Perceptual Video Coding Based on SSIM-Inspired Divisive Normalization

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

***Abstract*— We propose a perceptual video coding framework based on the divisive normalization scheme, which is found to be an effective approach to model the perceptual sensitivity of biological vision, but has not been fully exploited in the context of video coding. At the macroblock (MB) level, we derive the normalization factors based on the structural similarity (SSIM) index as an attempt to transform the discrete cosine transform domain frame residuals to a perceptually uniform space. We further develop an MB level perceptual mode selection scheme and a frame level global quantization matrix optimization method. Extensive simulations and subjective tests verify that, compared with the H.264/AVC video coding standard, the proposed method can achieve significant gain in terms of rate-SSIM performance and provide better visual quality.**

***Index Terms*— Divisive normalization, H.264/AVC coding, perceptual video coding, rate distortion optimization, structural similarity (SSIM) index.**

I. Introduction

OVER the past decade, there has been an exponential increase in the demand for digital video services such as high-definition television, web-based television, video conferencing and video-on-demand. To facilitate these services, it demands to significantly reduce the storage space and bandwidth of visual content production, storage and delivery. Therefore, there has been a strong desire of powerful video coding techniques beyond H.264/AVC.

The main objective of video coding is to minimize the perceptual distortion D of the reconstructed video with the number of used bits R subjected to a constraint R_c. This can be expressed as

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

Manuscript received May 11, 2012; revised September 11, 2012, accepted November 9, 2012. Date of publication December 3, 2012; date of current version February 6, 2013. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, the Ontario Early Researcher Award Program, the Major State Basic Research Development Program of China 973 Program under Grant 2009CB320903, the National Science Foundation of China under Grant 60833013 and Grant 61121002, the National High-Tech R&D Program of China 863 Program under Grant SS2012AA010805, and the National Key Technology R&D Program of China under Grant 2011BAH08B01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James E. Fowler.

S. Wang, 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).

A. Rehman and Z. Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

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/TIP.2012.2231090

Central to such an optimization problem is the way in which the distortion D is defined because the quality of video can only be as good as it is optimized for. Since the ultimate receiver of video is the Human Visual System (HVS), the correct optimization goal should be perceptual quality. However, existing video coding techniques typically use the sum of absolute difference (SAD) or sum of square difference (SSD) as the model for distortion, which have been widely criticized in the literature for the lack of correspondence with perceptual quality [1]–[3]. For many years, there have been numerous efforts in developing subjective-equivalent quality models in an attempt to generate quality scores close to the opinions of human viewers. The more accurate the model is, the more distortion can be allowed without generating perceivable artifact, and the better compression can be achieved.

It is well known that the distortion introduced by quantization in lossy coding is content-dependent due to visual masking effects. By exploiting these effects, it is possible to design video coding algorithms which are able to reduce the coding bitrate for a given target perceptual quality. Many perceptual rate allocation techniques are developed based on human visual sensitivity models. The basic idea of these techniques is to allocate fewer bits to the areas or image components that can tolerate more distortions. In [4], [5], the authors exploited non-uniform spatial-temporal sensitivity characteristics and developed visual sensitivity models which are based on the visual cues such as motion and textural structures. In [6], motion attention, position, and texture structure models were used in the rate distortion optimization (RDO) process to adapt the Lagrange multiplier based on the content of each MB. In [7], a content complexity based RDO scheme was proposed for scalable video coding that also considers object-based features such as human subject and skin color to adjust the Lagrange multiplier.

Since the distortion in video coding mainly originates from quantization, many recent methods attempt to incorporate the properties of the HVS into the quantization process [8]–[13]. Because HVS has different sensitivities to different frequencies, the concept of frequency weighting has been incorporated in the quantization process in many picture coding standards from JPEG to H.264/AVC high profile [9]–[12]. In [8], [14], foveated vision models were employed for optimizing the quantization parameter and Lagrange multiplier. However, these methods are based on near threshold perceptual models, but practical video coding typically

works in a suprathreshold range [15]–[17], where the perceptual quality behavior is poorly predicted from the threshold level.

The structural similarity (SSIM) index [18] has become a popular image quality measure in recent years in various image/video processing areas due to its good compromise between quality evaluation accuracy and computation efficiency. For example, it has been incorporated into motion estimation, mode selection and rate control schemes [19]–[30]. For intra frame coding, SSIM-based RDO schemes were proposed in [19]–[21]. In [22]–[24], the authors developed SSIM-based RDO schemes for inter frame prediction and mode selection. One major advantage of utilizing SSIM index in RDO is that, unlike MSE, the SSIM index is totally adaptive according to the reference signal [18] and therefore the RDO will be automatically adapted to the properties of the video content. However, in these RDO schemes [19]–[24], the properties of video frames are not directly accounted for in determining the Lagrange multiplier. To address this issue, content-adaptive Lagrange multiplier selection schemes were proposed in [25]–[28]. These algorithms employed an adaptive rate-SSIM curve to describe the relationship between SSIM and rate to approximate the R-D characteristics. In [31], adaptive SSIM and rate models are established to develop an SSIM based RDO scheme, where the SSIM model is derived from a reduced-reference image quality assessment algorithm.

In this work, we aim to transform the optimization process in (1) into a perceptually uniform domain by incorporating the divisive normalization framework. It has already been shown that the main difference between SSIM and MSE lies in the locally adaptive divisive normalization process [32]. In general, divisive normalization transform is recognized as a perceptually and statistically motivated non-linear image representation [33], [34]. It is shown to be a useful framework that accounts for the masking effect in the HVS, which refers to the reduction of the visibility of an image component in the presence of neighboring components [35], [36]. It has also been found to be powerful in modeling the neuronal responses in the human perceptual systems [37]–[39]. Divisive normalization has been successfully applied in image quality assessment [40], [41], image coding [42], video coding [43] and image denoising [34], [44].

The main contributions of our work are as follows:

1) We propose a divisive normalization scheme to transform the discrete cosine transform (DCT) domain residuals which are obtained after prediction to a perceptually uniform space based on a DCT domain SSIM index.
2) Following the divisive normalization scheme, we define a new distortion model and propose a novel perceptual RDO scheme for mode selection.
3) In the divisive normalized domain, we propose a frame-level quantization matrix selection approach so that the normalized coefficients of different frequencies share the same R-D relationship.

II. SSIM-Inspired Divisive Normalization

Block motion compensated inter-prediction technique plays an important role in existing hybrid video codecs. In this work, we follow this framework, where previously coded frames are used to predict the current frame and only residuals after prediction are coded.

A. Divisive Normalization Scheme

Assume $C(k)$ to be the k^{th} DCT transform coefficient of a residual block, then the normalized coefficient is computed as $C(k)' = C(k)/f(k)$, where $f(k)$ is a positive normalization factor for the k^{th} subband that will be discussed later.

The quantization process of the normalized residuals for a given predefined quantization step Q_s can be formulated as

$$\begin{aligned} Q(k) &= \text{sign}\{C(k)'\}\text{round}\left\{\frac{|C(k)'|}{Q_s} + p\right\} \\ &= \text{sign}\{C(k)\}\text{round}\left\{\frac{|C(k)|}{Q_s \cdot f(k)} + p\right\} \end{aligned} \tag{2}$$

where p is the rounding offset in the quantization.

At the decoder, the de-quantization and reconstruction of $C(k)$ is performed as

$$\begin{aligned} R(k) &= R(k)' \cdot f(k) = Q(k) \cdot Q_s \cdot f(k) \\ &= \text{sign}\{C(k)\}\text{round}\left\{\frac{|C(k)|}{Q_s \cdot f(k)} + p\right\} \cdot Q_s \cdot f(k). \end{aligned} \tag{3}$$

The purpose of the divisive normalization process is to convert the transform residuals into a perceptually uniform space. Thus the factor $f(k)$ determines the perceptual importance of each of the corresponding transform coefficient. The proposed divisive normalization scheme can be interpreted in two ways. An adaptive normalization factor is applied, followed by quantization with a predefined fixed step Q_s. Alternatively, an adaptive quantization matrix is defined for each MB and thus each coefficient is quantized with a different quantization step.

In the context of computational neuroscience as well as still image processing and coding, several different approaches have been used to derive the normalization factor, which may be defined as the sum of the squared neighboring coefficients plus a constant [42], or derived from a local statistical image model [45]. In this work, our objective is to optimize the SSIM index, therefore, we employ a model based on the DCT domain SSIM index.

The DCT domain SSIM index was first presented in [46]:

$$\begin{aligned} \text{SSIM}(\mathbf{x}, \mathbf{y}) = &\left(1 - \frac{(X(0) - Y(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right) \\ &\times \left(1 - \frac{\frac{\sum_{k=1}^{N-1}(X(k) - Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1} + C_2}\right) \end{aligned} \tag{4}$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients of the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. C_1 and C_2 are used to avoid instability when the means and variances are close to zero and N denotes the block size. The DCT domain SSIM index is composed of the product of two terms, which are the

normalized squared errors of DC and AC coefficients, respectively. Moreover, the normalization is conceptually consistent with the light adaptation (also called luminance masking) and contrast masking effect of the HVS [47]–[49]. Eq. (4) can be re-written as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \left(1 - \left(\frac{X(0)}{\sqrt{\eta_{dc}}} - \frac{Y(0)}{\sqrt{\eta_{dc}}}\right)^2\right) \times \left(1 - \frac{1}{N-1}\sum_{k=1}^{N-1}\left(\frac{X(k)}{\sqrt{\eta_{ac}}} - \frac{Y(k)}{\sqrt{\eta_{ac}}}\right)^2\right) \quad (5)$$

where

$$\eta_{dc} = X(0)^2 + Y(0)^2 + N \cdot C_1 \quad (6)$$

$$\eta_{ac} = \frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1} + C_2. \quad (7)$$

Eq. (5) suggests that the DCT domain SSIM index can be computed from normalized MSE of DC and AC coefficients. This inspires us to use SSIM-based divisive normalization for perceptual video coding.

In the video coding scenario, let $P(k)$ be the prediction signal of the k^{th} subband in DCT domain, then the SSIM index can be rewritten as in (8)

$$\begin{aligned}\mathrm{SSIM}(\mathbf{x},\mathbf{y}) &= \left\{1 - \frac{((C(0)+P(0)) - (R(0)+P(0)))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right\} \times \left\{1 - \frac{\frac{\sum_{k=1}^{N-1}((C(k)+P(k)) - (R(k)+P(k)))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2}\right\} \\ &= \left\{1 - \frac{(C(0)-R(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right\} \times \left\{1 - \frac{\frac{\sum_{k=1}^{N-1}(C(k)-R(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2}\right\}. \quad (8)\end{aligned}$$

$$\begin{aligned}\mathrm{SSIM}(\mathbf{x},\mathbf{y}) &= \left\{1 - \frac{(C(0)' \cdot f_{dc} - R(0)' \cdot f_{dc})^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right\} \times \left\{1 - \frac{\frac{\sum_{k=1}^{N-1}(C(k)' \cdot f_{ac} - R(k)' \cdot f_{ac})^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2}\right\} \\ &\approx \left\{1 - \frac{(C(0)' - R(0)')^2}{\mathbb{E}(\sqrt{X(0)^2 + Y(0)^2 + N \cdot C_1})^2}\right\} \times \left\{1 - \frac{\frac{\sum_{k=1}^{N-1}(C(k)' - R(k)')^2}{N-1}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2})^2}\right\}. \quad (9)\end{aligned}$$

Since the local statistics do not change significantly within each MB, we divide each MB into l sub-MBs for DCT transform and $X_i(k)$ denotes the k^{th} DCT coefficient in the i^{th} sub-MB. As the SSIM index differentiates between the DC and AC coefficients, we use separate normalization factors for AC and DC coefficients, which are defined as

$$f_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2 + Y_i(0)^2 + N \cdot C_1}}{\mathbb{E}(\sqrt{X(0)^2 + Y(0)^2 + N \cdot C_1})} \quad (10)$$

$$f_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2+Y_i(k)^2)}{N-1} + C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2})} \quad (11)$$

where $\mathbb{E}(\cdot)$ denotes the mathematical expectation operator. The expectations are over the whole frame, and thus do not affect the relative normalization factors across space within the same frame.

As a result of the use of f_{dc} and f_{ac}, the normalized DCT coefficients for residuals can be expressed as

$$C(k)' = \begin{cases}\frac{C(0)}{f_{dc}}, & k = 0 \\ \frac{C(k)}{f_{ac}}, & \text{otherwise}\end{cases} \quad (12)$$

$$R(k)' = \begin{cases}\frac{R(0)}{f_{dc}}, & k = 0 \\ \frac{R(k)}{f_{ac}}, & \text{otherwise.}\end{cases} \quad (13)$$

Therefore, the SSIM index in the divisive normalization framework can be expressed as in (9), which implies that in the divisive normalization space, the SSIM index is dependent on the difference of the normalized signals but not adaptive to the local normalized signals themselves and therefore all the MBs can be treated as perceptually identical. Since the clearly visible distortion regions will be perceptually more apparent [50], transforming all the coefficients into the perceptually uniform domain is also a convenient approach to improve the perceptual quality according to the philosophy behind distortion-based pooling scheme [51].

The divisive normalization factor is spatially adaptive and dependent on the content of the MB and determines the relative perceptual importance of each MB. The MBs which are less important are quantized more coarsely as compared to the more important MBs. The expected values of DC and AC energy are used as the reference point to determine the importance of each MB. The MBs with higher energy than the mean value have effectively larger quantization step and vice versa. By doing so, we are borrowing bits from the regions which are perceptually less important and using them for the regions with more perceptual relevance, as far as SSIM is concerned, so that all the regions in the frame conceptually have uniform perceptual distortion. It is important to note that the reference point, mean AC and DC energies, is highly dependent on the content of the video frame. The frames with significant texture regions are likely to get more perceptual improvement because the texture regions are the main beneficiaries of the spatially adaptive normalization process.

The calculation of divisive normalization factors for DC and AC coefficients are demonstrated in Fig. 1, where darker MBs indicate smaller normalization factors. As the flower textures can mask more distortions, we assign larger normalization factors to the AC coefficients in these regions. However, since the luminance values in these regions are relatively lower, we assign smaller normalization factors to the DC coefficients.

(a) (b) (c)

Fig. 1. Visualization of spatially adaptive divisive normalization factors for *Flower*@CIF. (a) Original frame. (b) Normalization factors for DC coefficients for each MB. (c) Normalization factors for AC coefficients for each MB.

These are conceptually consistent with the light adaptation and contrast masking effects of the HVS.

B. Perceptual Rate Distortion Optimization for Mode Selection

The RDO process in video coding can be expressed by minimizing the perceived distortion D with the number of used bits R subject to a constraint R_c. This can be converted to an unconstrained optimization problem by

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R \tag{14}$$

where J is called the Rate Distortion (RD) cost and λ is known as the Lagrange multiplier that controls the trade-off between R and D.

Here we replace the conventional SAD and SSD with a new distortion model that is consistent with the residual normalization process. As illustrated in Fig. 2, for each MB, the distortion model is defined as the SSD between the normalized DCT coefficients, which is expressed as

$$\begin{aligned} D &= \sum_{i=1}^{l}\sum_{k=0}^{N-1}(C_i(k)' - R_i(k)')^2 \\ &= \sum_{i=1}^{l}\frac{(X_i(0)-Y_i(0))^2}{f_{dc}^2} + \frac{\sum_{k=1}^{N-1}(X_i(k)-Y_i(k))^2}{f_{ac}^2}. \end{aligned} \tag{15}$$

Based on (14), the RDO problem is given by

$$\begin{aligned} \min\{J\} \ \text{where} \ \ J = &\sum_{i=1}^{l}\sum_{k=0}^{N-1}(C_i(k)' - R_i(k)')^2 \\ &+\lambda_{H.264}\cdot R \end{aligned} \tag{16}$$

where $\lambda_{H.264}$ indicates the Lagrange multiplier defined in H.264/AVC coding with the predefined quantization step Q_s.

From the residual normalization point of view, the distortion model calculates the SSD between the normalized original and distorted DCT coefficients, as shown in Fig. 2. Therefore, we can still use the Lagrange multiplier defined in H.264, $\lambda_{H.264}$, in this perceptual RDO scheme.

C. Sub-Band Level Normalization Factor Computation

In this sub-section, we show that the proposed method in section II-A can be improved further by fine tuning the DCT normalization matrix so that each AC coefficient has a different normalization factor. Motivated by the fact that the normalized DCT coefficients of residuals of different frequencies have different statistical distributions, we propose a frame level quantization matrix selection algorithm considering the perceptual quality of the reconstructed video. To begin with, we model the normalized transform coefficients x with Laplace distribution, which has been proved to achieve a good trade-off between model fidelity and complexity [52]:

$$f_{Lap}(x) = \frac{\Lambda}{2}\cdot e^{-\Lambda\cdot|x|} \tag{17}$$

where Λ is called the Laplace parameter.

From (14), 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

$$\frac{dJ}{dR} = \frac{dD}{dR} + \lambda = 0 \tag{18}$$

which yields

$$\lambda = -\frac{dD}{dR} = -\frac{\frac{dD}{dQ_s}}{\frac{dR}{dQ_s}}. \tag{19}$$

In [52], Laplace distribution based rate and distortion models were established to derive λ for each frame dynamically. However, all the transform coefficients were modeled with a single distribution and the variation in the distribution between DCT sub-bands was ignored. Here we model the distortion and rate in a similar way as in [52], where D is obtained by summing the perceptual distortion in each quantization interval and R is calculated with the help of the entropy of the normalized coefficients. Let $c_{i,j}^m$ be the DCT coefficient in the $(i,j)^{th}$ sub-band of the m^{th} block and $\hat{c}_{i,j}^m$ the reconstructed coefficient of the same position in the decoder, the perceptual

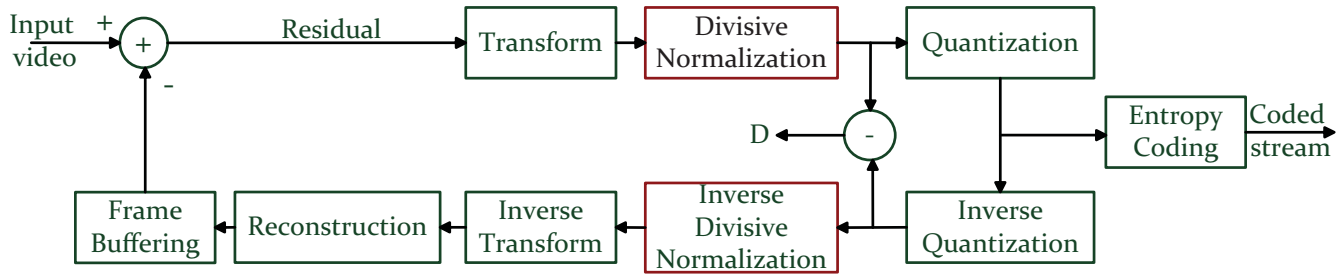


Fig. 2. Framework of the proposed scheme.

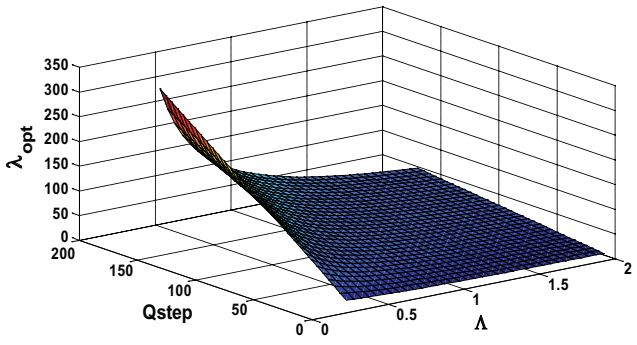


Fig. 3. Relationship between the optimal λ and (Λ, Qstep).

distortion for this sub-band $D_{i,j}$ is defined as

$$D_{i,j} = \frac{1}{N_B}\sum_{m=1}^{N_B}\left(\frac{c_{i,j}^m}{f_{i,j}^m} - \frac{\hat{c}_{i,j}^m}{f_{i,j}^m}\right)^2 = \frac{1}{N_B}\sum_{m=1}^{N_B}\left(c_{i,j}^{m'} - \hat{c}_{i,j}^{m'}\right)^2 \quad (20)$$

where N_B is the number of DCT blocks in each frame and $f_{i,j}^m$ represents the normalization factor for the $(i, j)^{th}$ sub-band of the m^{th} block; $c_{i,j}^{m'}$ and $\hat{c}_{i,j}^{m'}$ are the normalized coefficients of $c_{i,j}^m$ and $\hat{c}_{i,j}^m$, respectively.

More specifically, the perceptual distortion defined in (20) is equivalent to the MSE in the divisive normalization domain. If $x_{i,j}$ denotes the normalized coefficient in the $(i, j)^{th}$ sub-band, then $D_{i,j}$ can be modeled in the divisive normalization domain according to the quantization process in H.264/AVC, which is given by

$$D_{i,j} \approx \int_{-(Q_s-\gamma Q_s)}^{(Q_s-\gamma Q_s)} x_{i,j}^2 f_{Lap}(x_{i,j})dx_{i,j} + 2\sum_{n=1}^{\infty}\int_{nQ_s-\gamma Q_s}^{(n+1)Q_s-\gamma Q_s} \times(x_{i,j} - nQ_s)^2 f_{Lap}(x_{i,j})dx_{i,j} \quad (21)$$

where γ is the rounding offset. Subsequently, we model the rate of the $(i, j)^{th}$ sub-band by calculating its entropy [53]:

$$R_{i,j} = -P_0 \cdot \log_2 P_0 - 2\sum_{n=1}^{\infty} P_n \cdot \log_2 P_n \quad (22)$$

where P_0 and P_n are the probabilities of the transformed residuals quantized to the zero-th and n-th quantization levels, respectively, which can be modeled by the Laplace distribution

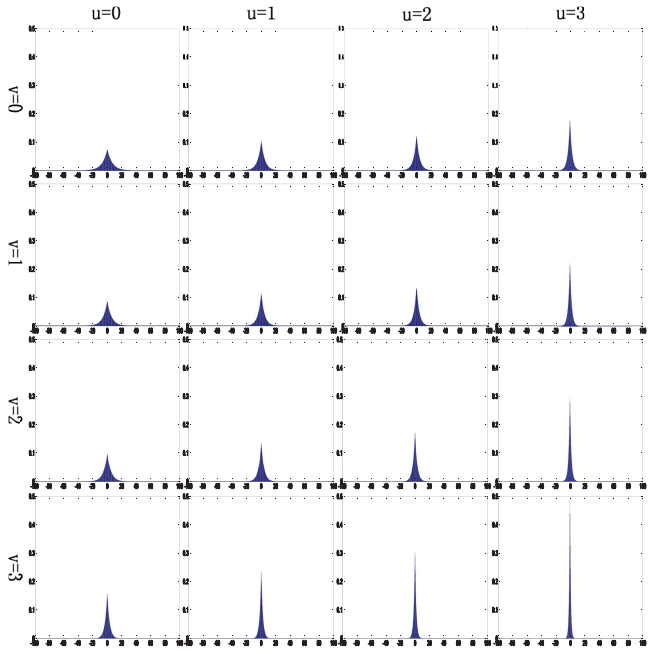


Fig. 4. Laplace distributions for DCT subband coefficients (Bus@CIF).

as

$$P_0 = \int_{-(Q_s-\gamma Q_s)}^{(Q_s-\gamma Q_s)} f_{Lap}(x_{i,j})dx \quad (23)$$

$$P_n = \int_{nQ_s-\gamma Q_s}^{(n+1)Q_s-\gamma Q_s} f_{Lap}(x_{i,j})dx. \quad (24)$$

Since the rounding offset can be regarded as a constant value for each frame, by incorporating (22) into (19), we conclude that the optimal Lagrange multiplier which controls the trade-off between R and D is a function of the Laplace parameter and the quantization step only, which is given by

$$\lambda_{opt} = f(\Lambda, Q_s). \quad (25)$$

The λ_{opt} for each (Λ, Q_s) is shown in Fig. 3, which confirms the idea that λ_{opt} increases monotonically with Q_s but decreases monotonically with Λ. It suggests that for the same λ_{opt} but different Λ, we will have different Q_s values.

Fig. 4 shows that the distribution of the normalized transform coefficients in different sub-bands have similar shape but different widths [53], [54], thus their optimal Lagrange multipliers should be different. However, in the current hybrid video coding framework, directly adjusting λ_{opt} for each

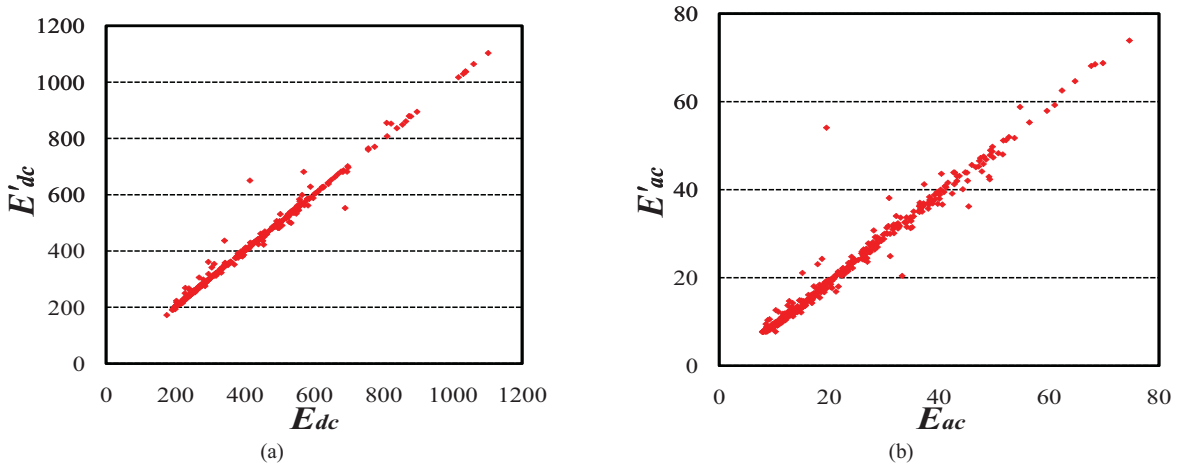


Fig. 5. (a) Relationship between E_{dc} and E'_{dc} at QP = 30 for Bus@CIF. (b) Relationship between E_{ac} and E'_{ac} at QP = 30 for Bus@CIF.

subband is impractical because the Lagrange multiplier needs to be uniform across the whole frame in RD optimization. To overcome this, we generate a uniform λ_{opt} for each subband by modifying Q_s values. Given the optimal λ_{opt}, the optimal quantization step for the $(i, j)^{th}$ sub-band is calculated as

$$Q_{i,j} = g(\lambda_{opt}, \Lambda_{i,j}). \tag{26}$$

In our implementation, we keep λ of the DC coefficients unaltered and modify Q_s of the AC coefficients. To obtain the optimal $Q_{i,j}$, we build a look-up table based on Fig. 3.

D. Implementation Issues

In video coding, the normalization factors defined in (11) and (11) need to be computed at both the encoder and the decoder. However, before coding the current frame, the distorted MBs are not available, which creates a chicken or egg causality dilemma. Moreover, at the decoder side, the original MB is not accessible either. Therefore, the normalization factors defined in (11) and (11) cannot be directly applied in practice. To overcome this problem, we propose to make use of the predicted MB, which is available at both the encoder and the decoder for the calculation of the normalization factors. As such, we do not need to transmit any additional overhead information to the decoder.

$$\begin{aligned}
E_{dc} &= \frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2 + Y_i(0)^2 + N\cdot C_1} \\
E'_{dc} &= \frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2 + N\cdot C_1} \\
E_{ac} &= \frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2 + Y_i(k)^2)}{N-1} + C_2} \\
E'_{ac} &= \frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(2\cdot Z_i(k)^2)}{N-1} + C_2}.
\end{aligned} \tag{27}$$

The relationship between E_{dc} and E'_{dc} as well as E_{ac} and E'_{ac} are illustrated in Fig. 5, where E_{dc}, E'_{dc}, E_{ac} and E'_{ac} are defined in (27). In these equations, $Z_i(k)$ is the k^{th} DCT coefficient of the i^{th} prediction sub-MB for each mode. We can observe dependency between the DC and AC energy values of the original and predicted MBs. Therefore, the DC and AC energy of the original MB can be approximated with the help of the corresponding energy of the prediction MB. Consequently, the approximation of the normalization factors can be determined by

$$f'_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2 + N\cdot C_1}}{\mathbb{E}(\sqrt{2Z(0)^2 + N\cdot C_1})} \tag{28}$$

$$f'_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(Z_i(k)^2 + s\cdot Z_i(k)^2)}{N-1} + C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(Z(k)^2 + s\cdot Z(k)^2)}{N-1} + C_2})}. \tag{29}$$

For intra mode, we use the MB at the same position in the previously coded frames.

In order to compensate for the loss of AC energy, we use a factor s to bridge the difference between the energy of AC coefficients in the prediction MB and the original MB, which can be defined as

$$s = \frac{\mathbb{E}(\sum_{k=1}^{N-1} X(k)^2)}{\mathbb{E}(\sum_{k=1}^{N-1} Z(k)^2)}. \tag{30}$$

As depicted in Fig. 6, we can approximate s by a linear relationship with Q_s, which can be modeled empirically as

$$s = 1 + 0.005\cdot Q_s. \tag{31}$$

In order to compute the normalization factors for DC and AC coefficients, as defined in (28) and (29), the DC and AC energy of the prediction MB should firstly be calculated. The DCT is an orthogonal transform that obeys Parseval's theorem. Thus we will have the following relations between the DCT coefficients and the spatial domain mean and variance:

$$\mu_x = \frac{\sum_{i=0}^{N-1} x(i)}{N} = \frac{X(0)}{\sqrt{N}} \tag{32}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{N-1} X(i)^2}{N-1} \tag{33}$$

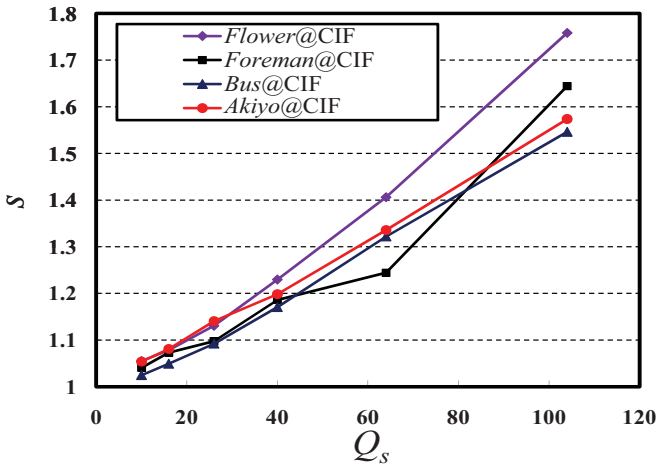


Fig. 6. Relationship between s and Q_s for different sequences.

Therefore, to calculate the normalization factors in (28) and (29), in the actual implementation for both the encoder and decoder, it is not necessary to perform the actual DCT transform. Instead, we only need to compute the mean and variance of the prediction block in spatial domain.

In our implementation, we combine the frame-level quantization matrix selection and divisive normalization together and employ one quantization matrix to achieve two goals. Analogous to [10], the quantization matrix for 4×4 DCT transform is defined as

$$WS_{ij} = 16 \cdot \begin{bmatrix} f'_{dc} \cdot \omega_{0,0} & f'_{ac} \cdot \omega_{0,1} & f'_{ac} \cdot \omega_{0,2} & f'_{ac} \cdot \omega_{0,3} \\ f'_{ac} \cdot \omega_{1,0} & f'_{ac} \cdot \omega_{1,1} & f'_{ac} \cdot \omega_{1,2} & f'_{ac} \cdot \omega_{1,3} \\ f'_{ac} \cdot \omega_{2,0} & f'_{ac} \cdot \omega_{2,1} & f'_{ac} \cdot \omega_{2,2} & f'_{ac} \cdot \omega_{2,3} \\ f'_{ac} \cdot \omega_{3,0} & f'_{ac} \cdot \omega_{3,1} & f'_{ac} \cdot \omega_{3,2} & f'_{ac} \cdot \omega_{3,3} \end{bmatrix} \quad (34)$$

where

$$\omega_{i,j} = Q_{i,j}/Q_s. \quad (35)$$

The Laplace parameter $\Lambda_{i,j}$ and the expectation of the energy (as indicated in (11)) should be available before coding the current frame. However, these quantities can only be obtained after coding it. As they are approximately constants during a very short period of time, we estimate them by averaging their corresponding values from previous frames coded in the same manner

$$\hat{\Lambda}^t_{i,j} = \frac{1}{N_f} \sum_{n=1}^{N_f} \Lambda^{t-n}_{i,j} \quad (36)$$

where t indicates the frame number and N_f represents the number of previous frames used. Practically, N_f is set to be 3 in this paper.

At the decoder, the Laplace distribution parameters of the normalized coefficients in each sub-band are not available. To address this issue, we transmit the frame-level quantization matrix to the decoder. As the statistics of frames in a short time do not change considerably, we empirically define a threshold to determine whether to refresh the quantization matrix, which is expressed as

$$\omega^t = \begin{cases} \omega^{t-1} & \sum(\omega^t_{i,j} - \omega^{t-1}_{i,j})^2 < T_r \\ \omega^t & \text{otherwise} \end{cases} \quad (37)$$

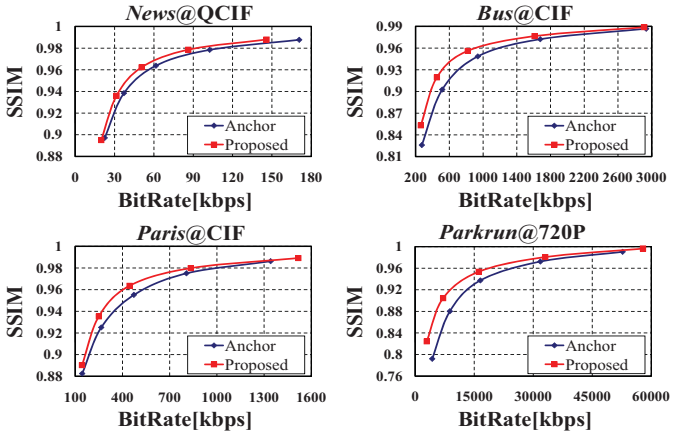


Fig. 7. Rate-SSIM Performance comparisons (Anchor: H.264/AVC).

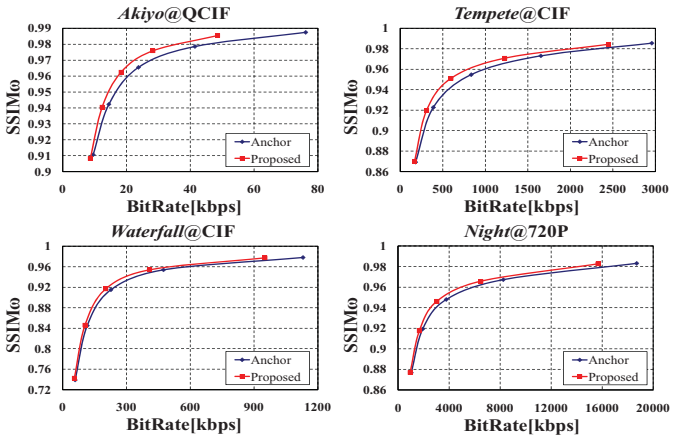


Fig. 8. Rate-SSIM$_\omega$ Performance comparisons (Anchor: H.264/AVC).

where

$$\omega = 16 \cdot \begin{bmatrix} \omega_{0,0} & \omega_{0,1} & \omega_{0,2} & \omega_{0,3} \\ \omega_{1,0} & \omega_{1,1} & \omega_{1,2} & \omega_{1,3} \\ \omega_{2,0} & \omega_{2,1} & \omega_{2,2} & \omega_{2,3} \\ \omega_{3,0} & \omega_{3,1} & \omega_{3,2} & \omega_{3,3}. \end{bmatrix} \quad (38)$$

We set the threshold T_r to be 100 to balance the transmitted bits and the accuracy of the matrix. Empirically, we find this to be a non-sensitive parameter as the quantization matrix of each frame is very stable and the transmission of the matrix takes only a small number of bits.

III. Validations

To validate the proposed scheme, we integrate it into H.264/AVC reference software JM15.1 [55]. All test video sequences are in YCbCr 4:2:0 format. The common coding configurations are set as follows: all available inter and intra modes are enabled; five reference frames; one I frame followed by all P frames; high complexity RDO and fixed quantization parameters (QP).

A. Objective Performance Evaluation of the Proposed Scheme

The RD performance is measured in two cases: SSIM of Y component only and SSIM of Y, Cb and Cr components,

TABLE I
PERFORMANCE OF THE PROPOSED ALGORITHMS (COMPARED WITH H.264/AVC VIDEO CODING)

Sequence	QP_1={18, 22, 26, 30}				QP_2={26, 30, 34, 38}			
	$\Delta SSIM$	ΔR	$\Delta SSIM_\omega$	ΔR_ω	$\Delta SSIM$	ΔR	$\Delta SSIM_\omega$	ΔR_ω
Akiyo (QCIF)	0.0038	–20.5%	0.0044	–23.0%	0.0091	–14.0%	0.0084	–14.6%
Bridge-close (QCIF)	0.0066	–33.1%	0.0069	–28.3%	0.0289	–42.5%	0.0241	–42.6%
Carphone (QCIF)	0.0022	–12.9%	0.0027	–14.1%	0.0040	–8.2%	0.0042	–9.2%
Coastguard (QCIF)	0.0034	–7.0%	0.0027	–6.6%	0.0094	–9.0%	0.0075	–8.7%
Container (QCIF)	0.0024	–10.5%	0.0007	–3.9%	0.0046	–12.3%	0.0034	–10.9%
Grandma (QCIF)	0.0063	–20.0%	0.0066	–21.5%	0.0131	–14.6%	0.0119	–15.0%
News (QCIF)	0.0033	–15.7%	0.0034	–15.1%	0.0078	–13.2%	0.0077	–13.4%
Salesman (QCIF)	0.0041	–12.6%	0.0050	–14.3%	0.0136	–12.2%	0.0127	–12.7%
Akiyo (CIF)	0.0029	–20.5%	0.0032	–23.4%	0.0043	–12.5%	0.0042	–13.4%
Bus (CIF)	0.0048	–17.1%	0.0041	–14.6%	0.0205	–23.7%	0.0170	–23.2%
Coastguard (CIF)	0.0033	–7.4%	0.0028	–7.4%	0.0119	–11.7%	0.0097	–11.7%
Flower (CIF)	0.0036	–23.0%	0.0052	–24.7%	0.0092	–19.2%	0.0111	–22.1%
Mobile (CIF)	0.0014	–9.2%	0.0020	–9.7%	0.0056	–14.0%	0.0058	–13.8%
Paris (CIF)	0.0036	–15.0%	0.0025	–10.1%	0.0109	–17.9%	0.0091	–15.9%
Tempete (CIF)	0.0023	–13.4%	0.0035	–15.9%	0.0084	–14.7%	0.0084	–15.2%
Waterfall (CIF)	0.0038	–13.1%	0.0042	–12.7%	0.0132	–10.5%	0.0118	–10.5%
BigShip (720P)	0.0040	–11.8%	0.0036	–12.10%	0.0051	–7.3%	0.0044	–7.5%
Night (720P)	0.0030	–13.0%	0.0031	–14.1%	0.0064	–11.5%	0.0060	–12.0%
Spincalendar (720P)	0.0046	–19.9%	0.0024	–11.60%	0.0035	–13.8%	0.0017	–9.1%
Parkrun (720P)	0.0084	–3.9%	0.0066	–15.2%	0.0317	–36.5%	0.0257	–35.4%
Average	0.0039	–15.0%	0.0038	–14.9%	0.0111	–16.0%	0.0097	–15.8%

respectively. To apply SSIM to all three color components, we combine the SSIM indices of these components by [56]

$$SSIM_\omega = W_Y \cdot SSIM_Y + W_{Cb} \cdot SSIM_{Cb} + W_{Cr} \cdot SSIM_{Cr} \tag{39}$$

where $W_Y = 0.8$, $W_{Cb} = 0.1$ and $W_{Cr} = 0.1$ are the weights assigned to Y, Cb and Cr components, respectively. These quantities for the whole video sequence are obtained by simply averaging the respective values of individual frames. The method proposed in [57] is used to calculate the differences between two RD curves.

We use two different sets of QP values in the experiments: $QP_1 = \{22, 26, 30, 34\}$ and $QP_2 = \{26, 30, 34, 38\}$, which represent high bit-rate and low bit-rate coding configurations, respectively. From Table I, it can be observed that over a wide range of test sequences with resolutions from QCIF to 720P, the proposed scheme achieves average rate reduction of 15.0% for QP_1 and 16.0% for QP_2 for fixed SSIM values and the maximum coding gain is 42.5%. It can also be observed that our scheme performs better when there exist significant statistical differences between different regions in the same frame, for example, in the cases of *Bus* and *Flower*. This is likely because these frames allow us to borrow bits more aggressively from the regions with complex texture or high contrast (high normalization factor) and allocating them to the regions with relatively simple textures (low normalization factor).

The R-D performances for sequences with various resolutions are shown in Figs. 7 and 8. It can be observed that the proposed scheme achieves better R-D performance over the full range of QP values. Moreover, the gains become more significant at middle bit-rates. The reason may be that at high bit rate, the quantization step is small and thus the differences of quantization steps among the MBs are not significant, while at low bit rate, since the AC coefficients are severely distorted, the normalization factors derived from the prediction frame do not precisely represent the properties of the original frame.

TABLE II
COMPLEXITY OVERHEAD OF THE PROPOSED SCHEME

Sequences	ΔT in Encoder	ΔT in Decoder
Akiyo (QCIF)	1.20%	8.97%
News (QCIF)	1.17%	11.30%
Mobile (QCIF)	1.34%	5.3%
Bus (CIF)	1.16%	9.16%
Flower (CIF)	1.11%	8.75%
Tempete (CIF)	0.96%	7.38%
Average	1.16%	8.48%

When evaluating the coding complexity overhead, we calculate ΔT as

$$\Delta T = \frac{T_{pro} - T_{H.264}}{T_{H.264}} \times 100\% \tag{40}$$

where $T_{H.264}$ and T_{pro} indicate the total coding time for the sequence with H.264/AVC and the proposed schemes, respectively. Table II shows the computational overhead for both encoding and decoding. The coding time is obtained by encoding 100 frames of IPPP GOP structure with Intel 2.83 GHz Core processor and 4GB random access memory. As indicated in Section II-D, we do not need to perform DCT

TABLE III

SSIM INDICES AND BIT RATES OF TESTING SEQUENCES USED IN THE SUBJECTIVE TEST I. (SIMILAR BIT RATE BUT DIFFERENT SSIM VALUES)

Sequences	H.264/AVC		Proposed	
	SSIM	Bit Rate	SSIM	Bit Rate
Bridge-close (QCIF)	0.8892	29.56	0.9216	29.07
Bus (CIF)	0.8259	273.7	0.8531	262.03
Flower (CIF)	0.9121	317.8	0.9170	296.43
Mobile (CIF)	0.9462	631.89	0.9532	630.69
Paris (CIF)	0.8825	144.2	0.8902	142.59
Parkrun (720P)	0.7921	4311.6	0.8527	3768.34

TABLE IV

SSIM INDICES AND BIT RATES OF TESTING SEQUENCES USED IN THE SUBJECTIVE TEST II. (SIMILAR SSIM VALUES BUT DIFFERENT BIT RATE)

Sequences	H.264/AVC		Proposed	
	SSIM	Bit Rate	SSIM	Bit Rate
Bridge-close (QCIF)	0.8777	23.35	0.8764	12.76
News (QCIF)	0.9784	102.51	0.9786	86.18
Waterfall (CIF)	0.9619	474.09	0.962	408.79
Mobile (CIF)	0.9462	631.89	0.9467	537.78
Night (720P)	0.9845	18706.85	0.9839	15671.46
Bigship (720P)	0.9018	1552.8	0.9015	1390.08

transform at either the encoder or the decoder. Therefore, it is observed that the encoding overhead is negligible (1.16% on average). The complexity of the decoder is increased by 8.48% on average.

B. Subjective Performance Evaluation of the Proposed Scheme

To further validate our scheme, we carried out two subjective quality evaluation tests based on a two-alternative-forced-choice (2AFC) method. This method is widely used in psychophysical studies [58], [59], 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. For each subjective test, we selected six pairs of sequences with different resolutions. In the first test, the sequences were compressed by H.264/AVC and the proposed method at the same bit rate but with different SSIM levels. In the second test, the sequences were coded to achieve the same SSIM levels (where the proposed scheme uses much lower bit rates). Tables III and IV list all the test sequences as well as their SSIM values and bit rates. In the 2AFC test, each pair is repeated four times with random order. As a result, in each test we obtained 24 2AFC results for each subject. Eight subjects participated in the experiments.

The results of the two subjective tests are reported in Figs. 9 and 10, respectively. In each figure, the percentage by which the subjects are in favor of the H.264/AVC against the proposed scheme are shown. We also plot the error bars ($\pm$ one standard deviation between the measurements) over the eight subjects and over the six sequences. As can be observed in Fig. 9, the subjects are inclined to select the proposed method for better video quality. On the contrary, for the second test in Fig. 10, it turns out that for almost all cases the percentage is close to 50% and nearly all error bars cross the 50% line. These results provide useful evidence that the proposed method achieves the same level of quality with lower bit rates or creates better quality video at the same bit rates.

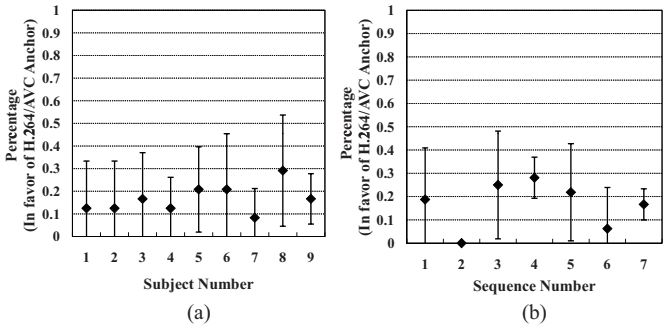


Fig. 9. Subjective test 1: Similar bit rate with different SSIM values. (a) Mean and standard deviation (shown as error-bar) of preference for individual subject (1∼8: subject number, 9: average). (b) Mean and standard deviation (shown as error-bar) of preference for individual sequence (1∼6: sequence number, 7: average).

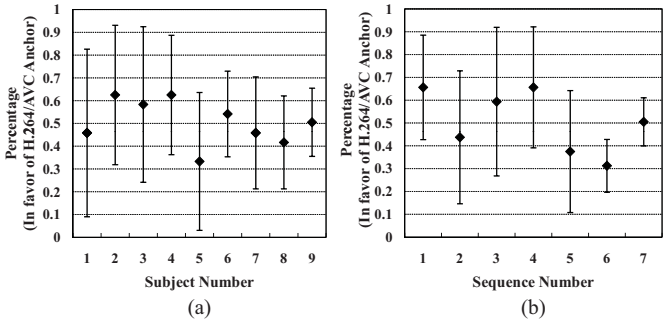


Fig. 10. Subjective test 2: Similar SSIM with different bit rates. (a) Mean and standard deviation (shown as error-bar) of preference for individual subject (1∼8: subject number, 9: average). (b) Mean and standard deviation (shown as error-bar) of preference for individual sequence (1∼6: sequence number, 7: average).

C. Comparisons With State-of-the-Art Algorithms

To show the advantage of our divisive normalization scheme, the performance comparisons of the proposed scheme, the state of the art SSIM based RDO scheme [31] and standard quantization matrix based video coding scheme in H.264/AVC are shown in Fig. 11. In this experiment, IPP GOP structure and CABAC coding techniques are used. The QP values range from 23 to 38 with an interval of 5. For most of the sequences, the proposed divisive normalization scheme achieves better coding performance. As discussed before, our scheme performs better especially for the sequences with significant statistical differences in the same frame, such as *Flower* and *Bus*. On average, compared with the SSIM based RDO scheme [31], the proposed scheme achieves better rate reduction of –17.7% vs –13.0%.

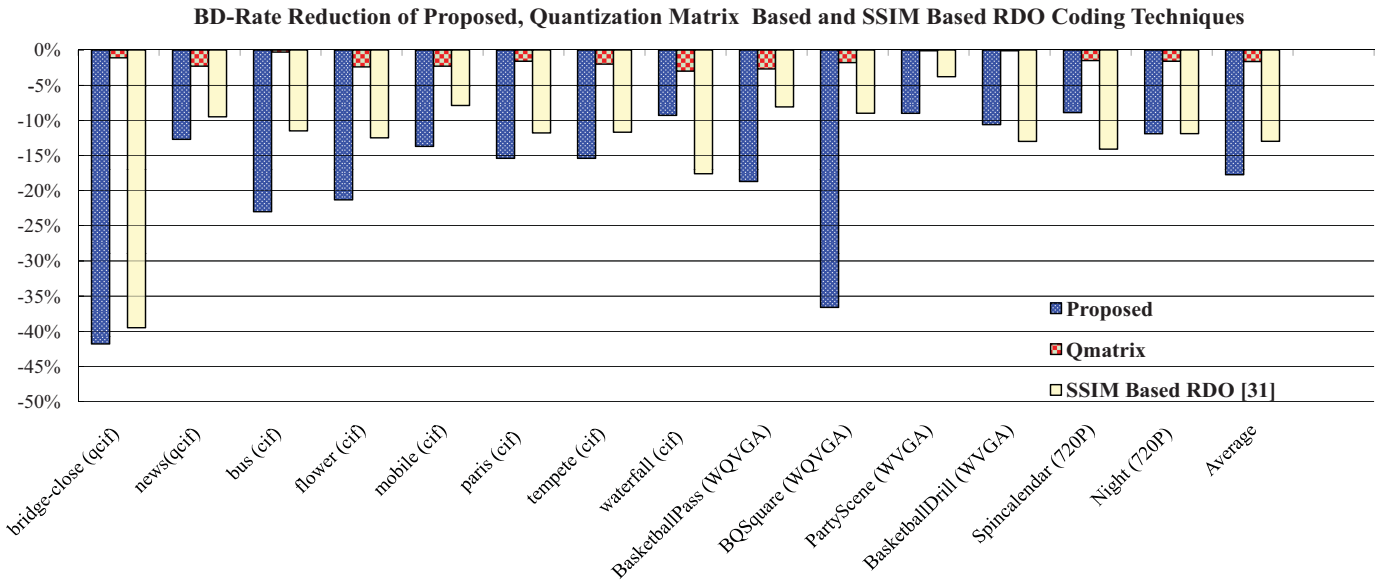


Fig. 11. Performance comparisons of the proposed, quantization matrix and the SSIM based RDO coding techniques. (Anchor: conventional H.264/AVC)

IV. Conclusion

We propose an SSIM-inspired novel residual divisive normalization scheme for perceptual video coding. The novelty of the scheme lies in normalizing the transform coefficients based on the DCT domain SSIM index and defining a new distortion model for the subsequent rate distortion optimization. We show two applications based on this divisive normalization scheme, which are MB-level mode selection and frame-level quantization matrix selection, respectively. The proposed scheme demonstrates superior performance as compared to H.264/AVC video codec by offering significant rate reduction, while keeping the same level of SSIM values. Visual quality improvement is also achieved by the proposed scheme.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments that significantly helped them in improving the presentation of the manuscript of this paper.

References

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

[2] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment* (Syntheses Lectures on Image, Video and Multimedia Processing). San Rafael, CA: Morgan Claypool, Mar. 2006.

[3] 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.

[4] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 11–18, Feb. 2006.

[5] C.-W. Tang, "Spatial temporal visual considerations for effcient video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Jan. 2007.

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

[7] 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.

[8] 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.

[9] J. Chen, J. Zheng, and Y. He, "Macroblock-level adaptive frequency weighting for perceptual video coding," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 775–781, May 2007.

[10] A. Tanizawa and T. Chujoh, "Adaptive quantization matrix selection," Toshiba, Geneva, Switzerland, Tech. Rep. T05-SG16-060403-D-0266, Apr. 2006.

[11] T. Suzuki, P. Kuhn, and Y. Yagasaki, "Quantization tools for high quality video," ISO/IEC, Washington, DC, Tech. Rep. MPEG ITU-T VCEG JVT-B067, Jan. 2002.

[12] T. Suzuki, K. Sato, and Y. Yagasaki, "Weighting matrix for JVT codec," ISO/IEC, Washington, DC, Tech. Rep. MPEG ITU-T VCEG JVT-C053, May 2002.

[13] J. Malo, J. Gutierrez, I. Epifanio, F. Ferri, and J. M. Artigas, "Perceptual feedback in multigrid motion estimation using an improved DCT quantization," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1411–1427, Oct. 2001.

[14] Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[15] T. Pappas, T. Michel, and R. Hinds, "Supra-threshold perceptual image coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 1996, pp. 237–240.

[16] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG 2000," *Signal Process. Image Commun.*, vol. 17, no. 1, pp. 85–104, Jan. 2001.

[17] D. Chandler and S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 397–410, Apr. 2005.

[18] 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.

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

[20] 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.

[21] 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.*, May 2005, pp. 2673–2678.

[22] 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. Acoust., Speech, Signal Process.*, vol. 2. Feb. 2006, pp. 913–916.

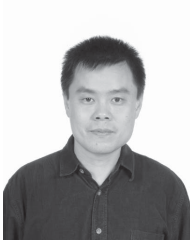
[23] C. Yang, R. Leung, L. Po, and Z. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 4. Sep. 2009, pp. 291–295.
[24] C. Yang, H. Wang, and L. Po, "Improved inter prediction based on structural similarity in H.264," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, vol. 2. Aug. 2007, pp. 340–343.
[25] 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.
[26] H. Chen, Y. Huang, P. Su, and T. Ou, "Improving video coding quality by perceptual rate-distortion optimization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 1287–1292.
[27] 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, pp. 1–6.
[28] Y. Huang, T. Ou, and H. H. Chen, "Perceptual-based coding mode decision," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 393–396.
[29] T. Ou, Y. Huang, and H. Chen, "A perceptual-based approach to bit allocation for H.264 encoder," *Proc. SPIE Visual Commun. Image Process.*, vol. 7744, p. 77441B, Jul. 2010.
[30] 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.
[31] S. Wang, A. Rehman, Z. 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.
[32] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1488–1499, Apr. 2012.
[33] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, 2000, pp. 855–861.
[34] S. Lyu and E. P. Simoncelli, "Statistically and perceptually motivated nonlinear image representation," *Proc. SPIE*, vol. 6492, pp. 67–91, Jan. 2007.
[35] J. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Opt. Soc. Amer.*, vol. 11, no. 6, pp. 1710–1719, 1994.
[36] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer.*, vol. 14, no. 9, pp. 2379–2391, 1997.
[37] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neurosci.*, vol. 4, no. 8, pp. 819–825, Aug. 2001.
[38] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neural Sci.*, vol. 9, no. 2, pp. 181–198, 1992.
[39] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vis. Res.*, vol. 38, pp. 743–761, Mar. 1998.
[40] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.
[41] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.
[42] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Non-linear image representation for efficient perceptual coding," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 68–80, Jan. 2006.
[43] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-inspired divisive normalization for perceptual video coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1657–1660.
[44] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[45] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, 2000, pp. 855–861.
[46] 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.
[47] J. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Opt. Soc. Amer.*, vol. 11, no. 6, pp. 1710–1719, 1994.
[48] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer.*, vol. 14, no. 9, pp. 2379–2391, 1997.
[49] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vis. Res.*, vol. 38, pp. 743–761, Mar. 1998.
[50] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, p. 011006, Jan. 2010.
[51] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[52] 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.
[53] X. Zhao, J. Sun, S. Ma, and W. Gao, "Novel statistical modeling, analysis and implementation of rate-distortion estimation for H.264/AVC coders," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 5, pp. 647–660, May 2010.
[54] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.
[55] *Joint Video Team (JVT) Reference Software*. (2010) [Online]. Available: http://iphome.hhi.de/suehring/tml/download/old-jm
[56] 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.
[57] G. Bjontegaard, "Calculation of average PSNR difference between RD curves," in *Proc. 13th Meeting ITU-T Q.6/SG16 VCEG*, Austin, TX, Apr. 2001.
[58] S. Winkler, "Analysis of public image and video database for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.
[59] *Recommendation 500-10: Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Standard Rec. BT.500, 2012.

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 with Peking University, Beijing, China. He was a Visiting Student with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, from 2010 to 2011.

He was with Microsoft Research Asia, Beijing, as an Intern, in 2011. 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, and the M.Sc. degree in communications engineering from Technical University Munich, Munich, Germany, in 2007 and 2009, respectively. He is currently pursuing the Ph.D. degree with the University of Waterloo, Waterloo, ON, Canada.

He has been a Research Assistant with the Department of Electrical and Computer Engineering, University of Waterloo, since 2009. In 2011, he was with the Video Compression Research Group, 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. He has authored or co-authored more than 100 papers with over 10,000 citations (Google Scholar). His current research interests include image processing, coding, quality assessment, computational vision, pattern analysis, multimedia communications, and biomedical signal processing.

Dr. Wang was a recipient of the 2009 IEEE Signal Processing Society Best Paper Award, the ICIP 2008 IBM Best Student Paper Award, and the 2009 Ontario Early Researcher Award. He has been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2009 and *Pattern Recognition* since 2006, and a Guest Editor of *Signal, Image and Video Processing* since 2011. He was an Associate Editor of the *IEEE Signal Processing Letters* from 2006 to 2010, and a Guest Editor of the *IEEE Journal of Selected Topics in Signal Processing* from 2007 to 2009, and the *EURASIP Journal of Image and Video Processing* from 2009 to 2010.

Siwei Ma (S'03–M'12) received the B.S. degree from Shandong Normal University, Jinan, China, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1999 and 2005, respectively.

He joined the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, where he is currently an Associate Professor. He was a Post-Doctorate Researcher with the University of Southern California, Los Angeles, from 2005 to 2007. He has authored or co-authored over 100 technical articles in refereed journals and proceedings on 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.

He is currently a Professor of computer science with Peking University, Beijing, China. He was a Professor of computer science with Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or co-authored over 600 technical articles in refereed journals and conference proceedings, and five books on image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Dr. Gao has been on the Editorial Board 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 was the Chair of a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE ICME and the ACM Multimedia, and was on the advisory and the technical committees of numerous professional organizations.