On Rate Distortion Optimization Using SSIM

Chuohao Yeo, *Member, IEEE,* Hui Li Tan, *Member, IEEE,* and Yih Han Tan, *Member, IEEE*

***Abstract*—In this paper, we present a method for performing rate-distortion optimization (RDO) using a perceptual visual quality metric, the structural similarity index (SSIM), as the target of optimization. Rate-distortion optimization is widely used in modern video codecs to make various encoder decisions to optimize the rate-distortion tradeoff. Typically, the distortion measure used is either sum-of-square error or sum-of-absolute distance, both of which are convenient when used in the RDO framework but not always reflective of a perceptual visual quality. We show that SSIM can be used as the distortion metric in the RDO framework in a simple, yet effective, manner by scaling the Lagrange multiplier used in RDO based on the local variance in that region. The experimental results on the H.264/AVC reference software show that compared to traditional RDO approaches, for the same SSIM score, the proposed approach can achieve an average rate reduction of about 9% and 14% for random access and low-delay encoding configurations. At the same time, there is no significant change in the encoding runtime.**

***Index Terms*—Perceptual-based coding, rate-distortion optimization (RDO), structural similarity index (SSIM), video coding.**

I. Introduction

THE AIM of video compression has always been to reproduce the source video with the best possible perceptual visual quality at any specified bitrate. However, perceptual visual quality is hard to quantify in practice [1]. Moreover, advanced video encoders that use rate-distortion optimization (RDO) techniques to perform mode decisions [2], [3] require a distortion metric that can easily be computed. The sum-of-square error (SSE) and sum-of-absolute distance (SAD) are, therefore, usually used as the distortion metric since they are convenient and well-understood quantities.

However, it is well known that both SSE and SAD are not good measures of perceptual quality [1]. Several models of human visual perception have been proposed in the past and applied to the problem of measuring perceptual visual quality. Unfortunately, the state of understanding of the human visual system is still limited. Furthermore, these models attempt to model highly nonlinear perceptual systems with restricted and simplified psychovisual models. The complexities involved in computing metrics based on these models may also make them unsuitable for use in practical video encoders.

Recently, simpler perceptual visual quality metrics with good correlation to human perception have been proposed, including structural similarity index (SSIM) [4], visual information fidelity [5], visual signal-to-noise ratio [6], natural image contour evaluation [7], local feature-based visual security [8], and local edge gradients [9]. These developments have given hope that we could make progress toward video encoders that directly optimize for visual quality [1].

Manuscript received April 5, 2012; revised August 9, 2012 and October 3, 2012; accepted October 16, 2012. Date of publication January 16, 2013; date of current version June 27, 2013. This paper was recommended by Associate Editor R. Rinaldo.

The authors are with the Signal Processing Department, Institute for Infocomm Research, Singapore (e-mail: chyeo@i2r.a-star.edu.sg; hltan@i2r.a-star.edu.sg; yhtan@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2240918

A. Problem Statement

For ease of introducing the problem that we wish to address, we consider the RDO problem at a block level. In RDO, the goal is to minimize a distortion metric subject to a rate constraint [2], [3], that is

$$\min_{\Phi} D(\Phi) \text{ s.t. } R(\Phi) \leq R_c$$

where Φ is the set of decisions made for the block, $D(\Phi)$ and $R(\Phi)$ are the distortion and rate, respectively, achieved by using Φ, and R_c is the rate constraint. The distortion metric used is typically the SSE or SAD.

This constrained optimization problem is typically solved by using the Lagrange multipliers method and minimizing the Lagrangian [2], [3], that is

$$\min_{\Phi} J(\Phi;\lambda) = D(\Phi) + \lambda R(\Phi)$$

where $J(\Phi;\lambda)$ is the Lagrangian and λ is the Lagrange multiplier. In principle, this can be solved by finding for each λ, the Φ that results in the minimum Lagrangian, and then by choosing the solution that results in the minimum distortion while still meeting the rate constraint, which typically happens at $R(\Phi) = R_c$. In practice, λ is typically fixed either through a quantization parameter (QP) relationship [2] or through rate control.

In this paper, we wish to solve the same RDO problem but using a perceptually motivated distortion metric such as SSIM. This would enable video encoders to explicitly optimize for a perceptual visual quality instead of through a proxy such as SSE or SAD.

B. Contributions

Our contributions are the following. First, we introduce a novel SSIM-based distortion metric that is derived from SSIM using a high resolution quantization assumption. This proposed metric, which we term dSSIM, is a function of mean-square-error (MSE) between the source and reconstructed signals and the source signal variance. Therefore, there is no need to compute the reconstructed signal variance or the cross-covariance between the source and reconstructed signals when using dSSIM. Second, we show that by using dSSIM in the

RDO framework described above in Section I-A, the optimization problem can be reformulated as a standard SSE-based RDO problem with a scaled Lagrange multiplier, in which the scaling is based on the local variance of the source region. We also derive and describe how this scaling should be done. Third, we derive a simple method for explicitly computing an appropriate global Lagrange multiplier to use, based on the default QP-Lagrange multiplier used in the H.264/AVC reference software. In particular, our proposed method does not require any preprocessing or multipass encoding. Finally, we have performed extensive simulations using a H.264/AVC-based implementation, and our experimental results show that our proposed method can effectively improve compression performance in terms of SSIM over the H.264/AVC reference software with the results that are comparable to other state-of-the-art SSIM-based RDO methods [10], [11]. We have also conducted subjective viewing tests to validate our approach.

An earlier version of this paper has been presented in [12]; compared to our earlier work, we have included an empirical validation of the relationship that is derived between SSIM and MSE, performed extensive experiments including the case of not testing a range of QPs when performing mode decision for each macroblock (MB), and conducted a subjective evaluation of the proposed method. We have also included a comprehensive review of the related work that either optimizes SSIM within the RDO decision framework or performs a perceptually motivated adaptive Lagrange multiplier selection.

The remainder of this paper is organized as follows. Section II discusses related work in applying SSIM or other perceptual metrics within the RDO framework, as well as other perceptually motivated encoding methods such as adaptive quantization. Then, we derive the relationship between SSIM and MSE and the proposed dSSIM distortion metric in Section III. Section IV describes how to use dSSIM within the RDO framework and how to compute an appropriate global Lagrange multiplier to use. Experimental setup and results are presented in Section V before concluding in Section VI.

II. Related Work

In this section, we review two main classes of related work. The first are those that attempt to optimize SSIM directly during the encoding process. The second are those that use adaptive quantization and/or adaptive Lagrange multiplier selection to take into account perceptual characteristics during the encoding process.

A. SSIM-Based Optimization

A number of previous works have incorporated optimizing SSIM into the encoder. These typically involve either directly maximizing SSIM or minimizing a distortion metric based on SSIM, i.e., (1-SSIM). Wang *et al.* [13] proposed an iterative bit allocation process for a bitplane-based image coder such as set partitioning in hierarchical trees, with the aim of maximizing the minimum SSIM over the reconstructed image.

Mai *et al.* [14] first proposed replacing SSE with (1-SSIM) as the distortion function used in RDO for Intra coding for H.264/AVC. They also proposed a Lagrange multiplier to be used, which can be computed as a function of QP, but it was not explained clearly how this relationship was derived. Their experiments on all-Intra coding showed that 2.2%–6.5% rate savings is possible while maintaining the same average SSIM score. Mai *et al.* [15], [16] later extended the idea of using (1-SSIM) as the distortion metric to fast Intra mode decision and block matching motion estimation.

Yang *et al.* [17] also used (1-SSIM) as the distortion metric in the Inter mode decision process, as well as in the block matching motion estimation process for H.264/AVC. From their experimental results for P-frames coding, an average of 15% rate savings is possible over a range of QP, but there is an average increase in encoding runtime of about 140%. It was also reported that the Lagrange multipliers used are scaled by a factor determined experimentally for each QP [17]. In a later work, Yang *et al.* [18] proposed using SAD or sum of absolute transform distance as the metric within block matching motion estimation as is the usual case, but using (1-SSIM) as the distortion metric within Inter mode decision. Furthermore, an appropriate Lagrange multiplier is computed using the same implicit rate-QP model used in the H.264/AVC reference software, JM, and an SSIM-QP model constructed experimentally from a training sequence.

Huang *et al.* [19], [10] also proposed using (1-SSIM) as the distortion metric within the RDO framework for H.264/AVC. Their main contribution is a method for computing an appropriate Lagrange multiplier, using an assumption that the tangent at a particular operating point on the rate-SSIM curve for SSIM-based RDO would have the same gradient as the tangent at the closest operating point on the rate-SSIM curve for SSE-based RDO. In their scheme, a key frame is encoded using SSE-based RDO at two QPs to obtain 2 (rate, SSIM) operating points. A power function rate-SSIM model is then fitted on these two points to find the model parameters. Subsequent frames will then use this rate-SSIM model, and use as the Lagrange multiplier the slope of the tangent at the point that is closest to the (rate, SSIM) point of the previous coded frame. Key frames are selected to be either the first frame or frames with scene changes. Experimental results show about 11.5% average rate savings with a 22.5% average increase in encoding time [19].[1] A similar approach was used earlier for Intra coding [20].

Wang *et al.* [21] proposed a SSIM-QP model and a model for rate as a function of residual coefficient statistics for use within a SSIM-based RDO framework for H.264/AVC, which also uses (1-SSIM) as the distortion metric. Both models are derived in the transform domain assuming a Laplace distribution for the transformed residuals, with the model parameters being estimated using statistics from three previous coded frames. The two proposed models are then used to compute the Lagrange multiplier. The same approach was later repeated, but using a reduced-reference SSIM-QP model in the transform domain and a simplified rate-QP model [11], [22]; again, the model parameters are estimated from previously encoded frames. An MB level Lagrange multiplier adjustment based on motion and local contrast weighting was

[1]No similar encoding runtime comparison numbers were provided in [10].

also proposed in a later work [11]. Experimental results show an average rate reduction of about 13% with a 6% encoding runtime increase [11].

Our work differs from the above methods in several ways. First, instead of using (1-SSIM) as the distortion metric, we use an approximation of 1/SSIM, as will be described later. Second, we do not require computing SSIM explicitly during the RDO process that eliminates some computation overhead. Third, no model parameters need to be estimated, so our proposed method requires neither special initialization nor multipass encoding. Finally, because of our formulation of the SSIM-based RDO as a local scaling of the Lagrange multiplier, the proposed method can be used in a consistent fashion throughout the entire encoding pipeline whenever rate-distortion tradeoffs need to be made, including mode decision, motion estimation, and rate-distortion optimized quantization.

B. Perceptual-Based Adaptive Quantization and Lagrange Multiplier Selection

An early adaptive quantization method was described in the rate control and quantization control chapter of MPEG-2 Test Model 5 under Step 3, where the quantization step size is scaled according to the spatial activity in the MB relative to its average over the previous coded frame [23]. In this method, the spatial activity is a function of the variance of each subblock within the MB. Tang [24] proposed a motion attention model, a visual sensitivity model, and a visual masking model for use during video encoding; the models are then used to select a QP for each MB.

Tsai *et al.* [25] proposed a visual attention model based on frame differences, and this model is used to scale the Lagrange multiplier used for each MB. Pan *et al.* [26] proposed adapting the frame-level Lagrange multiplier for scalable video coding by adding a spatial gradient dependent adjustment and a motion vector dependent adjustment. Furthermore, the Lagrange multiplier used for each MB is scaled based on a local gradient coherence term and output from a skin pixel detector. A similar approach was also proposed by Sun *et al.* [27], where the Lagrange multiplier for each MB is constructed to be a weighted sum of a local gradient coherence based term and a motion attention model based term. A later work then incorporated a position model [28]. Chen *et al.* [29] proposed the use of a foveated just-noticeable-distortion (JND) model that is incorporated into a video encoder to adjust QP and Lagrange multiplier for each MB based on their foveated JND weighting.

While our proposed method involves locally adapting the Lagrange multiplier and QP, the adaptation is driven by SSIM-based optimization, instead of simply relying on perceptually motivated heuristics.

III. SSIM Approximation Using MSE

The SSIM between two image regions is defined as [4]

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

where x and y are the two image regions to be compared, μ_x and μ_y are the means of x and y, respectively, σ_x^2 and σ_y^2 are the variances of x and y, respectively, σ_{xy} is the cross-covariance between x and y, and c_1 and c_2 are two constants used for numerical stability. From [4], the default values to be used are $c_1 = (\kappa_1 L)^2$ and $c_2 = (\kappa_2 L)^2$, where $\kappa_1 = 0.01$ and $\kappa_2 = 0.03$. L is the peak value of the image, and is, in general, given by $L = 2^{\text{bitdepth}} - 1$. Here, we would denote the original image region by x and the reconstructed image region by y.

We use the following additive distortion model for y:

$$y = x + e$$

where e is the reconstruction error due to lossy quantization. We model e as a random variable with variance σ_e^2. Assuming high-resolution quantization and smooth residual probability densities, we make the approximation that e is uncorrelated with x and has zero mean (i.e., $\mu_e = 0$) [30]. Note that MSE can be computed as

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

where N is the number of pixels in the region, and the index i denotes the ith pixel 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 under the high-resolution quantization approximation

$$\mu_y \approx \mu_x \tag{3}$$

$$\sigma_y^2 \approx \sigma_x^2 + \sigma_e^2 \tag{4}$$

$$\sigma_{xy} \approx \sigma_x^2. \tag{5}$$

By substituting (3), (4), and (5) into (1), we can obtain an approximation of SSIM as

$$\begin{aligned}\text{SSIM} &\approx \frac{2\sigma_x^2 + c_2}{2\sigma_x^2 + \sigma_e^2 + c_2} \\ &\approx \frac{2\sigma_x^2 + c_2}{2\sigma_x^2 + \text{MSE} + c_2}\end{aligned} \tag{6}$$

where the last line follows from (2) for reasonably large values of N. Since all the quantities in (6) are positive, under these assumptions, $0 < \text{SSIM} \leq 1$, and we can define a distortion metric based on SSIM as follows:

$$\begin{aligned}\text{dSSIM} &= \frac{1}{\text{SSIM}} \\ &\approx 1 + \frac{\text{MSE}}{2\sigma_x^2 + c_2}.\end{aligned} \tag{7}$$

Note that with this approximation, $\text{dSSIM} \geq 1$.

(7) 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; in other words, the more textured a region, the higher the tolerable MSE. Therefore, for the same perception of visual degradation, the MSE can be higher in a textured region, compared to a smooth region.

We note also that previous works have sought the relationship between MSE or PSNR and SSIM. For example, Horè and Ziou [31] have presented a relationship between PSNR

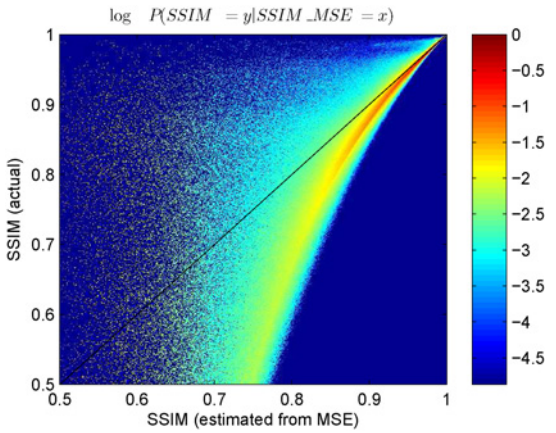


Fig. 1. Conditional probability density of SSIM, (1), given the approximation computed from MSE, (6). Note that for ease of viewing, the probability densities are plotted on a log scale, with the accompanying colormap showing the magnitudes.

and SSIM. Their corresponding relationship between MSE and SSIM depends on σ_{xy} instead of σ_x^2, i.e., $\frac{1}{\text{SSIM}} = 1 + \frac{\text{MSE}}{2\sigma_{xy}+c_2}$. In contrast, (7) is more convenient to use within an RDO framework since the source variance σ_x^2 only has to be computed once per block, whereas σ_{xy} needs to be computed for the reconstruction of a block for each possible mode decision.

A. Validation

We have carried out experiments to validate the relationship between SSIM and MSE, as derived in (7). Eleven video sequences of 720p, CIF, and QCIF resolutions were encoded using H.264/AVC at QPs ranging from 12 to 42. Ten frames spaced 1 s apart from each other were selected; for each 8×8 block of pixels, we compute SSIM using both the original definition in (1) and the approximation in (6).

Fig. 1 shows the conditional probability density of SSIM (1), given the approximation computed from MSE (6). This plot suggests that the approximation is good at high values of SSIM (above 0.95), and degrades at lower values of SSIM. To give further insight into how the approximation behaves at different QP settings, we also show the plot of the correlation coefficient between the two quantities over various QPs in Fig. 2(a), and the root mean-square-error (RMSE) between them over various QPs in Fig. 2(b). As might be expected, the approximation is good at low QPs, and degrades at high QPs.

IV. Proposed Method for SSIM-Based RDO

A. Basic Concept

In a block-based encoder, the RDO decision for each block can be performed by minimizing distortion subject to a rate constraint using the Lagrange multiplier method [2], [3]. When SSE is the distortion metric, this is done by optimizing

$$J_{\text{SSE}} = \text{SSE} + \lambda_{\text{SSE}} R = N \cdot \text{MSE} + \lambda_{\text{SSE}} R$$

for an appropriately chosen Lagrange multiplier λ_{SSE}. Note the use of the SSE subscript when SSE is used as the distortion metric.

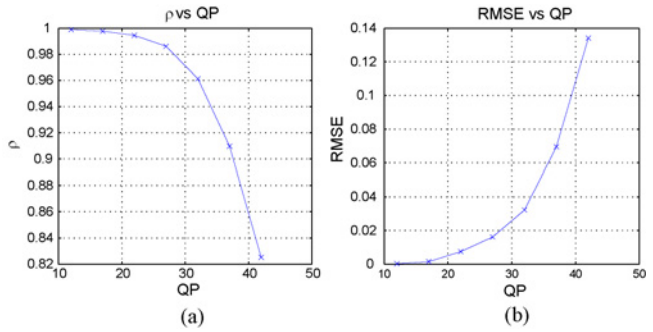


Fig. 2. Performance of SSIM approximation computed from MSE, (6), at various QPs when test videos are encoded by H.264/AVC. (a) Correlation coefficient versus QP. (b) RMSE versus QP.

To incorporate SSIM into RDO, we use dSSIM as defined in (7) as the distortion metric and optimize the following cost function for a block with N pixels:

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

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

$$J = \text{SSE} + \left(2\sigma_x^2 + c_2\right)\lambda R \tag{8}$$

again for some appropriately chosen Lagrange multiplier λ. Note that to keep the notation concise, we omit subscripts for J and λ when dSSIM is used as the distortion metric.

(8) offers us a convenient way to incorporate SSIM into the RDO decision process by using a local scaling of λ that depends on the local source variance. This means that the entire RDO machinery can be retained with just a minor modification of the Lagrange multiplier. There is again an intuitive explanation for this procedure. Essentially, in a highly textured region, an additional rate would be penalized more than in a smooth region, which implies that a larger SSE can be tolerated.

B. Computing the Lagrange Multiplier

While we have shown how to optimize SSIM within an RDO decision framework, there remains the issue of choosing 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)[2] statistics is the same whether SSE or SSIM is to be optimized.

Recall that when SSE is used, the optimization problem is to minimize total distortion subject to a constraint on the total

[2]Here, the DFD refers to the residual image that remains after performing motion compensation or some other type of prediction.

rate, i.e., [2], [3]

$$\begin{aligned} \min_{\Phi} \quad & \text{SSE} = \sum_i d_i \\ \text{s.t.} \quad & R = \sum_i r_i \le R_c \end{aligned}$$

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:

$$\min_{\{\phi_i\}_{i=1}^{M}} J_{\text{SSE}} = \sum_i d_i + \lambda_{\text{SSE}} \sum_i r_i = \sum_i (d_i + \lambda_{\text{SSE}} r_i) \quad (9)$$

where M is the number of MBs within the video frame, and ϕ_i is the set of encoder decisions for the ith MB. Typically, this optimization is performed by ignoring dependences between MBs and solving for each MB the following unconstrained problem:

$$\min_{\phi_i} d_i + \lambda_{\text{SSE}} r_i.$$

In the H.264/AVC [32] reference software (JM), the Lagrange multiplier is computed as $\lambda_{\text{SSE}} = \beta \cdot 2^{(\text{QP}-12)/3}$ [33], with β being a scale constant. This is justified by assuming the following rate-distortion model for each MB [2]:

$$r(d) = N\alpha \log\left(\frac{\sigma^2}{d/N}\right) \quad (10)$$

where $r(d)$ is the rate used to code the MB, σ^2 is the variance of the DFD in the MB, d is the SSE distortion of the reconstructed MB, and α is a scale constant.

To solve (9), we set for each i

$$\frac{\partial J_{\text{SSE}}}{\partial d_i} = 1 + \lambda_{\text{SSE}} \frac{\partial r_i}{\partial d_i} = 0. \quad (11)$$

By using (10) in (11), we can solve for

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

where d_i^* and r_i^* are the optimal SSE and the 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

$$R_{\text{SSE}} = N\alpha \sum_i \log\left(\frac{\sigma_i^2}{\alpha\lambda_{\text{SSE}}}\right).$$

We can repeat the same exercise when dSSIM is used as the distortion metric instead. Using (7), we would optimize

$$\begin{aligned} \min_{\{\phi_i\}_{i=1}^{M}} J &= \sum_i \frac{d_i}{2\sigma_{x_i}^2 + c_2} + \lambda \sum_i r_i \\ &= \sum_i \left(\frac{d_i}{2\sigma_{x_i}^2 + c_2} + \lambda r_i\right) \end{aligned} \quad (12)$$

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

To solve (12), we again set for each i

$$\frac{\partial J}{\partial d_i} = \frac{1}{2\sigma_{x_i}^2 + c_2} + \lambda \frac{\partial r_i}{\partial d_i} = 0. \quad (13)$$

Using (10) in (13), we solve for

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

The total rate used is

$$R_{\text{SSIM}} = N\alpha \sum_i \log\left(\frac{\sigma_i^2}{\alpha\left(2\sigma_{x_i}^2 + c_2\right)\lambda}\right).$$

As mentioned, we will pick λ such that the rate would be the same regardless of whether MSE or dSSIM is used as the distortion metric, and this will be computed as a function of λ_{SSE}, the Lagrange multiplier that is used in JM. By setting $R_{\text{SSIM}} = R_{\text{SSE}}$, we obtain

$$\lambda = \lambda_{\text{SSE}} \cdot \exp\left(-\frac{1}{M}\sum_{i=1}^{M} \log\left(2\sigma_{x_i}^2 + c_2\right)\right). \quad (14)$$

This means that in using (8) to perform RDO decision, we will use for the ith MB a Lagrange multiplier of

$$\lambda_i = \frac{2\sigma_{x_i}^2 + c_2}{\exp\left(\frac{1}{M}\sum_{j=1}^{M} \log\left(2\sigma_{x_j}^2 + c_2\right)\right)} \lambda_{\text{SSE}}. \quad (15)$$

This is simply applying a scaling, which depends on the local source variance statistic normalized by the geometric mean of the same statistic computed over the entire frame, to the original Lagrange multiplier used in JM. This gives a concrete way of applying a small modification to the RDO process to maximize SSIM over the entire frame.

C. A QP Change Perspective

It is also useful to interpret the local scaling in the Lagrange multiplier as corresponding to a change in a QP for each MB. Let $\bar{\text{QP}}$ be the original (or master) QP, and QP_i be the implicit QP used in SSIM-RDO for the ith MB. Then, using the relation, $\lambda_{\text{SSE}} = \beta \cdot 2^{(\text{QP}-12)/3}$ [33], and $\lambda = \gamma_i \lambda_{\text{SSE}}$, where γ_i is the scaling applied to the Lagrange multiplier in JM for the ith MB as in (15), we find that

$$\Delta\text{QP}_i = \text{QP}_i - \bar{\text{QP}} = 3\log_2 \gamma_i. \quad (16)$$

Encoders such as JM can test a range of QPs when encoding each MB to find the QP that gives the best RD cost. However, since this increases encoder complexity, only a limited range of QPs would be tested if such testing of multiple QPs is done. In practice, we have found that it helps apply the above limit on the scaling of the Lagrange multiplier to correspond to the range of QP testing that is performed by the encoder. Specifically, if we wish to enforce $\Delta_{\min} \le \Delta\text{QP}_i \le \Delta_{\max}$, we can use (16) to find the corresponding limits on γ_i as $2^{\Delta_{\min}/3} \le \gamma_i \le 2^{\Delta_{\max}/3}$.

Alternately, we can use (16) to directly determine the ΔQP to be applied for each MB, that is

$$\Delta\text{QP}_i = 3\left(s_i - \frac{1}{M}\sum_{j-1}^{M} s_j\right) \quad (17)$$

function FRAME-SSIMRDO(Input frame, λ_{SSE})
▷ Pre-compute parameters
  for $i \leftarrow 1$ **to** M **do**
    $\sigma_{x_i}^2 \leftarrow$ variance of ith MB
  end for
  $D \leftarrow \exp\left(\frac{1}{M}\sum_{i=1}^{M}\log\left(2\sigma_{x_i}^2 + c_2\right)\right)$
▷ Perform RDO on frame
  for $i \leftarrow 1$ **to** M **do**
    $\lambda_i \leftarrow \frac{2\sigma_{x_i}^2 + c_2}{D}\lambda_{\text{SSE}}$
    $\phi_i \leftarrow$ MB_RDO(i,λ_i)
  end for
  return $\{\phi_i\}_{i=1}^{M}$
end function

Fig. 3. Pseudocode of the proposed method. MB_RDO(i,λ_i) refers to performing the conventional RDO mode decision on the ith MB using a Lagrange multiplier of λ_i.

where

$$s_i = \log_2\left(2\sigma_{x_i}^2 + c_2\right).$$

D. Summary of Algorithm and Complexity Analysis

Fig. 3 presents a pseudocode for the proposed method when encoding a frame. It clearly shows that the overheads necessary when compared to conventional RDO, e.g., [33], are: 1) computing the variance of each MB; 2) computing a normalization constant, D; and 3) computing the Lagrange multiplier to be used for each MB. In particular, 1) and 3) are only performed once per MB, while 2) is only performed once per frame. Furthermore, these are operations with relatively low complexity compared to the rest of the encoding process, e.g., motion estimation and mode decision.

Suppose that an encoder needs to pick one out of K different possible ways to encode an MB. In our proposed method, within the MB RDO process, the SSE is computed K times, once for each possible reconstruction; but the reconstruction variance and cross-covariance need not be computed.

In contrast, let us consider previous SSIM-based RDO methods, e.g., [10], [11], and look at the computations that are necessary within RDO. If (1-SSIM) is used as the distortion metric for RDO, as in [10] and [11], then for each MB, the source variance needs to be computed once, while the reconstruction variance and cross-covariance need to be computed K times—once for each possible reconstruction. Note that the computation of the source and reconstructed means are necessary for the computation of the various variance terms, and that depending on the implementation, it may also be necessary to compute the SSE K times. Furthermore, additional steps are needed to compute an appropriate Lagrange multiplier either by using multiple encoding passes, e.g., [10], or by model parameters fitting, e.g., [11].

V. EXPERIMENTAL RESULTS

To test the proposed method, we have implemented it in the H.264/AVC reference software, JM 17.2.[3] 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 (15); this involves computing $\sigma_{x_i}^2$, the variance of source pixels within each MB. Then, λ_i, the Lagrange multiplier to be used for encoding the ith MB, is scaled as in (15) subject to the constraints discussed in Section IV-C. λ_i is then used in all RDO processes such as mode-decision, RD-optimized quantization, and motion estimation. We used the same MB size as in H.264/AVC, i.e., $N = 16 \times 16 = 256$. No other changes to the encoder software were made.

TABLE I
DESCRIPTION OF TEST VIDEO SEQUENCES USED

Name	Resolution	Frames
mobcal_720p	1280 × 720	500
shields_720p	1280 × 720	500
stockholm_720p	1280 × 720	600
parkjoy_720p	1280 × 720	500
silent_cif	352 × 288	300
flower_cif	352 × 288	375
bus_cif	352 × 288	150
foreman_cif	352 × 288	300
salesman_qcif	176 × 144	449
carphone_qcif	176 × 144	382
container_qcif	176 × 144	300

TABLE II
ENCODER SETTING FOR EXPERIMENTS

Parameter	**Setting**
ProfileIDC	100 (High)
SearchRange	64
NumerReferenceFrames	5
BiPredMERefinements	3
SymbolMode	1 (CABAC)
WeightedPrediction	0 (Off)
RDPictureDecision	0 (Off)
RDOptimization	1 (High complexity mode)
UseRDOQuant	1 (On)
SearchMode	3 (EPZS)

As shown in Table I, we applied the proposed approach to the encoding of 4 720p (1280 × 720 pixels) test sequences, 4 CIF (352 × 288 pixels) test sequences, and 3 QCIF (176 × 144 pixels) test sequences.[4] We used three different configurations that target different applications: all-Intra for use in high-quality cinema scenarios, random access for use in storage applications, and low delay for use in video conferencing applications. In the all-Intra configuration, all the frames are encoded as Intra pictures without any temporal prediction. In the random access configuration, we use an eight-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 other notable encoder settings are shown in Table II. The encoding was carried out over a range of QPs: {20, 25, 30, 35}.

We will consider both PSNR and SSIM as quality metrics. The PSNR for a video is computed as the average PSNR across frames. Similarly, the SSIM for a video is computed as the average SSIM across frames. To help us interpret coding performances, we will show BD-rate [34] figures that

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

[4]The test sequences can be obtained from http://media.xiph.org/video/derf/.

compute the average rate difference between two rate-quality curves, with piecewise cubic polynomials being used to fit each rate-quality curve for computing the integral over the range of qualities covered by the curves. The baseline is 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 or 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 to understand the complexity of the proposed approach.

A. JM Experiments With QP Sweep

In this subsection, we present experimental results when QP sweep is enabled. In other words, during encoding, for each MB, the encoder will try a range of QP values and choose the one that gives the best RD cost. While this will lead to better compression results, it also results in an increase in encoding runtime. For these experiments, we sweep across five QP values, i.e., $\{\bar{\text{QP}}-2, \bar{\text{QP}}-1, \bar{\text{QP}}, \bar{\text{QP}}+1, \bar{\text{QP}}+2\}$, using the encoder option `RDOQ_QP_Num = 5`. The constraints on scaling of the Lagrange multiplier are computed based on the above range of QP values considered during RDO. The baseline JM17.2 also uses the same QP sweep.

Table III shows the simulation results for the all-Intra, random access, and low-delay encoding configurations. The key observation is that for the same SSIM, the proposed approach can give significant coding gains ranging from 4% to 19%. This means that to get the same perceptual quality, our method can use up to 19% less rate compared to the baseline. The average rate reductions for all-Intra, random access, and low-delay encoding configurations are 7.7%, 9.1%, and 13.5%, respectively. On the other hand, for the same PSNR, the proposed approach suffers some coding loss of up to 13%, with an average loss of 2.1%, 3.2%, and 2.9% 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. In other words, our method does not introduce any significant complexity increase into the encoding process. 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 a significant increase in encoding time.

Fig. 4 shows the SSIM-rate plots for the *parkjoy_720p* and *silent_cif* sequences. We see that for different GOP structures and content, there is an improvement in SSIM across the tested range of rates.

B. JM Experiments With Fixed QP

Performing a QP sweep for each MB during encoding may not be feasible in practice, especially when the encoder is operating under tight latency and power constraints. An alternative is to use (17) to determine the appropriate ΔQP to use in addition to using the scaled Lagrange multiplier. Thus, we repeat the above experiment, but using the fixed QP determined by (17), so only 1 QP value will be tested during the encoding of each MB, using the encoder option `RDOQ_QP_Num = 1`. The baseline JM17.2 also uses a fixed QP value corresponding to the user-specified master QP.

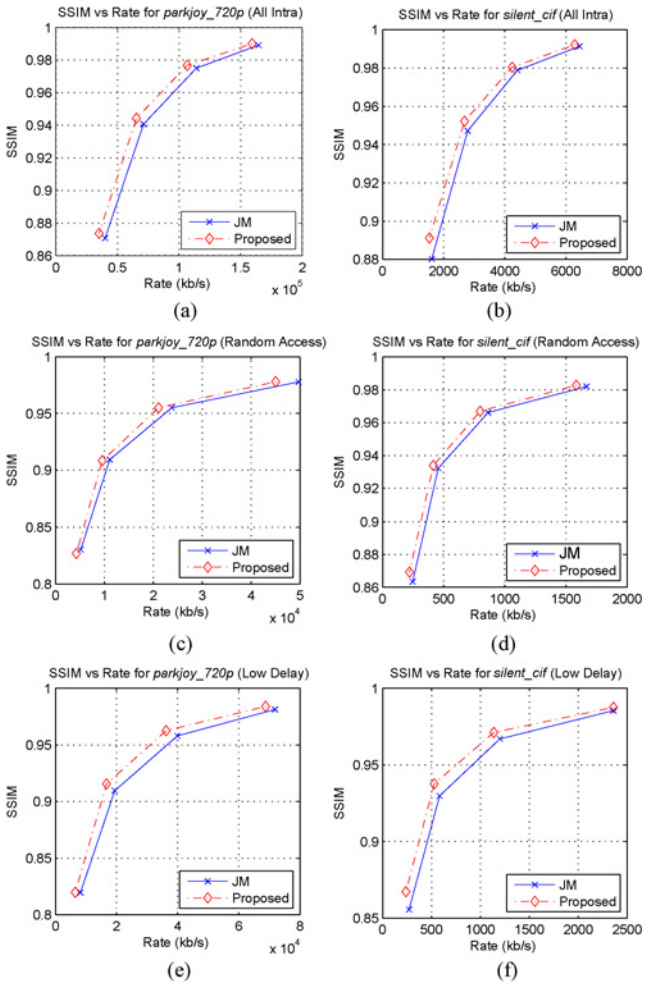


Fig. 4. SSIM versus rate plots comparing JM 17.2 with the proposed method, when using QP sweep at the encoder for different GOP structures. (a) *Parkjoy* (all intra). (b) *Silent* (all intra). (c) *Parkjoy* (random access). (d) *Silent* (random access). (e) *Parkjoy* (low delay). (f) *Silent* (low delay).

Table IV shows the results for the all-Intra, random access, and low-delay encoding configurations. As before, for the same SSIM, the proposed approach can give significant coding gains ranging from 3% to 21%. The average rate reductions for all-Intra, random access, and low-delay encoding configurations are 8.1%, 9.8%, and 14.0%, respectively. On the other hand, for the same PSNR, the proposed approach suffers some coding loss of up to 20%, with an average loss of 3.9%, 6.8%, and 6.0% for all-Intra, random access, and low-delay encoding configurations, respectively. Similarly, the encoding time of the proposed approach does not show any significant adverse impact, and is about the same as the baseline.

Fig. 5 shows the SSIM-rate plots for the *parkjoy_720p* and *silent_cif* sequences. Again, we see that for different GOP

TABLE III
BD-RATE (%) RESULTS FOR JM EXPERIMENTS WITH QP SWEEP (RDOQ_QP_Num = 5)

Sequence	All Intra			Random Access			Low Delay		
	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time
mobcal_720p	−7.6%	2.2%	101%	−9.6%	5.5%	102%	−15.4%	2.4%	100%
shields_720p	−9.8%	2.1%	101%	−12.6%	7.3%	102%	−19.3%	5.5%	101%
stockholm_720p	−9.0%	2.7%	101%	−9.5%	7.7%	101%	−17.2%	13.2%	101%
parkjoy_720p	−10.3%	1.3%	101%	−12.4%	0.7%	100%	−17.6%	−2.2%	99%
silent_cif	−5.7%	3.2%	101%	−4.9%	2.8%	100%	−4.1%	6.3%	101%
flower_cif	−7.2%	0.5%	100%	−10.1%	0.8%	100%	−15.6%	−0.6%	100%
bus_cif	−9.5%	2.3%	101%	−12.0%	1.4%	100%	−17.7%	−1.9%	100%
foreman_cif	−6.8%	2.8%	101%	−6.1%	2.4%	100%	−9.0%	2.3%	100%
salesman_qcif	−5.4%	2.6%	101%	−8.0%	−0.2%	101%	−14.0%	−3.6%	100%
carphone_qcif	−3.8%	2.1%	99%	−5.1%	3.0%	100%	−4.6%	5.2%	100%
container_qcif	−9.9%	1.7%	100%	−9.3%	4.0%	100%	−14.5%	5.6%	101%
Average	−7.7%	2.1%	101%	−9.1%	3.2%	101%	−13.5%	2.9%	100%

TABLE IV
BD-RATE (%) RESULTS FOR JM EXPERIMENTS WITH FIXED QP (RDOQ_QP_Num = 1)

Sequence	All Intra			Random Access			Low Delay		
	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time	BD-Rate (SSIM)	BD-Rate (PSNR)	Encoding Time
mobcal_720p	−7.6%	4.3%	101%	−10.2%	10.3%	102%	−14.9%	8.2%	101%
shields_720p	−9.9%	4.1%	101%	−12.4%	14.3%	101%	−18.8%	13.2%	101%
stockholm_720p	−9.3%	4.9%	100%	−7.4%	17.7%	100%	−15.5%	20.1%	101%
parkjoy_720p	−12.0%	2.2%	101%	−16.5%	0.8%	100%	−19.7%	−1.7%	100%
silent_cif	−5.2%	5.7%	100%	−4.6%	6.9%	101%	−3.5%	8.5%	101%
flower_cif	−8.7%	0.8%	98%	−12.7%	1.3%	101%	−17.3%	0.8%	99%
bus_cif	−10.3%	3.6%	100%	−15.2%	1.8%	101%	−20.6%	−1.1%	101%
foreman_cif	−6.6%	5.7%	99%	−5.7%	5.8%	101%	−9.0%	5.8%	101%
salesman_qcif	−5.0%	4.5%	100%	−9.8%	1.1%	100%	−15.9%	−4.1%	100%
carphone_qcif	−3.6%	4.1%	99%	−3.5%	6.8%	100%	−2.6%	8.1%	100%
container_qcif	−10.4%	3.4%	102%	−10.1%	7.9%	100%	−15.8%	7.8%	101%
Average	−8.1%	3.9%	100%	−9.8%	6.8%	101%	−14.0%	6.0%	100%

structures and content, there is an improvement in SSIM across a range of rates.

C. Comparison With Other Methods

To compare with other state-of-the-art algorithms that also aim to optimize SSIM during the RDO process, we ran our proposed method on the same settings as described by Wang *et al.* [11]. In particular, we tested using CAVLC entropy coding, five reference frames, high-complexity RDO mode decision, and RDOQ turned off for both IPP and IBP GOP structures. One hundred frames of each sequence are encoded, where the first frame is Intra coded and the remaining frames are Inter coded. We also compare two different sets of QPs, a low QP range of $\mathrm{QP}_{\mathrm{low}} = \{16, 20, 24, 28\}$ and a high QP range of $\mathrm{QP}_{\mathrm{high}} = \{24, 28, 32, 36\}$. The results are then compared with the methods proposed by Wang *et al.* [11] and Huang *et al.* [10] using the published results from [11].

Table V shows the results of this comparison. It can be observed that the performance of the proposed method has performance that is comparable to other state-of-the-art SSIM-based RDO procedures, but requires less drastic changes to the encoder RDO process. In addition, while our proposed method does not show any significant increase in encoding time, Wang *et al.* [11] reported an average overhead of about 6.5%, and Huang *et al.* [10] reported an average overhead of about 4.8%, excluding the need to perform an additional two encoding passes on certain frames.

D. Subjective Assessment

Table VI shows some visual quality results for the *bus_cif* sequence coded in the low-delay encoding configuration. The first column shows the original frame, while the second column shows the image difference between the decoded frame and the original frame after baseline encoding with JM. For comparison, two cropped regions are also shown at about 3× zoom. The third column shows the image difference of the decoded frame after encoding with JM at 23% less rate. The fourth column shows the image difference of the decoded frame after encoding with the proposed method that achieves the same SSIM as the second column but with a 23% less rate. Comparing the third and fourth columns, we see that while both use a 23% less rate compared to the second column, the visual quality in the fourth column is better, in particular, in the regions with trees where more detail is preserved; this is also suggested by the higher SSIM score. Finally, the last column shows the difference image of the decoded frame after encoding with the proposed method at the same rate as the second column. Again, as suggested by its higher SSIM score, the last column shows better visual quality than the second column.

Table VII shows similar visual quality results for the *flower_cif* sequence coded in the low delay encoding configuration, where the proposed method can achieve the same SSIM as JM but at a 27% less rate.

To provide some evidence of the perceptual improvements provided by our proposed method, we have also carried out

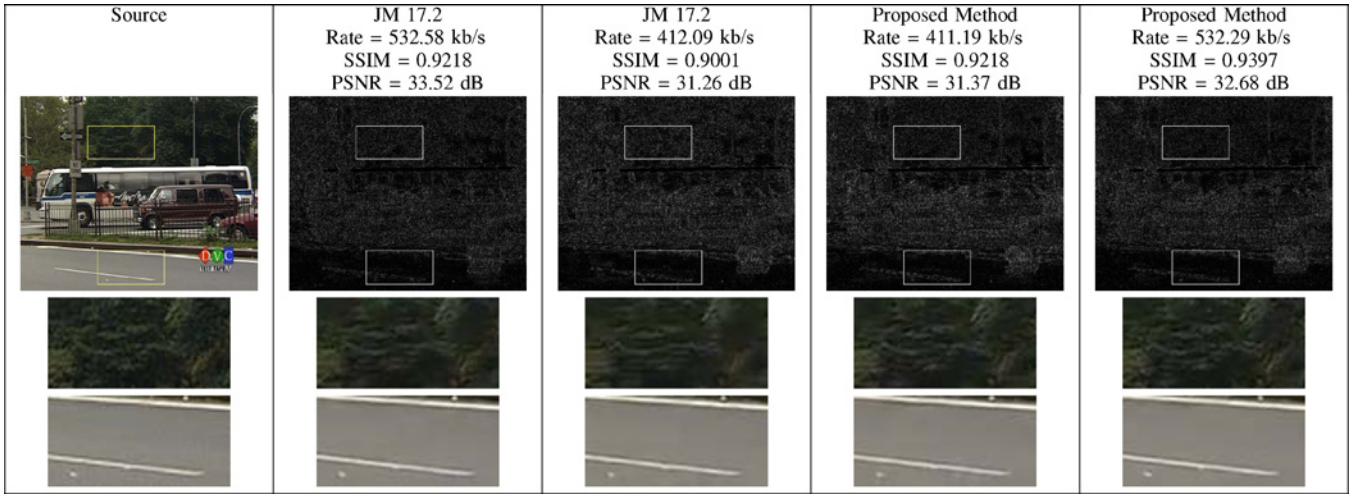
TABLE V

BD-RATE (%) RESULTS OF COMPARISONS WITH OTHER SSIM-BASED RDO METHOD

Sequence	**QP**	IPP			IBP		
		Proposed	Wang *et al.* [11]	Huang *et al.* [10]	Proposed	Wang *et al.* [11]	Huang *et al.* [10]
akiyo_cif	QP_{low}	−21.3%	−26.1%	−19.4%	−19.7%	−17.4%	−9.7%
	QP_{high}	−16.5%	−28.1%	−15.8%	−14.7%	−8.6%	−6.4%
bus_cif	QP_{low}	−14.9%	−7.8%	−6.0%	−15.5%	−2.0%	−4.0%
	QP_{high}	−20.2%	−14.9%	−13.3%	−19.1%	−7.6%	−5.3%
coastguard_cif	QP_{low}	−6.0%	−4.8%	−2.3%	−5.7%	−3.4%	−2.0%
	QP_{high}	−8.5%	−8.9%	−5.0%	−7.6%	−3.3%	−2.0%
silent_cif	QP_{low}	−6.4%	−9.6%	−5.3%	−6.8%	−4.6%	−4.3%
	QP_{high}	−5.3%	−12.4%	−6.8%	−5.0%	−6.8%	−4.6%
hall_cif	QP_{low}	−23.2%	−25.9%	−26.4%	−24.5%	−7.6%	−2.4%
	QP_{high}	−13.3%	−25.5%	−22.8%	−13.7%	−19.4%	−4.9%
MaD_cif	QP_{low}	−4.5%	−6.4%	−2.8%	−4.0%	−7.4%	−5.8%
	QP_{high}	−3.3%	−8.9%	−4.7%	−3.0%	−5.9%	−1.7%
spincal_720p	QP_{low}	−29.1%	−11.9%	−12.8%	−27.1%	−5.8%	−7.2%
	QP_{high}	−18.8%	−15.6%	−12.8%	−17.8%	−4.6%	−3.8%
night_720p	QP_{low}	−13.1%	−6.7%	−3.5%	−12.7%	−4.9%	−3.5%
	QP_{high}	−9.6%	−16.0%	−11.4%	−8.9%	−5.7%	−2.1%
Average	QP_{low}	−14.8%	−12.4%	−9.8%	−14.5%	−6.7%	−4.9%
	QP_{high}	−12.0%	−16.3%	−11.6%	−11.2%	−7.7%	−3.8%

TABLE VI

FRAME 149 OF SEQUENCE *Bus_Cif*, COMPRESSED USING THE LOW-DELAY ENCODING CONFIGURATION (BEST VIEWED IN COLOR)



subjective testing using a two-alternative forced choice (2AFC) test format, as was performed by Wang *et al.* [11]. In the 2AFC test format, within each basic test cell (BTC), the reference (source) video is first shown, followed sequentially by the reconstructed video encoded using a first method, and the reconstructed video encoded using a second method; the subject is then forced to pick the perceptually preferable method within 5 s. For this test, we selected six 720p videos, and encode each video using three different settings:

1) *SSIM-RDO*: using proposed SSIM-based RDO at rate R and SSIM S;
2) *JM-EQSSIM*: using JM-17.2 at rate $R+\Delta R$ and SSIM S;
3) *JM-EQRATE*: using JM-17.2 at rate R and SSIM $S-\Delta S$.

Table VIII shows the sequences and their rate and SSIM under each setting.[5] Two 2AFC tests were carried out: 1) comparing SSIM-RDO with JM-EQSSIM, and 2) comparing SSIM-RDO with JM-EQRATE.

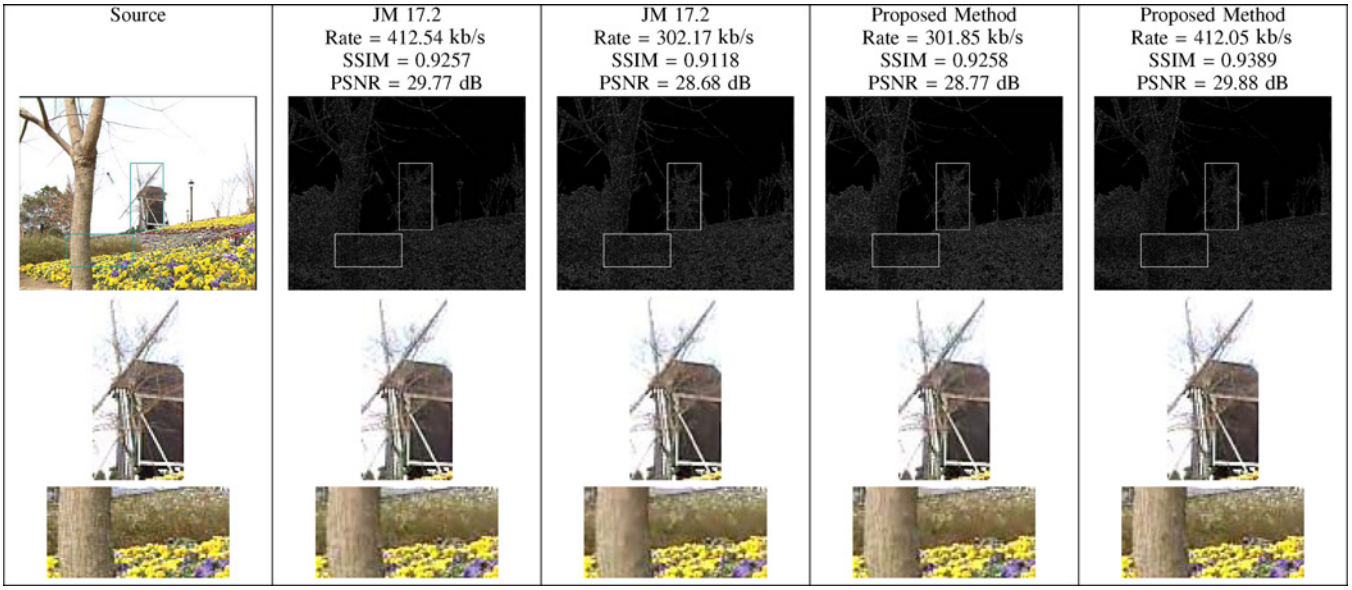
We performed the two comparison tests with 20 subjects, all of whom were not involved in performing this work. Two test sessions were carried out for each comparison test. In each test session, a training BTC is first shown using the *Johnny_720p* sequence, and the score of which is not used. Then, six BTCs are shown to cover the six test sequences, followed by another six BTCs. Each test session lasts about 8 min. Note that the order of appearance of the two methods was randomly generated for each test session. Therefore, in each comparison test, each subject was shown four BTCs for each of the six test sequences, and for each combination of comparison test and sequence, we obtained $20\times 4=80$ votes. Viewing was conducted in a room with minimal background illumination and a viewing distance of 60 cm was used.

Fig. 6 shows the results of the 2AFC test when comparing SSIM-RDO with JM-EQSSIM. We show the proportion of

[5]The bitstreams are available upon request.

TABLE VII

FRAME 219 OF SEQUENCE *Flower_Cif*, COMPRESSED USING THE LOW-DELAY ENCODING CONFIGURATION (BEST VIEWED IN COLOR)



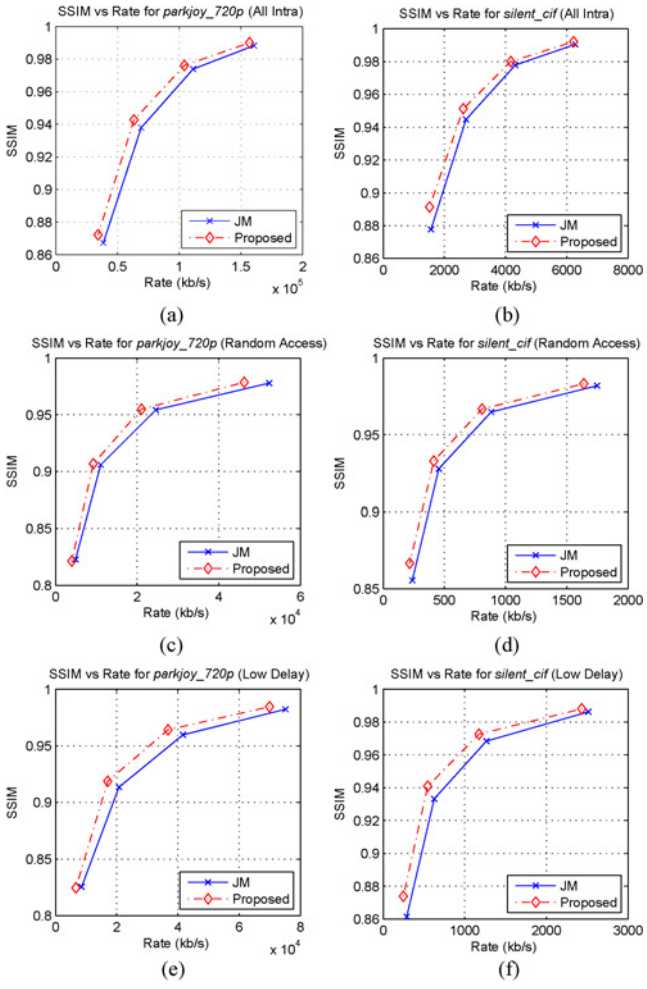


Fig. 5. SSIM versus rate plots comparing JM 17.2 with the proposed method, when using a fixed QP computed using (17) for different GOP structures. (a) *Parkjoy* (all intra). (b) *Silent* (all intra). (c) *Parkjoy* (random access). (d) *Silent* (random access). (e) *Parkjoy* (low delay). (f) *Silent* (low delay).

TABLE VIII

RATE AND SSIM FOR SEQUENCES USED IN SUBJECTIVE EVALUATION

Sequence	SSIM-RDO		JM-EQSSIM		JM-EQRATE	
	Rate (kb/s)	SSIM	Rate (kb/s)	SSIM	Rate (kb/s)	SSIM
mobcal_720p	1674	0.925	1974	0.925	1677	0.921
parkrun_720p	12 458	0.916	16 708	0.916	12 513	0.895
shields_720p	2003	0.904	2425	0.904	2009	0.898
parkjoy_720p	17 071	0.919	22 127	0.919	17 128	0.897
stockholm_720p	1534	0.884	1746	0.884	1533	0.879
Night_720p	2731	0.917	2890	0.917	2739	0.915

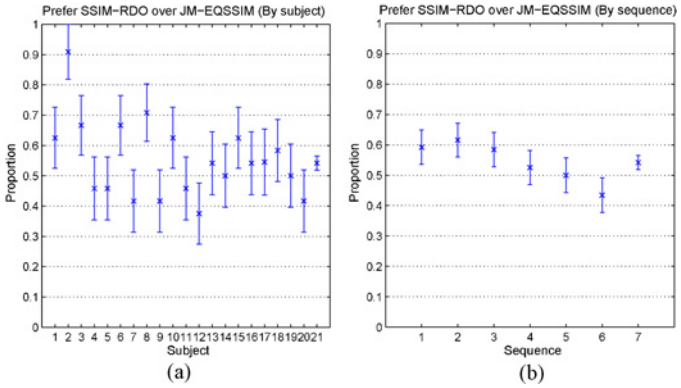


Fig. 6. Error bar plot of results from the 2AFC subjective test comparing SSIM-RDO with JM-EQSSIM, showing proportion that preferred SSIM-RDO. The error bars show ± one standard deviation. (a) Ordered by subject, where indices 1–20 correspond to the 20 subjects and index 21 corresponds to the overall average. (b) Ordered by sequence, where indices 1–6 correspond to the six sequences and index 7 corresponds to the overall average.

votes for SSIM-RDO, which is computed as the number of votes for SSIM-RDO divided by the number of BTCs being considered. While both methods gave the same SSIM scores for the reconstructed videos, the overall proportion preferring SSIM-RDO is about 54%. It should be noted that the SSIM-

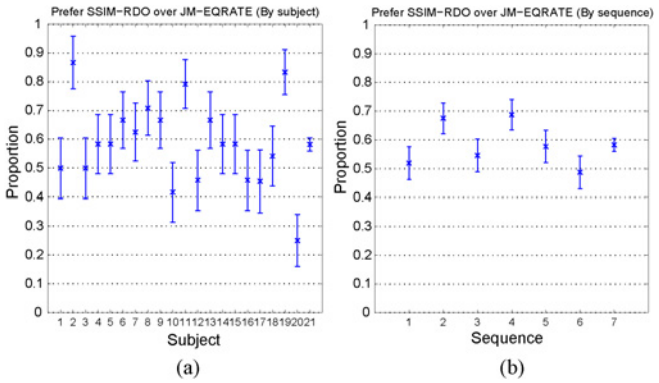


Fig. 7. Error bar plot of results from the 2AFC subjective test comparing SSIM-RDO with JM-EQRATE, showing proportion that preferred SSIM-RDO. The error bars show ± one standard deviation. (a) Ordered by subject, where indices 1–20 correspond to the 20 subjects and index 11 corresponds to the overall average. (b) Ordered by sequence, where indices 1–6 correspond to the six sequences and index 7 corresponds to the overall average.

RDO method uses a 16% less rate, on average, over the set of test sequences. Fig. 7 shows the 2AFC results when comparing SSIM-RDO with JM-EQRATE. In this case, both methods used an equal rate, while the results show that an overall proportion of about 58% preferred SSIM-RDO, with this being more than three standard deviations away from 50%. This suggests that over a variety of subjects and sequences and for the same rate, SSIM-RDO is preferred over JM-EQRATE. It is also interesting to note that the two sequences with the highest proportion preferring SSIM-RDO over JM-EQRATE, $parkrun\_720p$ and $parkjoy\_720p$, also have the largest gain in SSIM when SSIM-RDO is used.

VI. Conclusion

In this paper, we described an approach to incorporate the use of SSIM into RDO to optimize video encoding to target perceptual quality instead of MSE. This can be done by scaling the Lagrange multiplier, which was used in RDO decisions based on local variance statistics. We implemented the proposed approach and demonstrated that it achieved average coding gains of about 9% and 14% for random-access and low-delay encoding configurations, respectively, while maintaining the same perceptual quality as the H.264/AVC reference software encoder as measured by SSIM. At the same time, there was no increase in encoding complexity. This would be useful for video encoder practitioners who wish to achieve further compression gains, while maintaining the same perceptual quality with a low cost of implementation.

While the main goal of our proposed method was to improve compression performance when video quality was measured by SSIM, we also provided some evidence that this also led to perceptual quality improvements. It would be interesting to carry out a more comprehensive set of subjective viewing experiments to quantify this improvement. In addition, evaluating the performance of the approximation of SSIM based on MSE (6) as a perceptual metric would also be of value to the community.

We believe that this paper can be further extended in a number of ways. For example, 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 to have a more smoothly varying Lagrange multiplier over the frame. Also, our future work would also consider temporal qualities such as motion, e.g., [35], and incorporate those features into the RDO framework as well.

References

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

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

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

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

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

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

[7] D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 2217–2220.

[8] L. Tong, F. Dai, Y. Zhang, and J. Li, "Visual security evaluation for video encryption," in *Proc. Int. Conf. Multimedia*, 2010, pp. 835–838.

[9] H. Hofbauer and A. Uhl, "An effective and efficient visual quality index based on local edge gradients," in *Proc. Eur. Workshop Visual Inform. Process.*, 2011, pp. 162–167.

[10] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and H. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614–1624, Nov. 2010.

[11] S. Wang, A. Rehman, W. Wang, S. Ma, and W. Gao, "SSIM-motivated rate distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

[12] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate-distortion optimization using SSIM," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2012, pp. 833–836.

[13] Z. Wang, Q. Li, and X. Shang, "Perceptual image coding based on a maximum of minimal structural similarity criterion," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2. Oct. 2007, pp. 121–124.

[14] 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*. Berlin, Germany: Springer, 2005, pp. 435–441.

[15] 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 *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 3. Oct. 2005, pp. 2673–2678.

[16] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2. May 2006, pp. 913–916.

[17] C.-L. Yang, H.-X. Wang, and L.-M. Po, "Improved inter prediction based on structural similarity in H.264," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, Nov. 2007, pp. 340–343.

[18] C.-L. Yang, R.-K. Leung, L.-M. Po, and Z.-Y. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 4. Nov. 2009, pp. 291–295.

[19] H. H. Chen, Y.-H. Huang, P.-Y. Su, and T.-S. Ou, "Improving video coding quality by perceptual rate-distortion optimization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 1287–1292.

[20] Y.-H. Huang, T.-S. Ou, and H. H. Chen, "Perceptual-based coding mode decision," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jun. 2010, pp. 393–396.

[21] S. Wang, S. Ma, and W. Gao, "SSIM based perceptual distortion rate optimization coding," in *Proc. SPIE Visual Commun. Image Process. Conf.*, vol. 7744. 2010.

[22] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Rate-SSIM optimization for video coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2011, pp. 833–836.
[23] *Test Model 5*, ISO/IEC/JTC1/SC29/WG11, Mar. 1993.
[24] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Feb. 2007.
[25] C.-J. Tsai, C.-W. Tang, C.-H. Chen, and Y.-H. Yu, "Adaptive rate-distortion optimization using perceptual hints," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 1. Jun. 2004, pp. 667–670.
[26] F. Pan, Y. Sun, Z. Lu, and A. Kassim, "Complexity-based rate distortion optimization with perceptual tuning for scalable video coding," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2005, pp. 37–40.
[27] C. Sun, H.-J. Wang, T.-H. Kim, and H. Li, "Perceptually adaptive Lagrange multiplier for rate-distortion optimization in H.264," in *Proc. Future Generation Commun. Networking*, vol. 1. Dec. 2007, pp. 459–463.
[28] C. Sun, H.-J. Wang, and H. Li, "Macroblock-level rate-distortion optimization with perceptual adjustment for video coding," in *Proc. Data Compression Conf.*, Mar. 2008, p. 546.
[29] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.
[30] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Berlin, Germany: Springer, 1992.
[31] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.
[32] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.
[33] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.
[34] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD Curves*, ITU-T SC16/Q6, VCEG-M33, Austin, TX, USA, Apr. 2001.
[35] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1653–1658, Nov. 2010.

Chuohao Yeo (S'05–M'09) received the S.B. degree in electrical science and engineering and the M.Eng. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, USA, in 2002, and the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley (UC Berkeley), USA, in 2009.

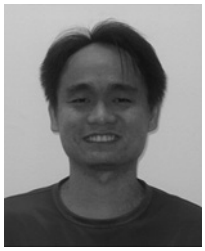
From 2005 to 2009, he was a Graduate Student Researcher with the Berkeley Audio Visual Signal Processing and Communication Systems Laboratory, UC Berkeley. Since 2009, he has been a Scientist with the Signal Processing Department, Institute for Infocomm Research, Singapore, where he leads a team that is actively involved in HEVC standardization activities. His current research interests include image and video processing, coding and communications, distributed source coding, and computer vision.

Dr. Yeo was a recipient of the Singapore Government Public Service Commission Overseas Merit Scholarship from 1998 to 2002, and a recipient of Singapore's Agency for Science, Technology, and Research National Science Scholarship from 2004 to 2009. He was a recipient of the Best Student Paper Award at SPIE VCIP 2007 and the Best Short Paper Award at ACM MM 2008.

Hui Li Tan (M'09) received the B.Sc. degree in applied mathematics from the National University of Singapore, Singapore, in 2007, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering.

Since 2007, she has been with the Signal Processing Department, Institute for Infocomm Research, Singapore, as a Research Engineer. Her current research interests include image/video quality assessment and compression.

Ms. Tan is a recipient of the A*STAR Scientific Staff Development Award, which funds her Ph.D. studies.



Yih Han Tan (M'11) received the B.Eng. (first-class Honors) and Ph.D. degrees in electrical engineering from the National University of Singapore, Singapore, in 2005 and 2010, respectively.

Since 2010, he has been with the Signal Processing Department, Institute for Infocomm Research, Singapore, as a Research Scientist. His current research interests include video processing and delivery.