and L is the peak value of the image. Here, we would denote the original image by x and the reconstructed image by y.

We use an additive distortion model for y, i.e., $y = x + e$, where e is the reconstruction error due to lossy quantization. We will assume that e is a random variable with zero mean and variance σ_e^2. Note that MSE can be computed as

$$\text{MSE} = \frac{1}{N}\Sigma_i \left(y_i - x_i\right)^2 = \frac{1}{N}\Sigma_i e_i^2 \tag{2}$$

where N is the number of pixels in the region, and the index i denotes individual pixels within a region. From the law of large numbers, as N gets large, $MSE \to \sigma_e^2$.

Now, we can also compute each of the terms in SSIM. It is easily verified that $\mu_y = \mu_x$, $\sigma_y^2 = \sigma_x^2 + \sigma_e^2$ and $\sigma_{xy} = \sigma_x^2$. By substituting these into (1), we can simplify the expression for SSIM into:

$$\text{SSIM} = \frac{2\sigma_x^2 + c_2}{2\sigma_x^2 + \sigma_e^2 + c_2} \tag{3}$$

Since all the quantities in (3) are positive, under these assumptions, $0 < \text{SSIM} \leq 1$, and we can also define a distortion

metric based on SSIM as follows:

$$\begin{aligned} \text{dSSIM} &= \frac{1}{\text{SSIM}} \\ &= 1 + \frac{\sigma_e^2}{2\sigma_x^2 + c_2} \\ &\approx 1 + \frac{\text{MSE}}{2\sigma_x^2 + c_2} \end{aligned} \tag{4}$$

where the last line follows from (2) for reasonably large values of N. Note that under this assumption, dSSIM ≥ 1.

(4) gives a convenient relationship between SSIM and MSE that can be used for RDO decisions. It also has an intuitive perceptual meaning, in that the perceptual distortion is the MSE scaled by the inverse variance of the local region. Therefore, for the same perception of visual degradation, the MSE can be higher in a textured region compared to a smooth region.

2.2. Using SSIM in RDO

Recall that for any block, the RDO decision can be done by optimizing a Lagrangian cost [8, 9]:

$$\tilde{J} = \text{SSE} + \tilde{\lambda} R = N \cdot \text{MSE} + \tilde{\lambda} R$$

for an appropriately chosen Lagrange multiplier $\tilde{\lambda}$. To incorporate SSIM into RDO, we use the dSSIM as defined in (4) to optimize the following:

$$\begin{aligned} J &= N \cdot \text{dSSIM} + \lambda R \\ &= N \left(1 + \frac{\text{MSE}}{2\sigma_x^2 + c_2}\right) + \lambda R \\ &= N + \frac{\text{SSE}}{2\sigma_x^2 + c_2} + \lambda R \\ &= N + \frac{1}{2\sigma_x^2 + c_2} \left(\text{SSE} + \left(2\sigma_x^2 + c_2\right) \lambda R\right) \end{aligned}$$

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

$$J = \text{SSE} + \left(2\sigma_x^2 + c_2\right) \lambda R \tag{5}$$

again for some appropriately chosen Lagrange multiplier λ.

(5) offers us a very convenient way to incorporate SSIM into the RDO decision process. In fact, all that is required is to do a local scaling of λ, depending on the local source variance. This means that the entire RDO machinery can be kept as is, with just a minor modification of the Lagrange multiplier. There is again an intuitive explanation for this procedure. For highly textured regions, additional rate would be penalized more than in a smooth region, which means that a larger SSE can be tolerated.

2.3. Picking λ for SSIM-RDO

While we have earlier shown how to optimize SSIM within the regular RDO mode decision framework, there is still an issue of how to choose an appropriate Lagrange multiplier, λ. Here, we present one possible approach based on keeping the overall rate of coding the frame the same, assuming that the displaced frame difference (DFD) statistics is the same whether MSE and SSIM is to be optimized.

Recall that when MSE is used, the optimization problem is to minimize total distortion subject to a constraint on the total rate, i.e., [8, 9]

$$\min_{\Phi} \text{SSE} = \Sigma_i d_i \text{ s.t. } R = \Sigma_i r_i \leq R_c$$

where Φ denote the set of encoder decisions (e.g., MB mode, QP), d_i is the SSE for the ith MB, and r_i is the rate used for the ith MB. This is solved by using the following unconstrained optimization problem:

$$\min_{\{\phi_i\}_{i=1}^{M}} \tilde{J} = \Sigma_i d_i + \tilde{\lambda} \Sigma_i r_i = \Sigma_i \left(d_i + \tilde{\lambda} r_i\right) \tag{6}$$

where M is the number of MBs, and ϕ_i is the set of encoder decisions for the ith MB. Typically, dependencies between MBs are ignored, and we solve for each MB the following unconstrained problem, $\min_{\phi_i} d_i + \tilde{\lambda} r_i$.

In H.264/AVC [10] JM, the Lagrange multiplier is computed as $\tilde{\lambda} = \beta \cdot 2^{(\text{QP}-12)/3}$ [11]. This is justified by assuming the following rate-distortion model for each MB [8]:

$$\frac{r(d)}{N} = \alpha \log \left(\frac{\sigma^2}{d/N}\right) \tag{7}$$

where σ^2 is the variance of the DFD in the MB, and d is the SSE in the MB. Note that the normalization by N is necessary to use the per-symbol characterization of the RD model even when we consider an entire MB.

To solve (6), we set for each i:

$$\frac{\partial \tilde{J}}{\partial d_i} = 1 + \tilde{\lambda} \frac{\partial r_i}{\partial d_i} = 0 \tag{8}$$

Using (7) in (8), we obtain:

$$\begin{aligned} d_i^* &= N\alpha\tilde{\lambda} \\ r_i^* &= N\alpha \log \left(\frac{\sigma_i^2}{\alpha\tilde{\lambda}}\right) \end{aligned}$$

where d_i^* and r_i^* are the optimal SSE and rate for the ith MB respectively, and σ_i^2 is the variance of the DFD for the ith MB. Therefore, the total rate used is:

$$R_{\text{SSE}} = N\alpha\Sigma_i \log \left(\frac{\sigma_i^2}{\alpha\tilde{\lambda}}\right)$$

We can repeat the same exercise when dSSIM is used as the objective function instead. Using (4), we would optimize

$$\min_{\{\phi_i\}_{i=1}^{M}} J = \Sigma_i \frac{d_i}{2\sigma_{x_i}^2 + c_2} + \lambda \Sigma_i r_i = \Sigma_i \left(\frac{d_i}{2\sigma_{x_i}^2 + c_2} + \lambda r_i\right) \tag{9}$$

where $\sigma^2_{x_i}$ is the local source variance for the ith MB.

To solve (9), we again set for each i:

$$\frac{\partial J}{\partial d_i} = \frac{1}{2\sigma^2_{x_i} + c_2} + \lambda \frac{\partial r_i}{\partial d_i} = 0 \tag{10}$$

Using (7) in (10), we obtain:

$$\begin{aligned} d_i^* &= \left(2\sigma^2_{x_i} + c_2\right) N\alpha\lambda \\ r_i^* &= N\alpha \log\left(\frac{\sigma_i^2}{\alpha\left(2\sigma^2_{x_i} + c_2\right)\lambda}\right) \end{aligned}$$

The total rate used is:

$$R_{\text{SSIM}} = N\alpha\Sigma_i \log\left(\frac{\sigma_i^2}{\alpha\left(2\sigma^2_{x_i} + c_2\right)\lambda}\right)$$

As mentioned, we will pick λ in order for the rate to be the same regardless of whether MSE or SSIM is used, and this will be computed as a function of $\tilde{\lambda}$, the Lagrange multiplier that is used in JM. By setting $R_{\text{SSIM}} = R_{\text{MSE}}$, we obtain:

$$\lambda = \tilde{\lambda} \cdot \exp\left(-\frac{1}{M}\Sigma_{i=1}^{M} \log\left(2\sigma^2_{x_i} + c_2\right)\right)$$

This means that in using (5) to perform RDO decision, we will use for the ith MB a Lagrange multiplier of:

$$\lambda_i = \frac{2\sigma^2_{x_i} + c_2}{\exp\left(\frac{1}{M}\Sigma_{i=1}^{M} \log\left(2\sigma^2_{x_i} + c_2\right)\right)}\tilde{\lambda} \tag{11}$$

This is simply applying a scaling that depends on the local source variance and some source variance statistic computed over the entire frame to the original Lagrange multiplier used in JM. This gives a very concrete way of applying a small modification to the RDO process in order to maximize SSIM over the entire frame.

3. EXPERIMENTAL RESULTS

3.1. Implementation

We implemented this approach in H.264/AVC JM 17.2[1]. Before encoding each frame, we first compute the denominator of the scaling to be applied to the original Lagrange multiplier used in JM, as in (11); this involves computing the variance of the source pixels within each MB. Then, before encoding each MB, the Lagrange multiplier to be used is scaled as in (11).This scaled Lagrange multiplier is then used in all RDO processes such as mode-decision, RD-optimized quantization and motion estimation. No other changes to the encoder software is necessary.

[1]Available from `http://iphome.hhi.de/suehring/tml/download/`

Table 1: BD-Rate (%) results for all-intra setting

Sequence	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time
silence_cif	-5.7%	3.2%	100%
flower_cif	-6.9%	0.5%	99%
bus_cif	-9.3%	2.3%	100%
foreman_cif	-7.0%	2.8%	98%
salesman_qcif	-5.4%	2.6%	100%
carphone_qcif	-3.9%	2.2%	99%
container_qcif	-9.9%	1.7%	99%
Average	-6.9%	2.2%	99%

3.2. Experimental setup

We applied the proposed approach to encoding of 4 CIF test sequences ("silence_cif", "flower_cif", "bus_cif", "foreman_cif") and 3 QCIF test sequences ("salesman_qcif", "carphone_qcif", "container_qcif"). We used 3 different configurations that target different applications: all-intra frame encoding for use in high quality digital cinema application, random access for use in storage applications, and low-delay for use in video conferencing applications. In the all-intra frame encoding, all the frames are encoded as intra pictures without any temporal prediction. For random access configurations, we use a 8-frame hierarchical B-picture structure, with an I-picture approximately every second. In the low-delay configuration, only the first frame is coded as an I-picture, with the rest being P-pictures. The encoding was carried out over a range of QPs: 20, 25, 30 and 35.

To help us understand coding performances, we will show BD-rate [12] figures for both PSNR and SSIM distortion metrics with respect to JM 17.2 without any modifications, but using the same encoding configuration. A negative BD-rate implies that the proposed approach brings coding gains, while a positive BD-rate implies that the proposed approach brings coding loss. These numbers can be interpreted as the average rate decrease/increase with respect to the baseline while maintaining the same PSNR or SSIM quality. We will also show the encoding time of the proposed approach as a percentage of the baseline in order to understand the complexity of the proposed approach.

3.3. Results

Tables 1, 2 and 3 show the results for the all-intra, random access and low-delay encoding configurations respectively.

The key observation is that for the same SSIM, the proposed approach can give significant coding gains ranging from 3% to 18%. The average rate gains for all-intra, random access and low-delay encoding configurations are 6.9%, 7.9% and 11.0% respectively. On the other hand, for the same PSNR, the proposed approach suffers some coding loss of up

Table 2: BD-Rate (%) results for random access setting

Sequence	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time
silence_cif	-4.8%	2.9%	100%
flower_cif	-10.3%	0.8%	99%
bus_cif	-12.0%	1.5%	99%
foreman_cif	-6.3%	2.4%	99%
salesman_qcif	-8.1%	-0.2%	100%
carphone_qcif	-4.9%	3.0%	99%
container_qcif	-9.3%	4.0%	100%
Average	-7.9%	2.1%	99%

Table 3: BD-Rate (%) results for low-delay setting

Sequence	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time
silence_cif	-3.8%	6.2%	100%
flower_cif	-15.3%	-0.3%	100%
bus_cif	-17.7%	-1.9%	100%
foreman_cif	-9.1%	-3.5%	99%
salesman_qcif	-12.7%	5.3%	100%
carphone_qcif	-3.4%	5.3%	100%
container_qcif	-14.6%	2.3%	101%
Average	-11.0%	1.9%	100%

to 5%, with an average loss of 2.2%, 2.1% and 1.9% for all-intra, random access and low-delay encoding configurations respectively. This is to be expected, since the optimization in the proposed approach is done with respect to SSIM, and is no longer optimal with respect to the PSNR metric. Finally, the encoding time of the proposed approach does not show any significant adverse impact, and is about the same as the baseline. This is unlike previous SSIM-based RDO methods, in which the computation of SSIM for all RD cost and the estimation of the Lagrange multiplier would lead to significant increase in encoding time.

4. CONCLUSIONS

In this paper, we have described a simple way to incorporate the use of SSIM into RDO in order to optimize video encoding to target perceptual quality instead of MSE. This can be done by scaling the Lagrange multiplier in RDO based on local statistics. We have implemented the proposed approach and demonstrated that it achieves substantial coding gains of up to 18% while maintaining the same perceptual quality as the H.264/AVC reference software encoder as measured by SSIM. At the same time, there is no increase in encoding complexity. This would be very useful for video encoder practitioners who wish to achieve further compression gains while maintaining the same perceptual quality at low cost.

In our current implementation, we only use the local variance of the luminance component in determining the adjustments. This can be extended to incorporate the local variance of the chrominance components by using a weighted average of the luminance and chrominance components in (11). Furthermore, when we compute the local variance, we might also choose to compute it over a region that extends a number of pixels from the current MB, in order to have more smoothly varying Lagrange mulitplier over the frame.

5. REFERENCES

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

[2] Zhi-Yi Mai, Chun-Ling Yang, and Sheng-Li Xie, "Improved best prediction mode(s) selection methods based on structural similarity in H.264 I-frame encoder," in *IEEE International Conference on Systems, Man and Cybernetics*, Oct 2005, vol. 3, pp. 2673–2678.

[3] Chun-Ling Yang, Rong-Kun Leung, Lai-Man Po, and Zhi-Yi Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, Nov 2009, vol. 4, pp. 291–295.

[4] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, and H H Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1614–1624, Nov 2010.

[5] S Wang, S Ma, and W Gao, "SSIM based perceptual distortion rate optimization coding," in *SPIE Visual Communications and Image Processing Conference*, 2010, vol. 7744.

[6] Zhou Wang, Qiang Li, and Xinli Shang, "Perceptual image coding based on a maximum of minimal structural similarity criterion," in *IEEE International Conference on Image Processing*, Oct 2007, vol. 2, pp. 121–124.

[7] Chang Sun, Hong-Jun Wang, Tai-Hoon Kim, and Hua Li, "Perceptually adaptive lagrange multiplier for rate-distortion optimization in H.264," in *Future Generation Communication and Networking*, Dec 2007, vol. 1, pp. 459–463.

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

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

[10] T Wiegand, G J Sullivan, G Bjontegaard, and A Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003.

[11] T Wiegand, H Schwarz, A Joch, F Kossentini, and G J Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, July 2003.

[12] G. Bjøntegaard, "Calculation of Average PSNR Differences between RD curves," in *ITU-T SC16/Q6, VCEG-M33*, Austin, USA, Apr 2001.