SSIM-Motivated Rate-Distortion Optimization for Video Coding

Shiqi Wang, Abdul Rehman, *Student Member, IEEE,* Zhou Wang, *Member, IEEE,* Siwei Ma, *Student Member, IEEE,* and Wen Gao, *Fellow, IEEE*

***Abstract*—We propose a rate-distortion optimization (RDO) scheme based on the structural similarity (SSIM) index, which was found to be a better indicator of perceived image quality than mean-squared error, but has not been fully exploited in the context of image and video coding. At the frame level, an adaptive Lagrange multiplier selection method is proposed based on a novel reduced-reference statistical SSIM estimation algorithm and a rate model that combines the side information with the entropy of the transformed residuals. At the macroblock level, the Lagrange multiplier is further adjusted based on an information theoretical approach that takes into account both the motion information content and perceptual uncertainty of visual speed perception. Finally, the mode for H.264/AVC coding is selected by the SSIM index and the adjusted Lagrange multiplier. Extensive experiments show that the proposed scheme can achieve significantly better rate-SSIM performance and provide better visual quality than conventional RDO coding schemes.**

***Index Terms*—H.264/AVC coding, Lagrange multiplier, rate-distortion optimization, reduced-reference image quality assessment, structural similarity (SSIM) index.**

I. Introduction

VIDEO CODECS are primarily characterized in terms of the throughput of the channel and perceived distortion of the reconstructed video. The main task of the video codec is to convey the sequence of images with minimum possible perceived distortion within available bit rate. Alternatively, it can be posed as a communication problem of conveying the sequence with minimum possible rate while maintaining a specific perceived distortion level. In both versions of the problem, the fundamental issue is to obtain the best tradeoff between the rate and perceived distortion. The process used to achieve this objective is commonly known as rate-distortion optimization (RDO), which can be expressed by minimizing the perceived distortion D with the number of used bits R subjected to a constraint R_c [1] as follows:

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c. \tag{1}$$

This is a typical constrained optimization problem which is generally solved using two methods: Lagrangian optimization and dynamic programming. In practice, the computation complexity of dynamic programming is often too high and is used only when direct Lagrangian optimization is difficult.

Lagrangian optimization technique converts the constrained optimization problem (1) to an unconstrained optimization problem [1], which can be expressed as

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R \tag{2}$$

where J is called the rate-distortion (RD) cost and the rate R is measured in number of bits per pixel. λ is known as the Lagrange multiplier which controls the tradeoff between R and D.

Since our knowledge of the human visual system (HVS) and statistics of natural images remains limited, the perceived distortion D is difficult to measure. In practice, distortion models such as sum of absolute difference (SAD) and mean-squared error (MSE) are used in most actual comparisons [2]. Many RDO algorithms were proposed along this line. The representative work includes RD-optimized transform [3], RD-optimized quantization [4], and the dependent joint RDO using soft decision quantization [5], [6]. However, the distortion measures such as SAD and MSE are widely criticized for not correlating well with perceived quality. Recently, a lot of work has been done to develop objective quality assessment measures which can accurately reflect the perceived distortion. The most prominent ones include the structural similarity (SSIM) index [7], visual information fidelity criterion [8], and visual signal-to-noise ratio [9]. Among them, SSIM has been preferred due to its best tradeoff among accuracy, simplicity, and efficiency [10]. The correlation of SSIM with mean opinion score, obtained using subjective tests, has been repeatedly proven in the literature. In this paper, we focus on solving (2), where SSIM is used to define the measure of perceived distortion and λ is adapted at both frame and macroblock

Manuscript received November 19, 2010; revised February 2, 2011, May 22, 2011 and July 6, 2011; accepted August 11, 2011. Date of publication September 15, 2011; date of current version April 2, 2012. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, in part by the Ontario Early Researcher Award Program, in part by the National Science Foundation of China, under Grants 60833013 and 60803068, and in part by the National Basic Research Program of China (973 Program), under Grants 2009CB320903 and 2009CB320904. This paper was recommended by Associate Editor M. Hannuksela.

S. Wang is with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China, and with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

A. Rehman and Z. Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zhouwang@ieee.org).

S. Ma and W. Gao are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: swma@pku.edu.cn).

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.2011.2168269

(MB) levels by taking the properties of the input sequences (statistical properties of residuals, structural information, motion information, etc.) into consideration.

In order to achieve optimal RD performance, it is very important to carefully choose λ and the best coding mode. To achieve a good balance between R and D, in the H.264/AVC [11] coding environment, the Lagrange multiplier is suggested to be [12] as follows:

$$\lambda = 0.85 \cdot 2^{\frac{Q_{H.264}-12}{3}} \tag{3}$$

where $Q_{H.264}$ is the quantization parameter (QP). This suggestion was proposed based on empirical results and typical RD models [1], [13]. It also suggests that λ is a function of QP only and therefore is independent of the frame properties, which simplifies the problem but may not result in optimal λ as some MBs could be more important compared to the others [14]. This motivated us to adapt λ according to the video sequences at both frame and MB levels.

In the literature, significant progress has been made to adapt λ on frame level when MSE is used as the distortion measure. In [15], Chen *et al.* developed an adaptive λ estimation algorithm by modeling the R and D in ρ domain, where ρ is defined as the percentage of zero coefficients among quantized transform residuals [16]. In [17], Laplace distribution-based rate and distortion models were established to derive λ for each frame dynamically.

Many rate control algorithms such as [18] and [19] showed that better performance and rate control can be achieved by modifying λ on MB level than having the same Lagrange multiplier for all MBs in a frame. In [20] and [21], the authors claimed that fixing the same Lagrange multiplier for the whole frame may not be accurate enough to capture the nature of motion, and therefore a context-adaptive Lagrange multiplier (CALM) selection scheme was introduced. However, all these methods ignored the perceptual aspect in the RDO scheme by adopting SAD/MSE as the measures of perceived distortion.

Recently, a number of video coding methods aiming to incorporate the properties of the HVS have been proposed. Yang *et al.* proposed a just noticeable distortion (JND) model for motion estimation and residue filtering process in [22] and [23]. A foveated JND model was employed in [24] for optimizing the QP and Lagrange multiplier. To incorporate perceptual information into the MB-based adaptive RDO scheme, three distortion sensitivity models were built into the RDO framework in [25]. Pan *et al.* [26] proposed a content complexity-based Lagrange multiplier selection scheme for scalable video coding.

Since SSIM is proven to be more effective in quantifying the suprathreshold compression artifacts, such as artifacts that distort the structure of an image [27], it was incorporated into motion estimation, mode selection, and rate control in hybrid video coding [28]–[38]. For intra frame coding, new SSIM-based RDO schemes were proposed in [28]–[30]. In [31]–[33], the authors developed SSIM-based RDO schemes for inter frame prediction and mode selection. However, following the method proposed in [13], the Lagrange multiplier was determined only by QP values in these schemes. Recently, content-adaptive Lagrange multiplier selection schemes were proposed in [34]–[37]. These algorithms employed a rate–SSIM curve to describe the relationship between SSIM and rate, which is given by $D = \zeta R^{\varepsilon}$, where ζ and ε are two fitting parameters which account for the RD characteristics. Subsequently, the key frames are identified and encoded twice with MSE-based RDO in the sequences to obtain the best parameters ζ and ε. However, two-pass encoding of the key frames will bring more additional complexities to the encoder. More importantly, this scheme is based on the assumption of constant RD characteristics in a short time period and uses a periodic refreshment technique to refresh the parameters, which may not be accurate in general.

In this paper, we use SSIM as the distortion measure and propose an adaptive RDO scheme for mode selection. The three main contributions of our work are as follows.

1) We employ SSIM as the distortion measure in the proposed mode selection scheme, where both the current MB to be coded and neighboring pixels are taken into account to fully exploit the properties of SSIM.
2) At the frame level, we present an adaptive Lagrange multiplier selection scheme based on a novel statistical reduced-reference (RR) SSIM model and a source-side information combined rate model.
3) At the MB level, we present a Lagrange multiplier adjustment scheme, where the scale factor for each MB is determined by an information theoretical approach based on the motion information content and perceptual uncertainty of visual speed perception.

II. SSIM-Based RDO

Analogous to (2), the SSIM motivated RDO problem can be defined as

$$\min\{J\} \quad \text{where} \quad J = (1 - \text{SSIM}) + \lambda \cdot R. \tag{4}$$

The spatial domain SSIM index [7] is based on similarities of local luminance, contrast, and structure between a reference image and a distorted image. Given two local image patches $\mathbf{x}$ and $\mathbf{y}$, the local SSIM index is defined as

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where μ_x, σ_x, and σ_{xy} are the mean, standard deviation, and cross correlation between the two patches, respectively. C_1 and C_2 are used to avoid instability when the means and variances are close to zero. SSIM index of the whole image is obtained by averaging the local SSIM indices calculated using a sliding window.

In the conventional mode selection process, the final coding mode is determined by the number of entropy coding bits and the distortion of the residuals, while the properties of the reference image are ignored. Unlike MSE, the SSIM index is totally adaptive according to the reference signal [7]. Therefore, the properties of video sequences can also be exploited when using SSIM to define the distortion model.

In H.264/AVC, the encoder processes a frame of video in units of nonoverlapping MBs. However, SSIM index is

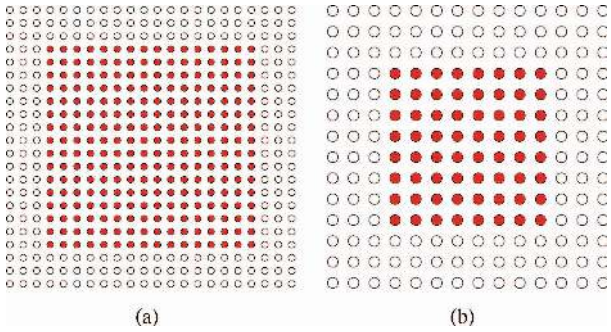

Fig. 1. Illustration of using surrounding pixels to calculate the SSIM index. Solid pixels: to be encoded. Hollow pixels: surrounding pixels from the input frame. (a) Y component. (b) Cb and Cr components.

meant to be calculated with the help of overlapping sliding windows, which are separated by one pixel. To bridge this gap, we calculate the SSIM index between the reconstructed MB and the original MB using an extended MB, which includes the current MB to be coded and the surrounding pixels, as illustrated in Fig. 1. Within this extended MB, we use a small sliding window which moves pixel by pixel to calculate the SSIM index. The size of the sliding window used to calculate SSIM is set to be 4×4. Therefore, we extend the MB boundaries for three pixels in each direction. For Y component, the SSIM index of the current 16×16 MB to be encoded is calculated within a 22×22 extended MB by using the sliding window. In case of 4:2:0 format, for Cb and Cr components the SSIM index is calculated within a 14×14 extended block. Additional benefit of this approach is that it helps us to alleviate the problem of discontinuities at the MB boundaries. When the MB is on the frame boundaries, we ignore the surrounding pixels in the distortion calculation and only use the MB to be coded for comparison.

Finally, SSIM indices of Y, Cb, and Cr components are weighted averaged to obtain a single measure of SSIM as follows:

$$\text{SSIM} = W_Y \cdot \text{SSIM}_Y + W_{Cb} \cdot \text{SSIM}_{Cb} + W_{Cr} \cdot \text{SSIM}_{Cr} \quad (6)$$

where W_Y, W_{Cb}, and W_{Cr} are the weights of Y, Cb, and Cr components, respectively, and are defined as $W_Y = 0.8$ and $W_{Cb} = W_{Cr} = 0.1$, respectively, [39].

III. Frame Level Lagrange Multiplier Selection

From (4), the Lagrange parameter is obtained by calculating the derivative of J with respect to R, then setting it to zero, and finally solving for λ as follows:

$$\frac{dJ}{dR} = -\frac{\text{dSSIM}}{dR} + \lambda = 0 \quad (7)$$

which yields

$$\lambda = \frac{\text{dSSIM}}{dR} = \frac{\frac{\text{dSSIM}}{dQ}}{\frac{dR}{dQ}} \quad (8)$$

where Q is the quantization step. This implies that, in order to estimate λ before actually encoding the current frame, we need to establish accurate SSIM and rate models.

In video coding, the most common models for the distribution of transformed residuals are Laplace distribution [17], generalized Gaussian distribution (GGD) [40], and Cauchy distribution [41]. Although GGD is a good statistical model to describe the discrete cosine transform (DCT) coefficients, it has more control parameters and closed-form expression of the distortion model cannot be obtained [40]. For Cauchy distribution, the mean and variance are not defined, which makes it inappropriate for this framework [17]. The Laplace distribution, which is a special case of GGD, does not suffer from these problems and achieves a good tradeoff between model fidelity and the complexity. Therefore, we model the transformed residuals x with the Laplace distribution given by

$$f_{\text{Lap}}(x) = \frac{\Lambda}{2} \cdot e^{-\Lambda \cdot |x|} \quad (9)$$

where Λ is called the Laplace parameter.

A. RR SSIM Model

SSIM is a full-reference (FR) measure that requires both the reference and distorted frames to compute. It cannot be directly applied in this framework because the distorted frame is not available. Therefore, we develop a RR quality assessment algorithm which requires a set of RR features extracted from the reference frame for SSIM estimation. The RR-SSIM estimation method based on a multiscale multiorientation divisive normalization transform (DNT) is proposed in [42] and achieves high SSIM estimation accuracy. However, it cannot be directly employed due to the high computational complexity of DNT. We use a similar approach here, but extract features from DCT coefficients instead.

FR DCT domain SSIM index was first presented by Channappayya *et al.* [43] as follows:

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \left\{1 - \frac{(X(0) - Y(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right\} \times \left\{1 - \frac{\sum_{k=1}^{N-1} (X(k) - Y(k))^2}{\sum_{k=1}^{N-1} (X(k)^2 + Y(k)^2) + N \cdot C_2}\right\} \quad (10)$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients for the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. This equation implies that the SSIM index is represented by the product of two terms, characterizing the distortions of the DC and AC coefficients, respectively. Moreover, the squared errors of DC and AC coefficients are normalized by their respective energy.

To develop the RR-SSIM model, we divide each frame into nonoverlapping blocks and the size of each block is set to be 4×4. Then DCT transform is performed on each block. In this way, we can obtain the statistical properties of the reference signal, which is consistent with the design philosophy of the SSIM index. Furthermore, we group the DCT coefficients having the same frequency from each 4×4 DCT window into one subband, which results in 16 subbands. Motivated by the DCT domain SSIM index, the new RR distortion measure

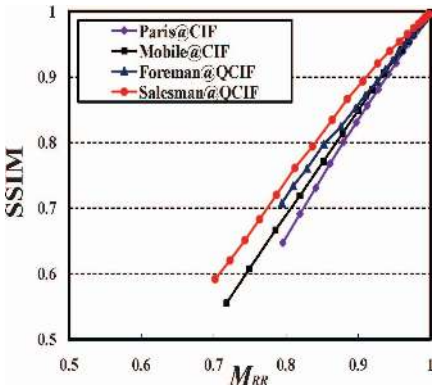

Fig. 2. Relationship between SSIM and M_{RR} for different sequences.

is defined as

$$M_{\mathrm{RR}} = \left(1 - \frac{D_0}{2\sigma_0^2 + C_1}\right)\left(1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{D_i}{2\sigma_i^2 + C_2}\right) \quad (11)$$

where σ_i is the standard deviation of the ith subband and N is the block size. D_i represents the MSE between the original and distorted frames in the ith subband, and is calculated as follows:

$$D_i = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} x_i^2 f_{Lap}(x_i)dx_i + 2\sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}(x_i - nQ)^2 f_{Lap}(x_i)dx_i \quad (12)$$

where γ is the rounding offset in the quantization. Fig. 2 presents the relationship between the RR distortion measure M_{RR} and the corresponding SSIM index for different sequences. The QP values in Fig. 2 cover a wide range from 0 to 50 with an interval of 2. The SSIM index and M_{RR} are calculated by averaging the respective values of individual frames. Interestingly, M_{RR} exhibits a nearly perfect linear relationship with SSIM. We regard this as an outcome of the similarity between their design principles. The clean linear relationship also helps us to design an SSIM predictor based on M_{RR} because the remaining job is just to estimate the slope and intercept of the straight line. More specifically, an RR-SSIM estimator can be written as

$$\hat{S} = \alpha + \beta \cdot M_{\mathrm{RR}}. \quad (13)$$

The proposed RR-SSIM model is totally based on the features extracted from the original frames in the DCT domain and the residuals. It can be observed from Fig. 2 that the slopes for different video sequences are different. Thus, before coding the current frame we should first estimate the parameters α and β. This requires the knowledge of two points on the straight line relating $\hat{S}$ and M_{RR}. We use (1, 1) as one of the points as it is always located on the line and also because it does not require any computation. This solves half of the problem as we still need $\hat{S}$ and M_{RR} of the second point. The SSIM index $\hat{S}$ and Laplace parameter for each subband Λ_i is not available

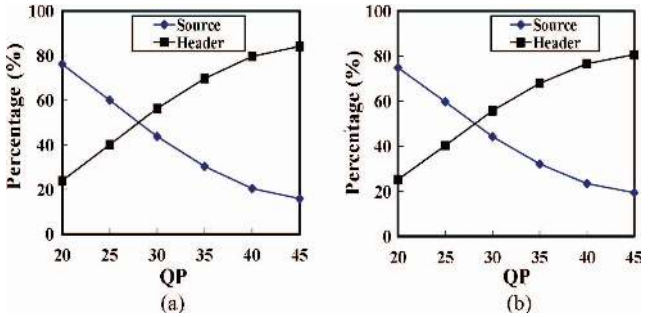

Fig. 3. Average percentages of header bits and source bits at various QPs. (a) *Foreman* (IPP). (b) *Foreman* (IBP).

since we have not encoded the frame yet. Therefore, we estimate them from the previous frames of the same type. The estimation details are provided in Section V. The distortion measure M_{RR} can be calculated by incorporating (12) into (11), and the standard deviation of the ith subband σ_i is calculated by DCT transform of the original frame. This procedure provides us with the second point required to find out α and β.

B. Proposed Rate Model

Our rate model is derived based on an entropy model that excludes the bit rate of the skipped blocks [17] as follows:

$$H = (1 - P_s)\cdot\left[-\frac{P_0 - P_s}{1 - P_s}\cdot\log_2\frac{P_0 - P_s}{1 - P_s} - 2\sum_{n=1}^{\infty}\frac{P_n}{1 - P_s}\cdot\log_2\frac{P_n}{1 - P_s}\right] \quad (14)$$

where P_s is the probability of the skipped blocks, P_0 and P_n are the probabilities of transformed residuals quantized to the zeroth and nth quantization levels, respectively, which can be modeled by the Laplace distribution as follows:

$$P_0 = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} f_{\mathrm{Lap}}(x)dx \quad (15)$$

$$P_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} f_{\mathrm{Lap}}(x)dx. \quad (16)$$

Subsequently, supposing the rate model in [17] to be R^*, a linear relationship between $ln(R^*/H)$ and $\Lambda \cdot Q$ is observed, where R^* is based on the assumption of negligible side information. However, in H.264/AVC, the side information (or header bits) may take a large portion of the total bit rate, especially in low bit rate video coding scenario [44], as illustrated in Fig. 3. Therefore, in our rate model, the side information is also taken into consideration. Notice that for the same quantization step, a larger Λ indicates small residuals, leading to a larger proportion of the side information. For total bit rate R, there is also an approximately linear relationship between $ln(R/H)$ and $\Lambda \cdot Q$, as can be seen in Figs. 4 and 5. Also, the relationship is totally consistent with the effect of dependent entropy coding and side information. In high bit rate video coding scenario, the effect of dependent entropy coding compensates the side information and $ln(R/H)$ approaches zero, while for low bit rate $ln(R/H)$ becomes

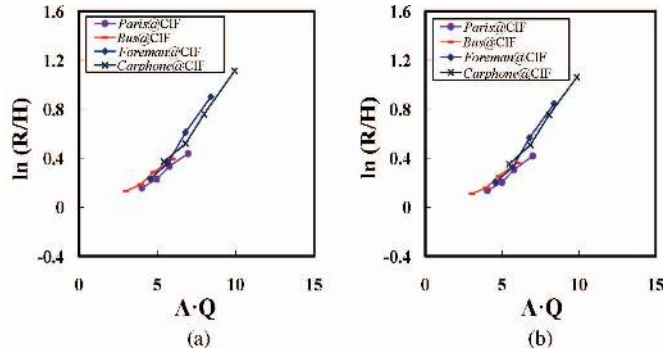

Fig. 4. Relationship between $ln(R/H)$ and $\Lambda \cdot Q$ for different sequences [group of picture (GoP) structure: IPP]. (a) CAVLC entropy coding. (b) CABAC entropy coding.

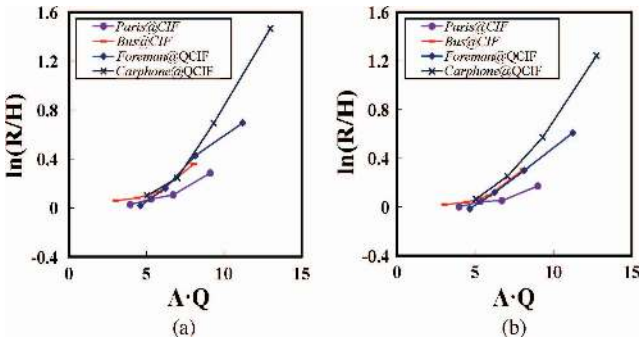

Fig. 5. Relationship between $ln(R/H)$ and $\Lambda \cdot Q$ for B frame of different sequences. (a) CAVLC entropy coding. (b) CABAC entropy coding.

larger because of the dominating effect of side information, as illustrated in Figs. 4 and 5.

Fig. 6 shows that the header bits change monotonically with the source bits. Consequently, the final rate model R can be approximated by

$$R = H \cdot e^{\xi \Lambda Q + \psi} \tag{17}$$

where ξ and ψ are two parameters to control the relationship between $ln(R/H)$ and $\Lambda \cdot Q$. It can be observed from Figs. 4 and 5 that the parameters ξ and ψ are not very sensitive to the video content. Also, for B frames the slope is smaller than that of the I and P frames. It is mainly due to the fact that in case of B frames the residuals are relatively smaller, resulting in a larger value of Λ. Therefore, for both context-adaptive variable length coding (CAVLC) and context-adaptive binary arithmetic coding (CABAC) entropy coding methods, ξ and ψ, are set empirically to be

$$\xi = \begin{cases} 0.03, & \text{B frame} \\ 0.07, & \text{otherwise} \end{cases} \quad \psi = \begin{cases} -0.07, & \text{B frame} \\ -0.1, & \text{otherwise.} \end{cases} \tag{18}$$

There is one limitation of the proposed rate model. At low bit rate, the skip mode is selected more often and hence the source rate of sequences coded at low bit rate is close to zero. The proposed rate model does not work well in such a situation because the side information modeling is based on the source rate. Efficient model of the side information is still an open problem.

Based on the statistical model of the transformed residuals, we obtain the final closed-form solutions of the R and D models. It is observed that the R and D models are functions of two sets of variables: Q and the other variables that describe the inherent properties of the video sequences such as Λ_i and σ_i. When Q varies within a small range, it can be regarded as independent of the other variables [17]. Consequently, before coding the current frame, the frame level Lagrange multiplier can be determined by incorporating the closed-form expressions of R and D into (8).

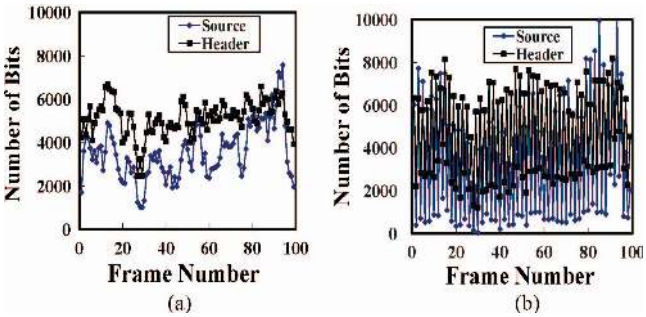

Fig. 6. Source bits and header bits for each frame at QP=30. (a) *Foreman* (IPP). (b) *Foreman* (IBP).

IV. MB Level Lagrange Multiplier Adjustment

Natural video sequence is not just a stack of independent still images, it also contains critical motion information that relates these images. Therefore, the frames in a natural video cannot be considered independently as far as HVS is concerned. Perception of motion information between frames plays an important role toward video quality assessment by HVS. In the conventional video coding framework, motion estimation is performed solely for motion compensation purposes in order to reduce the amount of data to be transmitted. Once the residual frame is calculated, all the MBs are considered equally for bit allocation. This does not conform with HVS, as perceptual information content is different in each MB that depends on the motion information content and perceptual uncertainty in video signals [14]. In [18], the relationship among the Lagrange multiplier λ, the corresponding rate R, and the distortion D was analyzed. A larger λ results in a higher D and a lower R and vice versa, which implies that we can influence the rate and perceptual distortion of each MB by adjusting its Lagrange multiplier. This motivated us to assign more bits to the MBs which are more important as far as perceptual information content is concerned. Lagrange multiplier is adjusted with the help of a spatiotemporal weighting factor, η, which increases with the information content and decreases with the perceptual uncertainty.

We employ the scheme proposed in [14] which uses an information communication framework to model the visual perception. We define the relative motion vector, v_r, as the difference between the absolute motion vector, v_a, and global background motion vector v_g: $v_r = v_a - v_g$.

In [45], the visual judgment of the speed of motion is modeled by combining some prior knowledge of the visual world and the current noisy measurements. Based on this approach, the motion information content is estimated by the

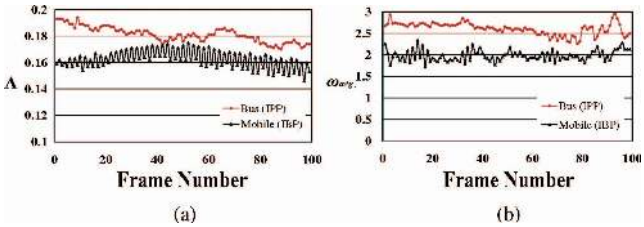

Fig. 7. Illustration of the parameters Λ and ω_{avg} for each frame. (a) Laplace distribution parameter Λ for each frame in *Bus* (IPP) and *Mobile* (IBP) with CIF format. (b) Average weight ω_{avg} for each frame in *Bus* (IPP) and *Mobile* (IBP) with CIF format.

self-information of the relative motion $I = \varphi \log v_r + \nu$, where φ, ν are the parameters of power-law function for the distribution of relative motion and are determined based on psychophysical study conducted in [45].

The perceptual uncertainty is estimated by the entropy of the likelihood function of the noisy measurement, which is given by $U = \log v_g - \tau \log c + \delta$, where τ and δ are the parameters of the log-normal distribution, used to determine perceptual uncertainty, determined based on psychophysical study [45]. The spatiotemporal importance weight function is given by

$$\omega = I - U = \varphi \log v_r + \nu - \{\log v_g - \tau \log c + \delta\}. \tag{19}$$

The contrast measure c can be derived by [14]

$$c = 1 - e^{-(c'/\phi)^{\kappa}} \tag{20}$$

$$c' = \frac{\sigma_p}{\mu_p + \mu_0} \tag{21}$$

where σ_p and μ_p are computed within the MB, representing the standard deviation and the mean, respectively. κ and ϕ are constants that control the slope and the position of the functions, respectively, [14], and are used to take into account the contrast response saturation effect at small and large contrast values. μ_0 is a constant to avoid instability near 0.

The global motion does not influence the perceptual weight of each MB, thus the weight for each MB is defined as follows:

$$\omega = \log\left(1 + \frac{v_r}{v_0}\right) + \log\left(1 + \frac{c}{c_0}\right) \tag{22}$$

where v_0 and c_0 are constants used to avoid unstable evaluation of the weight function when the relative motion v_r and the local contrast c may be close to zero. Note that this weight function increases monotonically with the relative motion and the local contrast, which is in line with the philosophy of visual attention. Consequently, the MBs with higher weights should be allocated more bits and vice versa. This motivated us to adjust the Lagrange multiplier by

$$\lambda' = \eta \cdot \lambda. \tag{23}$$

To determine the adjustment factor η for every MB, we calculate the weight based on the local information, then η is determined in a similar manner as in [19]

$$\eta = \left(\frac{\omega_{avg}}{\omega}\right)^{\epsilon}. \tag{24}$$

The parameter ω_{avg} represents the average weight of the current frame and ϵ is set to be 0.25 as in [19]. Following [14], we set v_0=0.32 and c_0=0.70.

Algorithm 1: Summary of the proposed RDO (GoP structure: IPP)

begin
  Calculate λ_i for the ith frame **switch** *the value of i*
  do
    case *0, 1, 2, 3*
      $\lambda_i \leftarrow \lambda_{\text{HR}}$
    end
    otherwise
      1) DCT transform of the input frame.
      2) $\lambda_i \leftarrow \begin{cases} \lambda_{\text{HR}}, & H = 0 \\ \frac{\frac{\text{dSSIM}}{dQ}}{\frac{dR}{dQ}}, & \text{otherwise.} \end{cases}$
    end
  end
end
begin
  For each MB in the frame
    1) Calculate the scale factor at MB level η.
    2) Adjust the Lagrange multiplier:
       $\lambda_i' \leftarrow \eta \cdot \lambda_i$.
    3) Calculate the RD cost for each Mode k:
       $J_k \leftarrow 1 - \text{SSIM}_k + \lambda_i' \cdot R_k$.
    4) Select the Mode j with minimal RD cost.
    5) Encode the MB with Mode j.
end
begin
  Update Λ_i, $\hat{S}$, Λ, ω_{avg}, and v_g.
end

V. Implementation Issues

The Lagrange parameter should be determined before coding the current frame in order to perform RDO. However, the parameters Λ_i,$\hat{S}$, Λ, ω_{avg}, and v_g can only be calculated after coding the current frame. As shown in Fig. 7, the parameters of the frames with the same coding type varies smoothly even for sequences of high motion. This is due to the fact that the inherent properties of the input sequences can be considered unchanged during a short period of time. Therefore, we estimate them by averaging their three previous values from the frames coded in the same manner, that is

$$\hat{\Lambda}_i^j = \frac{1}{3}\sum_{n=1}^{3} \Lambda_i^{j-n} \tag{25}$$

where the j indicates the frame number. The global motion vector, v_g, is derived using maximum likelihood estimation which finds the peak of the motion vector histogram [46].

To encode the first few frames, the adaptive Lagrange multiplier selection method is not used since it is difficult to estimate Λ_i, $\hat{S}$, Λ, ω_{avg}, and v_g. Motivated by the high rate λ selection method [1], [13], we derive a Lagrange multiplier based on the high bit rate assumption for such a situation.

With the high rate assumption, the SSIM index in the DCT domain can be approximated by [47]

$$\begin{aligned} &E[\text{SSIM}(\mathbf{x}, \mathbf{y})] \\ &\approx \{1 - E[X(0) - Y(0)]^2 \times E[\frac{1}{2X(0)^2 + N \cdot C_1}]\} \\ &\times \{1 - E[\sum_{k=1}^{N-1} (X(k) - Y(k))^2] \\ &\times E[\frac{1}{2\sum_{k=1}^{N-1} X(k)^2 + N \cdot C_2}]\} \end{aligned} \tag{26}$$

where E denotes the mathematical expectation operator. Furthermore, in (27), we use D_{dc}, D_{ac}, E_{dc}, E_{ac} to simplify this equation and the expectation of SSIM index can be rewritten as

$$\begin{aligned} D_{dc} &= E[X(0) - Y(0)]^2 \\ E_{dc} &= E[\frac{1}{2X(0)^2 + N \cdot C_1}] \\ D_{ac} &= E[\sum_{k=1}^{N-1} (X(k) - Y(k))^2] \\ E_{ac} &= E[\frac{1}{2\sum_{k=1}^{N-1} X(k)^2 + N \cdot C_2}] \end{aligned} \tag{27}$$

$$E[\text{SSIM}(x, y)] = (1 - E_{dc} \times D_{dc}) \times (1 - E_{ac} \times D_{ac}). \tag{28}$$

If the high rate assumption is valid, the source probability distribution can be approximated as uniform distribution and the MSE can be modeled by [48]

$$D = s \cdot Q^2. \tag{29}$$

The Lagrange multiplier based on the high rate assumption rate and MSE models is then given by [13]

$$\hat{\lambda}_{\text{HR}} = -\frac{dD}{dR} = c \cdot Q^2 \tag{30}$$

where c is a constant. Therefore, the general form of λ_{HR} for SSIM-based RDO can be derived by calculating the derivative of SSIM with respect to R (8), which leads to

$$\lambda_{\text{HR}} = \frac{d(E_{ac} \cdot E_{dc} \cdot D_{ac} \cdot D_{dc})}{dR} - \frac{d(E_{dc} \cdot D_{dc})}{dR} - \frac{d(E_{ac} \cdot D_{ac})}{dR}. \tag{31}$$

Although E_{ac} and E_{dc} are based on the properties of the frames, to provide a constant solution for SSIM-based RDO in the first few frames, we derive a general solution for them. Considering (29)–(31), the constant Lagrange multiplier for SSIM-based RDO can be expressed by

$$\lambda_{\text{HR}} = a \cdot Q^2 - b \cdot Q^4. \tag{32}$$

The values for a and b are determined empirically by experimenting with SSIM and the rate models as follows:

$$a = \begin{cases} 2.1 \times 10^{-4}, & \text{B frame} \\ 7 \times 10^{-5}, & \text{otherwise} \end{cases} \tag{33}$$

$$b = \begin{cases} 1.5 \times 10^{-9}, & \text{B frame} \\ 5 \times 10^{-10}, & \text{otherwise.} \end{cases} \tag{34}$$

In our rate model (17), the modeling of side information is totally based on the source rate. In the extreme case, e.g., when the source rate is zero, this rate model will fail because the header bit cannot be zero in the real video coding scenario. Therefore, we propose an escape method to keep a reasonable performance, where the Lagrange multiplier is given by

$$\lambda = \begin{cases} \lambda_{\text{HR}}, & H = 0 \\ \frac{\frac{\text{dSSIM}}{dQ}}{\frac{dR}{dQ}}, & \text{otherwise.} \end{cases} \tag{35}$$

We summarize the whole process of proposed RDO scheme for IPP coding structure in Algorithm 1. Similar process applies to IBP as well. It can be observed that the complexities introduced by the proposed method are only moderate. The additional computations are the DCT transform of the original frame, the calculation of the parameters (Λ_i, $\hat{S}$, Λ, ω_{avg}, and v_g) and the calculation of SSIM for each mode.

VI. Validations

To validate the accuracy and efficiency of the proposed perceptual RDO scheme, we integrate our mode selection scheme into the H.264/AVC reference software JM15.1 [49]. All test video sequences are in YCbCr 4:2:0 format. In this section, we present the results of three experiments which are used to validate various aspects of the proposed perceptual RDO algorithm. In the first experiment, we verify the proposed RR-SSIM model by comparing estimated SSIM values with actual SSIM values. In the second experiment, the performance of the proposed perceptual RDO algorithm is evaluated and compared with that of the conventional RDO scheme. In the third experiment, we compare the proposed method with state-of-the-art SSIM and MSE-based RDO schemes.

A. Comparison Between Estimated and Actual SSIM

In this section, we compare the estimated (RR) and actual (FR) values of the SSIM index for different sequences with a set of various QP values. The first frame is I-frame while all the rest are inter-coded frames. Equation (13) suggests that we first need to calculate the parameters α and β which vary across different video content. Thus, for each frame, we calculate the slope with the help of two points. $(\hat{S}, M_{\text{RR}})$ and $(1, 1)$, where the point $(\hat{S}, M_{\text{RR}})$ is obtained by setting QP=40, the middle point among the quantization steps used for testing the proposed scheme. Once α and β are determined, we can use (13) to estimate SSIM for other QP values. Fig. 8 plots the estimated and actual values of the SSIM index for various values of QP. It is observed that the proposed SSIM model is robust and accurate for different video contents with different resolutions. Moreover, we have also calculated the Pearson linear correlation coefficient (PLCC) and mean absolute error

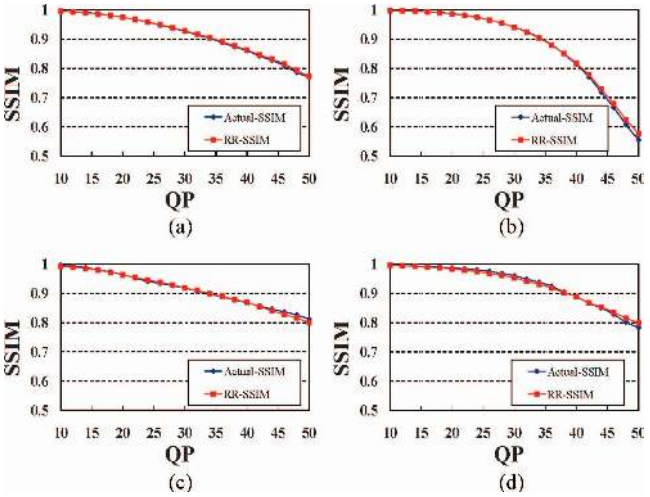

Fig. 8. Comparison of the actual FR-SSIM and estimated RR-SSIM values. (a) *Foreman* at CIF (IPP). (b) *Mobile* at CIF (IBP). (c) *Highway* at QCIF (IPP). (d) *Akiyo* at QCIF (IBP).

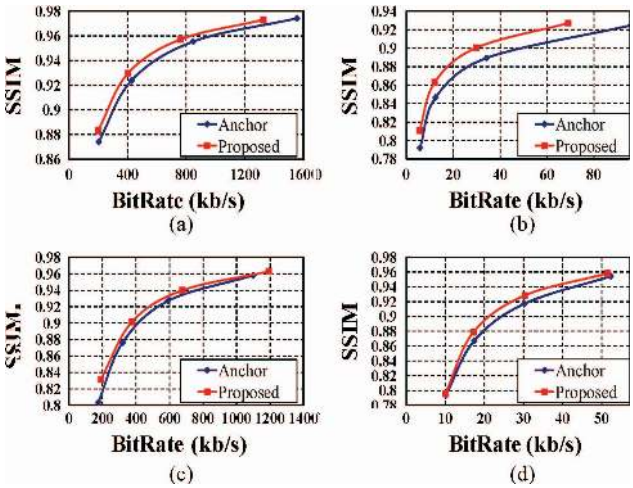

Fig. 9. Performance comparisons of different RDO algorithms for sequences with CABAC entropy coding method. (a) *Flower* at CIF (IPP). (b) *Bridge* at QCIF (IPP). (c) *Bus* at CIF (IBP). (d) *Salesman* at QCIF (IBP).

TABLE I
MAE AND PLCC BETWEEN FR-SSIM AND RR-SSIM ESTIMATION FOR DIFFERENT SEQUENCES

Sequences	GoP Structure	PLCC	MAE
Foreman (CIF)	IPP	0.999	0.002
News (CIF)	IPP	0.999	0.002
Mobile (CIF)	IBP	0.999	0.004
Paris (CIF)	IBP	0.999	0.003
Highway (QCIF)	IPP	0.998	0.003
Suize (QCIF)	IPP	0.998	0.004
Carphone (QCIF)	IBP	0.997	0.006
Akiyo (QCIF)	IBP	0.998	0.005
City (720P)	IPP	0.994	0.015
Crew (720P)	IBP	0.997	0.009
All		0.996	0.005

(MAE) between FR-SSIM and RR-SSIM which are given in Table I for ten different sequences. The values suggest that the proposed RR-SSIM model achieves high accuracy for different sequences.

B. Performance Evaluation of the Proposed Algorithms

We compare the RD performance of our proposed perceptual RDO algorithm and the conventional RDO with distortion measured in terms of SSIM, weighted SSIM, and peak signal-to-noise ratio (PSNR). The three quantities for the whole video sequence are obtained by simply averaging the respective values of individual frames. The size of sliding window to calculate the SSIM index is set to be 8×8. In this experiment, we employ the method proposed in [50] to calculate the differences between two RD curves.[1] Furthermore, the weighted SSIM index is defined as [14]

$$\mathrm{SSIM}_{\omega} = \frac{\sum_x \sum_y \omega(x,y)\mathrm{SSIM}(x,y)}{\sum_x \sum_y \omega(x,y)} \tag{36}$$

[1]Since R-SSIM curve exhibits a similar shape as R-PSNR curve, we use the same tool proposed in [50] to calculate the average of SSIM differences.

where $\omega(x, y)$ indicates the weight value for (x, y) as defined in (22). The SSIM indices of Y, Cb, and Cr components are combined according to (6). Since the SSIM_{ω} takes the motion information into account, it is more accurate for perceptual video quality assessment [14].

For coding complexity overhead evaluation, we calculate ΔT as follows:

$$\Delta T = \frac{T_{\mathrm{pro\_RDO}} - T_{\mathrm{org\_RDO}}}{T_{\mathrm{org\_RDO}}} \times 100\% \tag{37}$$

where $T_{\mathrm{org\_RDO}}$ and $T_{\mathrm{pro\_RDO}}$ indicate the total coding time with the conventional and the proposed SSIM-based RDO schemes, respectively.

To verify the efficiency of the proposed perceptual RDO method, extensive experiments are conducted on standard sequences in QCIF and CIF formats. In these experiments, RD performance of the conventional RDO coding strategy and the proposed SSIM motivated perceptual RDO coding strategy is compared. The common coding configurations are set as follows: all available inter and intra modes are enabled, five reference frames, one I frame followed by 99 inter frames, high complexity RDO, and the fixed QPs are set from 28 to 40. The results of the experiments are shown in Tables II and III, and the RD performances are compared in Fig. 9.

For IPP GoP structure, on average 15% rate reduction for fixed SSIM and 16% rate reduction while fixing weighted SSIM are achieved for both QCIF and CIF sequences. When the GoP structure is IBP, the rate reductions are 9% on average for fixed SSIM and 10% on average for fixed weighted SSIM. In general, there are three main reasons behind the improved performance. First, we use SSIM for RDO purposes, which is a better predictor of perceived quality by HVS as compared to ubiquitous MSE. Second, the Lagrange multiplier is calculated adaptively by the accurate RR-SSIM and rate models. Third, we consider the motion between the frames, which is an important information in visual perception of video signals, to further improve the rate distribution among the MBs considering the HVS. The lower gain of IBP coding scheme

TABLE II

PERFORMANCE OF THE PROPOSED ALGORITHMS (COMPARED WITH ORIGINAL RDO TECHNIQUE) FOR QCIF SEQUENCES AT 30 F/S

Sequences		CABAC					CAVLC				
		ΔSSIM	ΔR^a (%)	$\Delta SSIM_\omega$	ΔR^b (%)	ΔPSNR (dB)	ΔSSIM	ΔR^a (%)	$\Delta SSIM_\omega$	ΔR^b (%)	ΔPSNR (dB)
Akiyo	IPP..	0.0116	−17.85	0.0142	−19.83	0.13	0.0123	−19.33	0.0151	−21.09	0.21
	IBP..	0.0075	−5.77	0.0100	−8.93	−0.06	0.0091	−9.64	0.0116	−11.17	0.06
Bridge-close	IPP..	0.0171	−30.65	0.0192	−34.20	−0.02	0.0194	−35.64	0.0228	−41.12	0.01
	IBP..	0.0148	−29.11	0.0168	−32.77	−0.15	0.0150	−30.90	0.0177	−35.98	−0.17
Highway	IPP..	0.0108	−21.00	0.0127	−20.70	−0.26	0.0109	−21.78	0.0144	−23.09	−0.42
	IBP..	0.0043	−7.80	0.0057	−9.40	−0.49	0.0046	−10.91	0.0064	−12.82	−0.46
Grandma	IPP..	0.0188	−23.03	0.0219	−25.38	0.25	0.0192	−22.70	0.0220	−24.47	0.28
	IBP..	0.0158	−19.44	0.0192	−21.74	0.13	0.0164	−19.68	0.0198	−21.59	0.14
Container	IPP..	0.0088	−18.06	0.0088	−17.12	−0.10	0.0091	−17.63	0.0096	−17.01	−0.10
	IBP..	0.0048	−12.30	0.0054	−13.11	−0.47	0.0055	−11.04	0.0058	−10.72	−0.47
Salesman	IPP..	0.0189	−17.72	0.0199	−18.11	0.11	0.0200	−18.14	0.0210	−18.28	0.12
	IBP..	0.0103	−9.44	0.0125	−11.24	−0.21	0.0101	−9.25	0.0118	−10.39	−0.26
News	IPP..	0.0082	−12.76	0.0098	−11.82	−0.15	0.0078	−12.71	0.0096	−12.96	−0.19
	IBP..	0.0052	−7.36	0.0071	−8.56	−0.35	0.0046	−6.50	0.0061	−8.21	−0.38
Carphone	IPP..	0.0035	−6.29	0.0042	−7.21	−0.52	0.0034	−5.59	0.0042	−6.62	−0.45
	IBP..	0.0010	−2.45	0.0015	−3.55	−0.56	0.0010	−2.36	0.0019	−4.42	−0.56
Average	IPP..	0.0122	−18.42	0.0138	−19.30	−0.07	0.0128	−19.19	0.0148	−20.58	−0.07
	IBP..	0.0080	−11.71	0.0098	−13.66	−0.27	0.0082	−12.54	0.0101	−14.41	−0.26

[a]Rate reduction while maintaining SSIM.
[b]Rate reduction while maintaining weighted SSIM.

Fig. 10. Visual quality comparison between the conventional RDO and proposed RDO scheme, where the 40th frame (cropped for visualization) of the *Flower* sequence is shown. (a) Original. (b) H.264/AVC coded with conventional RDO; bit rate: 203.5 kbit/s; SSIM: 0.8710; PSNR: 25.14 dB. (c) H.264/AVC coded with proposed RDO; bit rate: 199.82 kbit/s; SSIM: 0.8805; PSNR: 24.57 dB.

may be explained by two reasons. First, the B frame is usually coded at relatively low bit rate while our proposed scheme achieves superior performance at high bit rate compared to low bit rate, as can be observed from Fig. 9. Second, the parameters estimation scheme proposed in Section V is not as accurate for this GoP structure because the frames of the same coding types are not adjacent to each other.

Rate reduction peaks for sequences with slow motion such as *Bridge*, in which case 35% of the bits can be saved for the same SSIM value of the received video. It is observed that for these sequences with larger Λ, the superior performance is mainly due to the selection of the MB mode with less bits. A similar phenomenon has also been observed in [17] and [18]. Another interesting observation is that the performance gain of the proposed method decreases at very low bit rate, such as the *Bridge* and *Salesman* in Fig. 9. It is due to the fact that at low bit rate a large percentage of MBs have already been coded with the best mode in the conventional RDO scheme, such as SKIP mode. Also, the limitation of the proposed rate model as stated in Section III also brings the limited performance gain at low bit rate. We have also compared the performance in terms of PSNR of the luminance component, which is shown in Tables II and III. Because our scheme is totally adaptive to the video sequences, for some sequences, such as *Akiyo* and *Container*, PSNR increases. However, on average PSNR decreases because our optimization objective is SSIM rather than PSNR.

To show the advantage of our frame-MB joint RDO scheme, the performance comparisons of the frame-level perceptual RDO (FP-RDO) and the frame-MB level perceptual RDO (FMP-RDO) are also listed in Table IV. As can be observed from Table IV, the weighted SSIM increases for sequences with high motion, such as *Flower*. However, the weighted SSIM decreases for constant sequences, such as *Silent*. This performance degradation mainly comes from the inter prediction technique used in video coding. For instance, the

TABLE III

PERFORMANCE OF THE PROPOSED ALGORITHMS (COMPARED WITH ORIGINAL RDO TECHNIQUE) FOR CIF SEQUENCES AT 30 F/S

Sequences		CABAC					CAVLC				
		ΔSSIM	ΔR^a (%)	$\Delta SSIM_\omega$	ΔR^b (%)	ΔPSNR (dB)	ΔSSIM	ΔR^a (%)	$\Delta SSIM_\omega$	ΔR^b (%)	ΔPSNR (dB)
Silent	IPP..	0.0109	−13.98	0.0118	−14.69	−0.18	0.0114	−14.13	0.0123	−14.85	−0.21
	IBP..	0.006	−7.79	0.0077	−9.96	−0.34	0.0063	−7.84	0.0074	−9.10	−0.37
Bus	IPP..	0.0134	−14.85	0.0122	−13.88	−0.70	0.0148	−15.61	0.0136	−14.89	−0.62
	IBP..	0.0083	−9.39	0.0087	−9.51	−0.66	0.0080	−8.63	0.0081	−8.49	−0.73
Mobile	IPP..	0.0047	−8.52	0.0053	−10.50	−0.58	0.0051	−9.52	0.0059	−11.76	−0.63
	IBP..	0.0017	−3.23	0.0026	−5.52	−0.64	0.0009	−1.77	0.0019	−4.35	−0.68
Paris	IPP..	0.0080	−12.07	0.0096	−14.35	−0.38	0.0076	−11.30	0.0090	−13.69	−0.43
	IBP..	0.0036	−5.17	0.0050	−7.36	−0.62	0.0029	−4.02	0.0043	−6.55	−0.36
Flower	IPP..	0.0076	−14.19	0.0068	−11.69	−0.57	0.0070	−13.31	0.0063	−10.86	−0.71
	IBP..	0.0035	−6.92	0.0029	−4.65	−0.47	0.0021	−4.01	0.0014	−1.78	−0.71
Foreman	IPP..	0.0023	−4.80	0.0020	−4.26	−0.75	0.0028	−5.72	0.0027	−5.11	−0.58
	IBP..	0.0008	−1.89	0.0008	−1.97	−0.55	0.0009	−1.66	0.0008	−1.65	−0.70
Tempete	IPP..	0.0072	−10.28	0.0083	−11.70	−0.35	0.0078	−11.27	0.0088	−12.48	−0.36
	IBP..	0.0031	−4.13	0.0040	−5.51	−0.41	0.0029	−4.26	0.0038	−5.56	−0.58
Waterfall	IPP..	0.0207	−15.51	0.0193	−14.22	−0.27	0.0237	−17.20	0.0226	−16.39	−0.22
	IBP..	0.0097	−9.37	0.0099	−9.98	−0.47	0.0092	−8.80	0.0093	−9.35	−0.46
Average	IPP..	0.0094	−11.78	0.0094	−11.91	−0.47	0.0100	−12.26	0.0102	−12.50	−0.47
	IBP..	0.0046	−5.99	0.0052	−6.81	−0.52	0.0042	−5.12	0.0046	−5.85	−0.57

[a]Rate reduction while maintaining SSIM. [b]Rate reduction while maintaining weighted SSIM.

TABLE IV

PERFORMANCE COMPARISON OF THE PROPOSED FP-RDO AND FM-PRDO CODING (ANCHOR: CONVENTIONAL RDO TECHNIQUE)

Sequences		CABAC				CAVLC			
		IPPPP		IBPBP		IPPPP		IBPBP	
		ΔR^a (%)	ΔR^b (%)	ΔR^a (%)	ΔR^b (%)	ΔR^a (%)	ΔR^b (%)	ΔR^a (%)	ΔR^b (%)
Flower (CIF)	FMP-RDO	−14.19	−11.69	−6.92	−4.65	−13.31	−10.86	−4.01	−1.78
	FP-RDO	−14.34	−11.43	−6.73	−4.05	−12.73	−9.75	−2.04	0.38
Waterfall (CIF)	FMP-RDO	−15.51	−14.22	−9.37	−9.98	−17.20	−16.39	−8.80	−9.35
	FP-RDO	−15.45	−14.43	−8.79	−9.47	−16.13	−15.48	−7.98	−8.62
Bus (CIF)	FMP-RDO	−14.85	−13.88	−9.39	−9.51	−15.61	−14.89	−8.63	−8.49
	FP-RDO	−14.71	−13.72	−8.95	−8.84	−16.05	−14.96	−8.72	−8.63
Silent (CIF)	FMP-RDO	−13.98	−14.69	−7.79	−9.96	−14.13	−14.85	−7.84	−9.10
	FP-RDO	−14.62	−15.28	−8.07	−9.79	−15.23	−15.59	−8.53	−9.85
Salesman (QCIF)	FMP-RDO	−17.72	−18.11	−9.44	−11.24	−18.14	−18.28	−9.25	−10.39
	FP-RDO	−17.09	−17.48	−8.44	−10.43	−18.17	−19.06	−8.28	−9.75
Carphone (QCIF)	FMP-RDO	−6.29	−7.21	−2.45	−3.55	−5.59	−6.62	−2.36	−4.42
	FP-RDO	−6.89	−7.31	−2.11	−3.43	−4.40	−5.86	−2.61	−4.85
Container (QCIF)	FMP-RDO	−18.06	−17.12	−12.30	−13.11	−17.63	−17.01	−11.04	−10.72
	FP-RDO	−17.23	−16.21	−12.41	−13.16	−18.20	−17.90	−11.89	−11.71
Bridge (QCIF)	FMP-RDO	−30.65	−34.20	−29.11	−32.77	−35.64	−41.12	−30.90	−35.98
	FP-RDO	−30.93	−34.24	−30.16	−33.88	−33.78	−39.32	−30.40	−35.48

[a]Rate reduction while maintaining of SSIM.
[b]Rate reduction while maintaining weighted SSIM.

TABLE V

SSIM INDICES AND BIT RATES OF TESTING SEQUENCES USED IN THE SUBJECTIVE TEST

Sequences		Conventional RDO		Proposed RDO	
		SSIM	Bit Rate (kbit/s)	SSIM	Bit Rate (kbit/s)
1	*Bus*	0.996	6032.68	0.9955	5807.44
2	*Hall*	0.9899	4976.36	0.99	4745.04
3	*Container*	0.9745	994.04	0.9754	883.72
4	*Tempete*	0.9726	1248.4	0.9707	1044.72
5	*Akiyo*	0.9711	97.81	0.9722	75.68
6	*Silent*	0.9655	457.68	0.9669	423.02
7	*Mobile*	0.9577	728.87	0.9572	703.34
8	*Stefan*	0.8956	179.42	0.8973	174.33

TABLE VI

ENCODING COMPLEXITY OVERHEAD OF THE PROPOSED SCHEME

Sequences	ΔT with CABAC (%)	ΔT with CAVLC (%)
Akiyo (QCIF)	5.21	5.72
News (QCIF)	5.18	5.60
Mobile (QCIF)	5.82	6.14
Silent (CIF)	7.04	7.46
Foreman (CIF)	6.79	7.03
Tempete (CIF)	7.04	7.13
Average	6.18	6.51

MB with higher weight in the current frame may get the prediction pixels from an unimportant MB in the pervious frame, which can cause more quantization errors. Our current work focuses on RDO frame by frame. The interrelationship between frames and the rate control at the GoP level will be studied in the future.

TABLE VII
PERFORMANCE COMPARISON OF USING DIFFERENT PREVIOUS FRAMES FOR PARAMETER ESTIMATION

Sequences			Three Previous Frames		Five Previous Frames		Seven Previous Frames	
			ΔSSIM	ΔR (%)	ΔSSIM	ΔR (%)	ΔSSIM	ΔR (%)
Akiyo (QCIF)	IPP	CABAC	0.0116	−17.85	0.0115	−16.91	0.0116	−18.57
		CAVLC	0.0123	−19.33	0.0120	−17.64	0.0118	−16.80
	IBP	CABAC	0.0075	−5.77	0.0078	−6.83	0.0069	−5.10
		CAVLC	0.0091	−9.64	0.0085	−8.41	0.0090	−9.26
Highway (QCIF)	IPP	CABAC	0.0108	−21.00	0.0103	−20.51	0.0102	−20.33
		CAVLC	0.0109	−21.78	0.0107	−20.41	0.0105	−19.70
	IBP	CABAC	0.0043	−7.80	0.0045	−8.13	0.0045	−8.24
		CAVLC	0.0046	−10.91	0.0048	−11.72	0.0045	−10.10
Mobile (CIF)	IPP	CABAC	0.0047	−8.52	0.0051	−9.22	0.0045	−8.01
		CAVLC	0.0051	−9.52	0.0047	−8.41	0.0053	−10.09
	IBP	CABAC	0.0017	−3.23	0.0015	−2.81	0.0015	−3.03
		CAVLC	0.0009	−1.77	0.0010	−1.89	0.0010	−2.01
Flower (CIF)	IPP	CABAC	0.0076	−14.19	0.0074	−13.87	0.0075	−13.90
		CAVLC	0.0070	−13.31	0.0068	−12.88	0.0072	−14.60
	IBP	CABAC	0.0035	−6.92	0.0032	−5.74	0.0033	−6.04
		CAVLC	0.0021	−4.01	0.0022	−4.58	0.0023	−4.60

Fig. 10 shows the original frame, H.264/AVC coded frame with the conventional RDO and H.264/AVC coded frame with the proposed RDO method. Note that the bit rates for the two coding methods are almost the same. However, since our proposed RDO scheme is based on SSIM index optimization, higher SSIM and lower PSNR are achieved. Furthermore, the quality of the reconstructed frame has been obviously improved by the proposed scheme. It can be observed that more information and details have been preserved, such as the branches on the roof. The visual quality improvement is due to the fact that we can select the best mode from perceptual point of view, resulting in more bits allocated to the areas which are more sensitive to our visual systems.

To further validate our scheme, we carried out a subjective quality evaluation test based on a two-alternative forced choice (2AFC) process that is widely used in psychophysical studies, where in each trial, a subject is shown a pair of video sequences and is asked (forced) to choose the one he/she thinks to have better quality. In our experiment, we selected eight pairs of sequences of CIF format that were coded by the conventional and the proposed RDO schemes to achieve the same SSIM levels (where the proposed scheme uses much lower bit rates). Table V lists all the test sequences as well as their SSIM values and bit rates. In the 2AFC test, each pair is repeated six times with random order. As a result, we obtained 48 2AFC results for each subject. Ten subjects participated in this experiment.

The subjective test results are reported in Fig. 11, which shows the percentage ϖ by which the subjects are in favor of the conventional RDO against the proposed RDO schemes. As can be observed in the figure, the overall percentage (the rightmost bar in the figure) is very close to 50% (52.5%), meaning that there is no significant perceptual difference of visual quality between the video sequences coded by the two schemes (though the proposed scheme uses much lower bit rates). In the figure, we also plot the variations of the percentage over the ten subjects and over the eight sequences, together with the error bars (±one standard deviation between the measurements). It turns out that for almost all cases the value of ϖ is close to 50% and all error bars cross the 50% line, showing the robustness of the measurement. These results provide useful evidence that the proposed method achieves the same level of quality with lower bit rates.

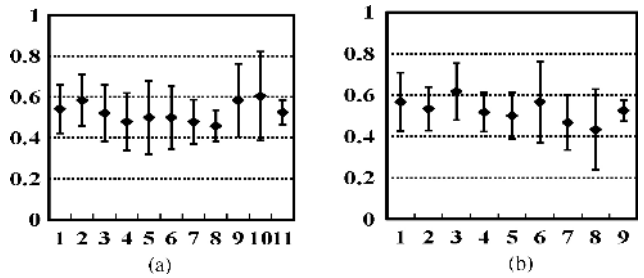

Fig. 11. Error-bar plot for the subjective test. (a) Error-bar plot with in units of ϖ and standard deviation for each subject (1–10: subject number; 11: average). (b) Error-bar plot with in units of ϖ and standard deviation for each test sequence (1–8: sequence number; 9: average).

Table VI reports the computation overhead of the proposed scheme with both CABAC and CAVLC entropy coding methods, where ΔT is calculated according to (37). The coding time is obtained by encoding 100 frames of IPPP GoP structure with Intel 2.83 GHz Core processor and 4 GB random access memory. On average the computation overhead is 6.3% for our scheme. As already indicated in [34] that the computation of SSIM index in the mode selection process causes about 5% overhead. Therefore, in our method the computation overhead is mainly due to the calculation of the SSIM index for each mode. We also observe that the overhead is stable for different video sequences.

Table VII lists the experimental results of using three, five, and seven previous frames to estimate the parameters in Section V, respectively. Both IPP and IBP GoP structures are tested and both CAVLC and CABAC entropy coding algorithms are employed. As indicated in Table VII, the final performance is not sensitive to the number of pervious frames used in the estimation. This can be explained by the stable properties of video sequences during a short period of time, as shown in Fig. 7. This suggests us to use three previous frames, as they are enough to capture the properties of the

TABLE VIII
PERFORMANCE COMPARISON WITH THE STATE OF THE ART RDO CODING ALGORITHMS FOR IPP GOP STRUCTURE
(ANCHOR: CONVENTIONAL RDO TECHNIQUE)

Sequences		Proposed		Huang *et al.*'s		Yang *et al.*'s		CALM		RDOQ	
		ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)
Akiyo (CIF)	QP_1	**0.0026**	**−26.11**	0.0020	−19.40	0.0004	−4.28	0	0.46	0.0001	−1.08
	QP_2	**0.0078**	**−28.06**	0.0056	−15.78	0.0024	−13.60	0	0.25	0	0.11
Bus(CIF)	QP_1	**0.0016**	**−7.77**	0.0011	−5.95	0.0015	−7.12	0	−0.04	0.0006	−2.20
	QP_2	**0.0099**	**−14.87**	0.0086	−13.25	0.0038	−6.03	0	−0.07	0.0007	−1.36
Coastguard (CIF)	QP_1	**0.0013**	**−4.77**	0.0004	−2.28	0.0005	−2.16	0	−0.06	0.0006	−1.54
	QP_2	**0.0076**	**−8.91**	0.0038	−5.04	0.0036	−3.97	−0.0002	0.3	0.0005	−0.80
Silent (CIF)	QP_1	**0.0026**	**−9.64**	0.0013	−5.28	−0.0002	0.04	0	−0.14	0.0012	−4.15
	QP_2	**0.0091**	**−12.43**	0.0046	−6.83	−0.0008	0.58	0	−0.05	0	−0.08
Hall (CIF)	QP_1	0.0034	−25.89	**0.0035**	**−26.41**	0.0013	−10.01	0	0.27	0.0005	−3.78
	QP_2	**0.0062**	**−25.46**	0.0059	−22.84	0.0003	−1.51	0	0.11	0.0002	−2.80
Mother_Dau (CIF)	QP_1	**0.0008**	**−6.43**	0.0004	−2.76	0	0.56	0	0.03	0.0003	−1.49
	QP_2	**0.0049**	**−8.94**	0.0022	−4.69	0.0015	−2.84	0	−0.3	0	−0.19
Spincalendar (720P)	QP_1	0.0028	−11.89	**0.0030**	**−12.78**	0.0021	−8.29	0	0.02	0.0022	−9.13
	QP_2	**0.0042**	**−15.57**	0.0040	−12.81	0.0006	−2.16	0	−0.43	0.0011	−2.50
Night (720P)	QP_1	**0.0019**	**−6.65**	0.0011	−3.45	−0.0002	0.85	0	0.14	0.0009	−4.70
	QP_2	**0.0062**	**−16.02**	0.0029	−11.38	0.0002	−0.96	0	0.09	0.0010	−2.05
Average	QP_1	**0.0021**	**−12.39**	0.0016	−9.79	0.0007	−3.8	0	0.09	0.0008	−3.51
	QP_2	**0.0070**	**−16.28**	0.0047	−11.58	0.0015	−3.81	0	−0.01	0.0004	−1.21

TABLE IX
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART RDO CODING ALGORITHMS FOR IBP GOP STRUCTURE
(ANCHOR: CONVENTIONAL RDO TECHNIQUE)

Sequences		Proposed		Huang *et al.*'s		Yang *et al.*'s		CALM		RDOQ	
		ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)	ΔSSIM	Δ*R* (%)
Akiyo (CIF)	QP_1	**0.0014**	**−17.39**	0.0007	−9.72	0.0003	−5.01	0	−0.49	0	−0.19
	QP_2	**0.0030**	**−8.56**	0.0022	−6.41	0.0015	−4.60	0	0.32	−0.0005	2.01
Bus (CIF)	QP_1	0.0004	−2.04	**0.0006**	**−3.95**	0.0003	−1.12	0	0.15	0.0002	−1.20
	QP_2	**0.0048**	**−7.58**	0.0036	−5.25	0.0038	−6.05	0	0.12	0.0021	−3.36
Coastguard (CIF)	QP_1	**0.0007**	**−3.41**	0.0003	−1.96	0.0005	−2.59	0	0.46	0.0006	−2.86
	QP_2	**0.0027**	**−3.31**	0.0011	−2.04	0.0009	−1.67	0	0.25	0.0014	−1.89
Silent (CIF)	QP_1	**0.0014**	**−4.64**	0.0013	−4.28	0	−0.03	0	0.06	0.0006	−2.75
	QP_2	**0.0050**	**−6.76**	0.0036	−4.60	0.0018	−2.11	0	0	0.0012	−1.73
Hall (CIF)	QP_1	**0.0009**	**−7.60**	0.0003	−2.41	0.0003	−2.72	0	0.21	0.0003	−2.09
	QP_2	**0.0031**	**−19.42**	0.0007	−4.87	0.0005	−3.27	0	0.43	0.0003	−2.51
Mother_Dau (CIF)	QP_1	**0.0009**	**−7.43**	0.0006	−5.80	0.0001	−1.23	0	−0.59	0.0003	−2.28
	QP_2	**0.0041**	**−5.94**	0.0007	−1.69	0.0015	−2.91	0.0001	−0.16	0.0003	−0.51
Spincalendar (720P)	QP_1	0.0006	−5.79	**0.0010**	**−7.18**	0.0004	−4.10	0	0.15	0.0005	−5.60
	QP_2	**0.0037**	**−4.59**	0.0021	−3.81	0.0009	−1.16	0	−0.53	0.0013	−2.57
Night (720P)	QP_1	**0.0013**	**−4.94**	0.0010	−3.51	0.0002	−0.91	0	-0.15	0.0010	−3.61
	QP_2	**0.0019**	**−5.73**	0.0006	−2.11	0.0004	−1.96	0	-0.23	0.0016	−3.33
Average	QP_1	**0.0010**	**−6.66**	0.0007	−4.85	0.0003	−2.21	0	-0.03	0.0004	−2.57
	QP_2	**0.0035**	**−7.74**	0.0018	−3.85	0.0014	−2.97	0	0.03	0.0010	−1.74

video sequences and to obtain an accurate estimation of the required parameters.

C. Comparisons with State-of-the-Art RDO Algorithms

In this experiment, the proposed scheme is compared with state-of-the-art RDO algorithms, including Huang *et al.*'s SSIM-based RDO algorithm [34], Yang *et al.*'s SSIM-based RDO algorithm [32], the CALM selection scheme [21], and the RD-optimized quantization (RDOQ) scheme [4]. For this experiment, both IPP and IBP GoP structures are employed and CAVLC entropy coding method is used. We use two different sets of QP values in the experiments: $QP_1 = \{16, 20, 24, 28\}$ and $QP_2 = \{24, 28, 32, 36\}$, where QP_1 indicates a high bit rate coding configuration. For each scheme, the improvement of the SSIM index as well as the rate reduction compared to the conventional RDO coding schemes are tabulated in Tables VIII and IX.

From Tables VIII and IX, it can be observed that over a wide range of bit rates, for most of the cases our scheme achieves better performance than state-of-the-art SSIM-based RDO methods. Specifically, when compared to Huang *et al.*'s method, on average the proposed scheme achieves better

rate reduction of 12.39% versus 9.79% for QP_1 and 16.28% versus 11.58% for QP_2 while maintaining the same SSIM values for IPP GoP structure. For IBP GoP structure, the performance gain is 6.66% versus 4.85% for QP_1 and 7.74% versus 3.85% for QP_2. We believe that there are three main factors that are responsible for the performance improvement. First, the proposed scheme uses more accurate statistical SSIM and rate models which are derived from the inherent properties of SSIM index and the video signals. Second, in this scheme, the Lagrange multiplier is derived adaptively for each frame. Finally, in the mode selection process, the surrounding pixels are employed to accurately obtain the SSIM index for each mode. The performances of the MSE-based RDO coding schemes are also given in Tables VIII and IX. Since their optimization objective is MSE rather than SSIM, there is no significant change of SSIM values in these schemes.

VII. Conclusion

We proposed an SSIM-motivated perceptual RDO scheme for H.264/AVC video coding with the aim of selecting the best coding mode and achieving the best rate–SSIM performance. The novelty of our approaches lies in the adaptive Lagrange multiplier selection methods at both frame and MB levels, where we incorporated a new RR-SSIM estimation algorithm and information theoretical methods that take motion information and perceptual uncertainty of visual speed perception into account. The superior performance of the proposed scheme was demonstrated using the reference software JM, which offered significant rate reduction, while keeping the same level of SSIM values. Visual quality improvement was also observed when compared with conventional RDO scheme.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments that significantly helped us in improving the presentation of this paper. The authors would also like to thank the authors of [34] for providing critical details of the implementation of their algorithms.

References

[1] 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.

[2] B. Girod, "Whats wrong with mean-squared error," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993, pp. 207–220.

[3] X. Zhao, L. Zhang, S. Ma, and W. Gao, "Rate-distortion optimized transform for intra-frame coding," in *Proc. IEEE Int. Conf. Acou., Speech Signal Process.*, Mar. 2010, pp. 1414–1417.

[4] M. Karczewicz, Y. Ye, and I. Chong, "Rate distortion optimized quantization," document VCEG-AH21, ITU-T Q.6/SG16 VCEG, Antalya, Turkey, Jan. 2008.

[5] E. H. Yang and X. Yu, "Rate distortion optimization for H.264 inter-frame video coding: A general framework and algorithms," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1774–1784, Jul. 2007.

[6] E. H. Yang and X. Yu, "Soft decision quantization for H.264 with main profile compatibility," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 122–127, Jan. 2009.

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

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

[9] D. Chandler and 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.

[10] Z. Wang and A. 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.

[11] *Advanced Video Coding*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 Part 10), 2010.

[12] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullian, "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.

[13] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. Int. Conf. Image Process.*, 2001, pp. 542–545.

[14] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Optic. Soc. Am. A*, vol. 24, pp. B61–B69, Dec. 2007.

[15] L. Chen and I. Garbacea, "Adaptive λ estimation in Lagrangian rate-distortion optimization for video coding," *Proc. SPIE*, vol. 6077, pp. 1–8, Jan. 2006.

[16] Z. He and S. Mitra, "Optimum bit allocation and accurate rate control for video coding via rho-domain source modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 10, pp. 840–849, Oct. 2002.

[17] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[18] M. Jiang and N. Ling, "On Lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 663–669, May 2006.

[19] M. Wang and B. Yan, "Lagrangian multiplier based joint three-layer rate control for H.264/AVC," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 679–682, Aug. 2009.

[20] J. Zhang, X. Yi, N. Ling, and W. Shang, "Context adaptive Lagrange multiplier (CALM) for rate-distortion optimal motion estimation in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 820–828, Jun. 2010.

[21] J. Zhang, X. Yi, N. Ling, and W. Shang, "Context adaptive lagrange multiplier (CALM) for motion estimation in JM-improvement," document JVT-T046, Joint Video Team (JVT) of ISO/IEC MPEG ITU-T VCEG, Jul. 2006.

[22] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-compensated residue pre-processing in video coding based on just-noticeable distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, Jun. 2005.

[23] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Process.: Image Commun.*, vol. 22, pp. 662–680, Aug. 2005.

[24] 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.

[25] C. Sun, H.-J. Wang, and H. Li, "Macroblock-level rate-distortion optimization with perceptual adjustment for video coding," in *Proc. IEEE DCC*, Mar. 2008, p. 546.

[26] F. Pan, Y. Sun, Z. Lu, and A. Kassim, "Complexity-based rate distortion optimization with perceptual tuning for scalable video coding," in *Proc. Int. Conf. Image Process.*, Sep. 2005, pp. 37–40.

[27] A. Brooks, X. Zhao, and T. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 121–132, Aug. 2008.

[28] B. Aswathappa and K. R. Rao, "Rate-distortion optimization using structural information in H.264 strictly intra-frame encoder," in *Proc. South Eastern Symp. Syst. Theory*, 2010, pp. 367–370.

[29] Z. Mai, C. Yang, L. Po, and S. Xie, "A new rate-distortion optimization using structural information in H.264 I-frame encoder," in *Proc. ACIVS*, 2005, pp. 435–441.

[30] Z. Mai, C. Yang, and S. 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.*, Oct. 2005, pp. 2673–2678.

[31] Z. Mai, C. Yang, K. Kuang, and L. Po, "A novel motion estimation method based on structural similarity for H.264 inter prediction," in *Proc. IEEE Int. Conf. Acou., Speech, Signal Process.*, vol. 2. May 2006, pp. 913–916.
[32] C. Yang, R. Leung, L. Po, and Z. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 4. Jun. 2009, pp. 291–295.
[33] C. Yang, H. Wang, and L. Po, "Improved inter prediction based on structural similarity in H.264," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, vol. 2. Nov. 2007, pp. 340–343.
[34] Y. H. Huang, T. S. Ou, P. Y. Su, and 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.
[35] H. Chen, Y. Huang, P. Su, and T. Ou, "Improving video coding quality by perceptual rate-distortion optimization," in *Proc. IEEE Int. Conf. Multimedia Exp.*, Jul. 2010, pp. 1287–1292.
[36] P. Su, Y. Huang, T. Ou, and H. Chen, "Predictive Lagrange multiplier selection for perceptual-based rate-distortion optimization," in *Proc. 5th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, Jan. 2010.
[37] Y. Huang, T. Ou, and H. Chen, "Perceptual-based coding mode decision," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 393–396.
[38] T. Ou, Y. Huang, and H. Chen, "A perceptual-based approach to bit allocation for H.264 encoder," *Proc. SPIE: Vis. Commun. Image Process.*, vol. 7744, pp. 1–10, Jul. 2010.
[39] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, pp. 121–132, Feb. 2004.
[40] J. Sun, W. Gao, D. Zhao, and Q. Huang, "Statistical model, analysis and approximation of rate-distortion function in MPEG-4 FGS videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 4, pp. 535–539, Apr. 2006.
[41] Y. Altunbasak and N. Kamaci, "An analysis of the DCT coeffcient distribution with the H.264 video coder," in *Proc. IEEE ICASSP*, vol. 3. May 2004, pp. 177–180.
[42] A. Rehman and Z. Wang, "Reduced-reference SSIM estimation," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 289–292.
[43] S. Channappayya, A. C. Bovik, and J. R. W. Heathh, "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008.
[44] D. Kwon, M. Shen, and C. Kuo, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 5, pp. 517–529, May 2007.
[45] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nat. Neurosci.*, vol. 9, no. 4, pp. 578–585, Apr. 2006.
[46] T. Vlachos, "Simple method for estimation of global motion paramenters using spares translational motion vector fields," *Electron. Lett.*, vol. 34, no. 1, pp. 90–91, Jan. 1998.
[47] S. Wang, S. Ma, and W. Gao, "SSIM based perceptual distortion rate optimization coding," *Proc. SPIE: Vis. Commun. Image Process.*, vol. 7744, pp. 1–10, Jul. 2010.
[48] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. 14, no. 5, pp. 676–683, Oct. 1968.
[49] *Joint Video Team (JVT) Reference Software* [Online]. Available: http://iphome.hhi.de/suehring/tml/download/old-jm
[50] G. Bjøntegaard, "Calculation of average PSNR difference between RD curves," document ITU-T Q.6/SG16, 13th VCEG-M33 Meeting, Austin, TX, Apr. 2001.

Shiqi Wang received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008. He is currently pursuing the Ph.D. degree in computer science from Peking University, Beijing, China.

From 2010 to 2011, he was a Vising Student with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From April 2011 to August 2011, he was with Microsoft Research Asia, Beijing, as an Intern. His current research interests include video compression, image and video quality assessment, and multiview video coding.

Abdul Rehman (S'10) received the B.S. degree in electrical engineering from the National University of Sciences and Technology, Rawalpindi, Pakistan, in 2007, and the M.S. degree in communications engineering from Technical University Munich, Munich, Germany, in 2009. Currently, he is pursuing the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada.

Since 2009, he has been a Research Assistant with the Department of Electrical and Computer Engineering, University of Waterloo. In 2011, he was with the Video Compression Research Group at Research in Motion, Waterloo. From 2007 to 2009, he was a Research and Teaching Assistant with the Department of Electrical Engineering and Information Technology, Technical University Munich. His current research interests include image and video processing, coding, communication and quality assessment, machine learning, and compressed sensing.

Zhou Wang (S'97–A'01–M'02) received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin, Austin, in 2001.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His current research interests include image processing, coding, and quality assessment, computational vision and pattern analysis, multimedia communications, and biomedical signal processing. He has more than 90 publications in these fields with over 7000 citations (Google Scholar). He is an author of *Modern Image Quality Assessment* (Morgan & Claypool, 2006).

Dr. Wang received the 2009 IEEE SIGNAL PROCESSING Best Paper Award, the ICIP 2008 IBM Best Student Paper Award (as a Senior Author), and the 2009 Ontario Early Researcher Award. He has served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2009 to present, the IEEE SIGNAL PROCESSING LETTERS from 2006 to 2010, and *Pattern Recognition* from 2006 to present. He was a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING from 2007 to 2009, the *EURASIP Journal of Image and Video Processing* from 2009 to 2010, and the *Signal, Image, and Video Processing* from 2011 to present.

Siwei Ma (S'03) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

From 2005 to 2007, he was a Post-Doctoral Researcher with the University of Southern California, Los Angeles. He then joined the Institute of Digital Media, Department of Electrical Engineering and Computer Science, Peking University, Beijing, where he is currently an Associate Professor. He has published over 70 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission.

Wen Gao (M'92–SM'05–F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

Currently, he is a Professor of computer science at Peking University, Beijing, China. Before joining Peking University, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He has published extensively including five books and over 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Dr. Gao has served or serves on the editorial boards of several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, the *EURASIP Journal of Image Communications*, and the *Journal of Visual Communication and Image Representation*. He has chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also has served on the advisory and technical committees of numerous professional organizations.