Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric

Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, and Homer H. Chen, *Fellow, IEEE*

***Abstract*—The rate-distortion optimization (RDO) framework for video coding achieves a tradeoff between bit-rate and quality. However, objective distortion metrics such as mean squared error traditionally used in this framework are poorly correlated with perceptual quality. We address this issue by proposing an approach that incorporates the structural similarity index as a quality metric into the framework. In particular, we develop a predictive Lagrange multiplier estimation method to resolve the chicken and egg dilemma of perceptual-based RDO and apply it to H.264 intra and inter mode decision. Given a perceptual quality level, the resulting video encoder achieves on the average 9% bit-rate reduction for intra-frame coding and 11% for inter-frame coding over the JM reference software. Subjective test further confirms that, at the same bit-rate, the proposed perceptual RDO indeed preserves image details and prevents block artifact better than traditional RDO.**

***Index Terms*—H.264, Lagrange multiplier, perceptual quality, rate-distortion optimization, structural similarity index, video codec.**

I. Introduction

The H.264 VIDEO coding standard [1] achieves better performance than the previous video coding standards and has become a key technology for multimedia communications and consumer electronics. The high performance of this standard is attributed to the adoption of the rate-distortion optimization (RDO) framework [2] and a variety of coding modes [3]. However, objective distortion metrics such as mean squared error (MSE) traditionally used in the RDO framework are poorly correlated with perceptual quality. Perceptual-based RDO is desired.

Recently, there has been an increasing amount of research on perceptual-based video coding [4]–[12]. Since video quality is ultimately judged by human eyes, perceptual-based video coding takes the characteristics of the human visual system (HVS) into account in the design of video coding systems.

The goal of RDO is to minimize the overall distortion D at a given rate R by an appropriate selection of the coding mode for each coding unit. If an additive distortion metric is used, D is obtained by summing the distortion of all coding units, and the RDO problem can be formulated as follows:

$$\min\left\{\sum_{i=1}^{N_u} d_i\right\} \text{subject to} \sum_{i=1}^{N_u} r_i \le R_c \tag{1}$$

where N_u is the number of coding units, d_i and r_i are the distortion and bit-rate of the ith coding unit, respectively, and R_c is the maximum bit-rate allowed. There are two popular approaches to this problem: Lagrangian technique and dynamic programming (DP) technique [13]. In practice, the R-D curve is composed of discrete operating R-D points. When the operating R-D points on the R-D curve are sparse and none of them meets the bit budget constraint, the DP technique can still guarantee optimality, but not the Lagrangian technique [13]. However, the computational complexity of the DP technique is too high for practical applications [13]. As a result, most video coding systems, including H.264, adopt the Lagrangian technique to solve the RDO problem [2], [13]–[16].

The basic idea of the Lagrangian technique is to convert a constrained optimization problem to an unconstrained one by reformulating the RDO problem as follows:

$$\min\{J\}, \text{ where } J = \sum_{i=1}^{N_u} d_i + \lambda \sum_{i=1}^{N_u} r_i. \tag{2}$$

In this formulation, J is the cost function, and λ is the so-called Lagrange multiplier that controls the tradeoff between distortion and bit-rate. For simplicity, it is assumed that the selection of the coding mode can be made independently for each coding unit without affecting the other units. One can obtain

$$\min\{J\} = \sum_{i=1}^{N_u} \min(d_i + \lambda r_i) \tag{3}$$

so that the minimum cost can be computed independently for each coding unit. It is worthwhile to point out that the optimal λ is the negative slope of the tangent to the R-D curve. Assuming that the R-D curve is convex and both R and D are

Manuscript received September 27, 2009; revised March 4, 2010 and July 18 2010; accepted September 2, 2010. Date of publication October 14, 2010; date of current version November 5, 2010. This work was supported in part by the National Science Council of Taiwan, under Contract NSC 98-2221-E-002-078-MY2, and in part by National Taiwan University, under Contract NTU97R0062-04. This paper was recommended by Associate Editor J. Ridge.

Y.-H. Huang is with MediaTek, Hsinchu 30078, Taiwan (e-mail: yihsin.huang@mediatek.com).

T.-S. Ou and P.-Y. Su are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: odeson24@gmail.com; b95901164@ntu.edu.tw).

H. H. Chen is with the Department of Electrical Engineering, the Graduate Institute of Communication Engineering, and the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan (e-mail: homer@cc.ee.ntu.edu.tw).

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.2010.2087472

differentiable everywhere, the minimum J can be derived by setting the derivative of J to zero as follows:

$$\frac{dJ}{dR} = \frac{dD}{dR} + \lambda = 0. \tag{4}$$

Therefore, we obtain

$$\lambda = -\frac{dD}{dR}. \tag{5}$$

Since the R-D characteristic of the video content cannot be obtained until the video is encoded and the video cannot be encoded unless the R-D characteristic is given (a well-known chicken and egg dilemma), the Lagrange multiplier cannot be directly calculated from (5). Methods have been proposed for modeling the R-D characteristic [7], [8], [15], [16]. However, none of them is both content-adaptive and computationally efficient. In addition, most of them are not perceptual-based.

The distortion metric used for measuring video quality in the RDO process has a profound impact on video coding. Although MSE and the like, such as sum of squared difference (SSD) and sum of absolute difference, have been widely adopted as distortion metric due to their simplicity, they are poorly correlated with human perception [17]. Distorted images with nearly identical MSE may have very different levels of perceptual distortion. To resolve this problem, many perceptual image/video quality assessment metrics have been proposed [18]–[21]. These metrics can replace the SSD in the RDO framework to improve the performance of a video encoder. Among the various perceptual image/video quality assessment metrics reported in the paper, the structural similarity (SSIM) index has been shown to be effective and computationally efficient [18].

In this paper, we use the well-known SSIM index as the distortion metric in the RDO framework and apply the resulting perceptual RDO to H.264 intra-frame and inter-frame coding. A novel predictive Lagrange multiplier estimation method is proposed to set the ground and resolve the chicken and egg dilemma of the perceptual-based RDO.

The remainder of the paper is organized as follows. In Section II, a brief introduction of the SSIM and a literature survey on Lagrange multiplier estimation are given. Section III describes the perceptual-based rate-distortion formulation of the cost function for H.264 mode decision. The Lagrange multiplier estimation method is detailed in Section IV. Section V describes the experimental results followed by a conclusion in Section VI.

II. Background and Related Work

The SSIM index and Lagrange multiplier estimation methods are reviewed in this section.

A. Structural Similarity Index

The basis of the SSIM index as a means for image quality assessment is that the HVS is highly adapted for extracting structural information from the image [18]. The SSIM index measures the SSIM as well as the luminance and contrast similarity between two images block by block. The SSIM index between two image blocks $\mathbf{x}$ and $\mathbf{y}$ is defined as follows:

$$SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y}) \cdot c(\mathbf{x}, \mathbf{y}) \cdot s(\mathbf{x}, \mathbf{y}) \tag{6}$$

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{7}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{8}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{9}$$

where $l(\mathbf{x}, \mathbf{y})$ compares the luminance of the two blocks, $c(\mathbf{x}, \mathbf{y})$ compares the contrast, $s(\mathbf{x}, \mathbf{y})$ measures the structural correlation, and μ_x and μ_y denote the sample means of $\mathbf{x}$ and $\mathbf{y}$, respectively, ${\sigma_x}^2$ and ${\sigma_y}^2$ denote the sample variance of $\mathbf{x}$ and $\mathbf{y}$, respectively, σ_{xy} is the sample cross-covariance between $\mathbf{x}$ and $\mathbf{y}$, and the three constants, C_1, C_2, and C_3, are introduced to avoid unstable behavior in regions of low luminance or low contrast [18]. The block size is typically 8 × 8, and the final SSIM value of is the averaged SSIM of all blocks.

The more similar two images are, the higher the SSIM index is. The maximum SSIM index is 1, which occurs when the two images are identical. The SSIM index marks itself a perceptual-based image quality assessment metric because it takes into account properties of HVS such as luminance adaptation and textural masking [17].

Recently, many image/video quality assessment methods based on the SSIM index have been developed [19]–[21] and adopted in many image/video applications [22], [23]. In the JM reference software version 15.1 [24], the SSIM index is an optional distortion metric.

B. Lagrange Multiplier Estimation for H.264

Lagrange multiplier estimation methods can be classified into two categories: heuristic [13] and analytical [15], [16]. Methods of the latter have better computational efficiency and prediction accuracy. Most methods proposed in the paper, including the one in the JM reference software, belong to the analytical category.

The JM reference software uses a rate model and a distortion model derived under the high rate assumption to determine the Lagrange multiplier [14], [15]. Basically, the rate-distortion function is derived as follows:

$$R(D) = a \log_2 \left(\frac{b}{D}\right) \tag{10}$$

where a and b are parameters, whose values depend on the content. On the contrary, using a uniform distribution to approximate the source probability within each quantization interval, one can obtain the distortion model as follows:

$$D = \frac{\Delta^2}{12} \tag{11}$$

where Δ is the quantization step size. Therefore, taking the first derivative of (10), we can determine the Lagrange multiplier by

$$\lambda = -\frac{dD}{dR} = c \times \Delta^2 \tag{12}$$

where c is a constant equal to $0.85 \times 2^{-8/3}$ in the JM reference software.

However, if the high-rate assumption of the above derivation is not true, this method may fail. An improved R-D model based on the Laplace distribution has been developed [16]. Although it outperforms the baseline R-D model [15], the formula for λ is complex and may not be easy to implement in practice [16].

On the contrary, perceptual-based video coding has recently gained increasingly more attention. Several studies on perceptual-based Lagrange multiplier estimation have been reported. For example, in [7] and [8], the Lagrange multiplier of current macroblock (MB) is slightly adjusted around the value predicted by the baseline R-D model according to HVS properties. These methods attempt to take HVS properties into account, but they still use SSD as the distortion metric in the RDO framework. Therefore, the improvement is limited. Mai *et al.* [4], [5] use the SSIM index as distortion metric for motion estimation and intra mode decision. The relation between the Lagrange multiplier, λ, and the quantization parameter, QP, is determined empirically as in [15]. Because the same λ is used for various sequences, this method is not content-adaptive.

III. Formulation of Perceptual-Based Mode Decision

We use the SSIM index as the distortion metric in the RDO framework and apply it to the mode decision of H.264 intra-frame and inter-frame coding. It is straightforward to express the cost function for H.264 mode decision in terms of the SSIM index.

For a MB of an intra frame, the mode decision process is carried out in two stages: prediction mode decision and MB mode decision. We define the cost function of mode decision as follows:

$$J(s, c, mode|QP) = D(s, c, mode|QP) + \lambda R(s, c, mode|QP) \tag{13}$$

where s and c denote the original and reconstructed image block, respectively, QP is the quantization parameter, and mode denotes either a prediction mode or a MB mode.

The H.264 High Profile supports three MB modes: intra 4×4 (I4 MB), intra 8×8 (I8 MB), and intra 16×16 (I16 MB). Each of them corresponds to a different block size as indicated in its name. In the H.264 Baseline Profile, only I4 MB and I16 MB are supported. For intra coding, nine prediction modes are supported in H.264 as follows:

$$\{Horizontal, Vertical, DC, Diagonal\_right, Diagonal\_left,$$
$$Horizontal\_up, Horizontal\_down, Vertical\_left, Vertical\_right\}.$$

Therefore, a block of the I4 MB mode is of size 4×4 and has nine prediction modes. The I8 MB mode has the same prediction modes as the I4 MB mode except that its block size is 8×8. However, I16 MB only has four prediction modes [1].

In perceptual-based RDO, we use the SSIM index instead of the SSD used by traditional video codecs to measure the distortion between the original image block and the reconstructed image block. The SSIM index is larger when the image quality is better. For convenience, we define the distortion metric D_{SSIM} by

$$D_{SSIM}(s, c, mode|QP) = 1 - SSIM(s, c). \tag{14}$$

Then, the cost function of mode decision can be expressed as

$$\begin{aligned} J(s, c, mode|QP) = {} & D_{SSIM}(s, c, mode|QP) \\ & +\lambda_{SSIM} R(s, c, mode|QP). \end{aligned} \tag{15}$$

This cost function is used in the MB mode decision and the prediction mode decision of I4 MB and I8 MB. The prediction mode decision for I16 MB does not involve RDO [1], and hence we need not discuss it. To be consistent with the coding structure of H.264, the SSIM index is calculated on a 4×4 block basis. The D_{SSIM} of a MB is approximated by the sum of the D_{SSIM} of each individual 4×4 block. This approximation allows the same λ to be used for both prediction mode decision and MB mode decision, resulting in a great simplification of the RDO process.

The traditional distortion metric allows the R-D model to be determined analytically, as indicated by (10) and (11). This nice feature, however, is lost when the traditional distortion metric is replaced by the SSIM index in the RDO framework. Specifically, the SSIM index cannot be related to the QP in a closed form, making it impossible to determine the R-D curve analytically. This is a fundamental issue of perceptual-based RDO. We propose a new approach to R-D modeling (Section IV). It allows us to determine, on the fly, an appropriate λ_{SSIM} in the encoding process.

IV. Predictive Lagrange Multiplier Estimation

We designed a novel predictive Lagrange multiplier estimation method for perceptual-based RDO. In this section, the observations that inspire this method are described first, followed by a detailed description of the proposed method.

A. Perceptual-Based RDO Versus MSE-Based RDO

As described in Section III, the SSD is replaced by the D_{SSIM} in the perceptual-based RDO framework. For simplicity, let D and λ denote D_{SSIM} and λ_{SSIM}, respectively, from now on unless it is otherwise stated. Likewise, when the term "R-D curve" is referred to from now on, it means that the quality of a compressed image is expressed by the SSIM index. For an R-D curve generated by the MSE-based RDO, this means that an additional step is applied upon the completion of the MSE-based RDO to measure the quality by the SSIM index.

We first compare the performance of the perceptual RDO framework with the MSE-based RDO framework. The result is shown in Fig. 1, which has five R-D curves. Each of the three short solid R-D curves represents a collection of R-D points corresponding to a range of manually selected λs for a given QP. These short R-D curves are obtained by perceptual-based RDO. The other two long dash R-D curves represent the

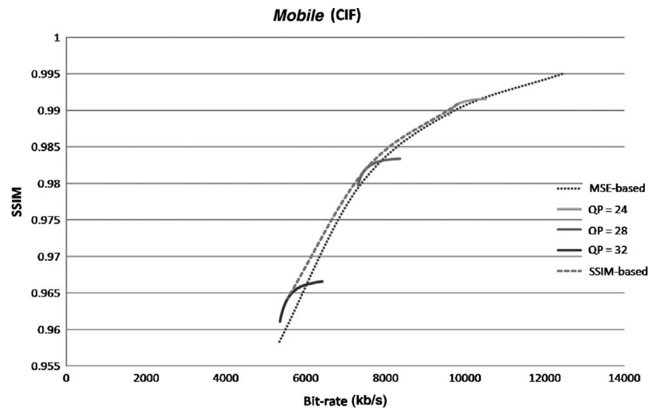


Fig. 1. Perceptual-based R-D curve versus the MSE-based R-D curve.

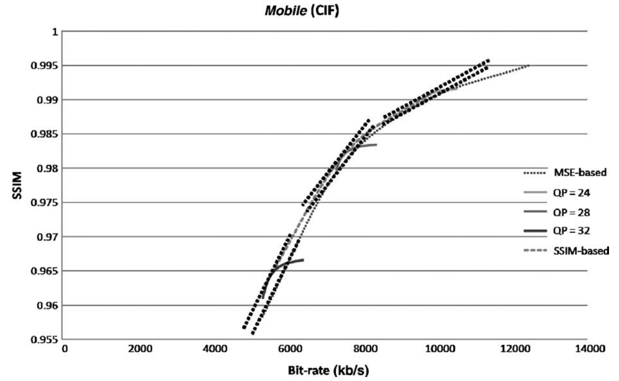


Fig. 2. Tangent lines of the optimal perceptual-based R-D curve and the optimal MSE-based R-D curve at various points.

collection of R-D points corresponding to various QPs. The upper long R-D curve is obtained by fitting an envelope to the three shorter R-D curves, and the lower one is obtained by traditional MSE-based RDO. We can see that the perceptual-based RDO is better than the MSE-based RDO when SSIM is used as quality metrics.

B. Observations

As discussed in Section I, the optimal λ corresponding to a point on the R-D curve can be obtained by computing the slope of the tangent to the R-D curve at that point. Fig. 2 shows several tangent pairs. In each tangent pair, there are two parallel tangents. One is the tangent to the perceptual-based R-D curve, and the other is the tangent to the MSE-based R-D curve at the point that is closest to the tangent point on the perceptual-based R-D curve. We can see that both tangents of a tangent pair have very similar slope. This is an important characteristic that can be exploited to relate the Lagrange multiplier of the perceptual-based RDO to that of the MSE-based RDO so that the former can be obtained from the latter.

Before doing that, we need to investigate a mathematical description of the MSE-based R-D curve. In [6], a linear function is used to describe the relation between the rate and the SSIM index. However, we find that power function leads to a better approximation of the relation. Fig. 3 shows the fitting of the R-D points obtained by the MSE-based RDO for four video sequences with a power function of the form $D = \alpha R^{\beta}$, where β is a negative number. As can be seen, the R-D points are well fitted by the power function. We use the R^2 [25] as a metric to evaluate the goodness of fitting. A higher R^2 value implies a better fitting, and the maximum R^2 value is 1, which occurs when the function perfectly fits all the data points. The R^2 statistics of the fitting results and the fitted coefficients α and β are given in Table I. It can be seen that the R^2 values of the fitting for different sequences are all close to 1. Therefore, we use the power function to approximate the MSE-based R-D curve in the perceptual R-D space. This is a crucial point of our approach.

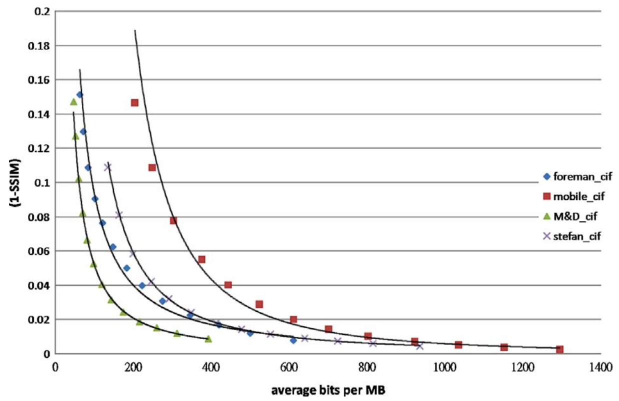


Fig. 3. R-D curve fitting for various test sequences using the power function $D = \alpha R^{\beta}$.

TABLE I
R-D Curve Fitting Using Power Function

Sequence (CIF)	*Foreman*	*Mobile*	*Mother and Daughter*	*Stefan*
α	27.10	20708.12	22.54	332.36
β	−1.229	−2.182	−1.319	−1.630
R^2	0.9854	0.9860	0.9993	0.9992

A closer look at Fig. 3 also indicates that the Lagrange multiplier varies with QP, as reported in [15] and [16], and the content since video sequences normally have diverse R-D characteristics. This is why we argue that the Lagrange multiplier estimation should be content-adaptive.

C. Proposed Method

We are now ready to describe the proposed Lagrange multiplier estimation method. First, an input frame is classified as a key frame or a non-key frame (Fig. 4). A key frame is encoded using the MSE-based RDO for two different QPs. This yields two distinct R-D points, (R_1, D_1) and (R_2, D_2). Then a curve fitting is applied to these two points to determine the two parameters α and β of the power function described in Section IV-B. These two parameters are directly computed as follows:

$$\alpha = D_1 \exp\left\{-\frac{\ln(R_1) \times \ln(D_2/D_1)}{\ln(R_2/R_1)}\right\} \tag{16}$$

$$\beta = \frac{\ln(D_2/D_1)}{\ln(R_2/R_1)}. \tag{17}$$

We call α and β the model parameters. Once the model parameters are obtained, the R-D model of the MSE-based

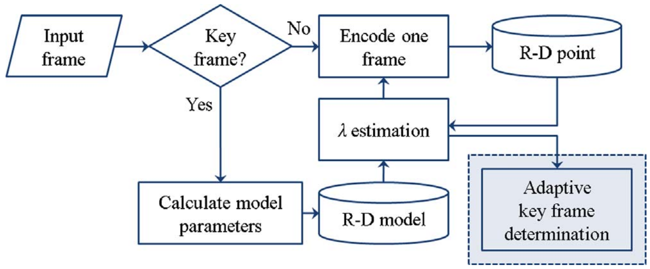


Fig. 4. Block diagram of video coding using the perceptual-based RDO framework.

R-D curve of a key frame is known. Since consecutive frames generally have high correlation and hence similar R-D characteristic, the R-D model of the key frame can be used as the predicted R-D model of the subsequent frames. Given the predicted R-D model of a frame, the λ value of the frame is obtained by calculating the slope of the tangent to the predicted R-D model at the point that is closest to the R-D point of the previous coded frame. The proposed method is predictive because the λ of each frame is obtained by using the R-D model of the previous coded key frame and the R-D point of the previous coded frame. Note that the first frame must be a key frame from which the initial R-D model is generated. Additional key frames can be assigned if needed.

D. Lagrange Multiplier Estimation by Gradient Descent

To determine the slope of the tangent to the R-D curve $\boldsymbol{C}$ of the current frame at the point p that is closest to the R-D point p_0 of the previous coded frame, we need to find the closest point p first. The process can be formulated as finding p such that

$$p = \arg\min_{q \in C}(d(p_0, q)) \tag{18}$$

where $d(\cdot, \cdot)$ denotes the two-norm distance between two points. Equation (18) is an optimization problem. Since the curve C can be well approximated by a power function (Section IV-B), $d(p_0, q)$ can be expressed as follows:

$$\begin{aligned} d(p_0, q) &= \sqrt{(R - R_{prev})^2 + (D - D_{prev})^2} \\ &= \sqrt{(R - R_{prev})^2 + (\alpha R^{\beta} - D_{prev})^2}. \end{aligned} \tag{19}$$

Because the objective function d (p_0, q) is convex, (18) can be solved by gradient descent. Once the point p is found, the Lagrange multiplier, which is the slope of the tangent at $p = (R_p, D_p)$, is readily obtained as follows:

$$\lambda_{SSIM} = -\frac{dD}{dR} = -\alpha\beta R^{\beta-1}\big|_{R=R_p}. \tag{20}$$

Since the initial point is required in the gradient descent algorithm, we choose it to be the point vertically projected to $\boldsymbol{C}$ from p_0.

E. Lagrange Multiplier Estimation by Slope Approximation

The gradient descent is an iterative approach, which is not preferable for many applications. Another approach to Lagrange multiplier estimation based on slope approximation is described here.

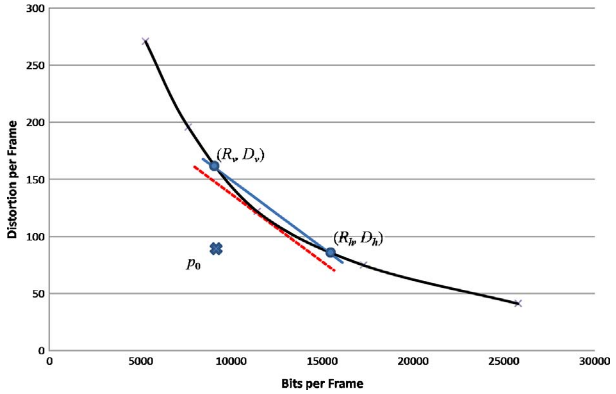


Fig. 5. Illustration of the slope of the tangent line and its approximation.

The basic idea of the approximation approach, as shown in Fig. 5, is that the line formed by the two points (R_h, D_h) and (R_v, D_v) projected from p_0 horizontally and vertically, respectively, to the R-D curve $\boldsymbol{C}$ is a good approximation of the slope of the tangent corresponding to the optimal λ (the dash line). Let $p_0 = (R_p, D_p)$ denote the R-D point of previous coded frame. Then, we have

$$R_v = R_p \tag{21}$$

$$D_v = \alpha R_p^{\beta} \tag{22}$$

$$R_h = (\frac{D_p}{\alpha})^{\frac{1}{\beta}} \tag{23}$$

$$D_h = D_p. \tag{24}$$

Therefore, we determine the Lagrange multiplier by

$$\lambda_{SSIM} = -\frac{D_v - D_h}{R_v - R_h}. \tag{25}$$

The advantage of this approach is simplicity. As opposed to the gradient descent approach, no iteration is required.

F. Periodic Refreshment

The predictive method for Lagrange multiplier estimation described above works well when the R-D characteristics of the first frame and the subsequent frames are similar. To handle sequences with varying R-D characteristics (due to, e.g., high motion or scene change) or to prevent error accumulation, we periodically refresh the R-D model by the key frame mechanism, similar to intra frame coding. There is a tradeoff between error accumulation and computational cost for periodic refreshment: the shorter the period, the higher the computational overhead. Therefore, the period should be appropriately selected according to the content. If the R-D characteristic of the video varies relatively slower, a longer period can be chosen, and vice versa. On the contrary, one may also apply conditional refreshment to the R-D model so that a new R-D model is set whenever it is necessary.

G. Inter-Frame Coding

In addition to the prediction and MB mode decisions of intra frame, the proposed method can be applied to MB mode decision of inter frames. For inter-frame coding, the Lagrange multiplier estimation method is the same as that of intra-frame coding except that the key frame is adaptively determined (shaded part in Fig. 4). Let λ_t be the weighted average of the λs of the first five frames after the key frame. If the relative change of λ to λ_t is larger than the threshold T_k, the next frame is set as a key frame. We leave T_k as a user-defined parameter.

V. Experimental Results

In this section, the experimental results for the proposed Lagrange multiplier estimation method are given. The algorithm is implemented in the JM reference software version 15.1 and run on a personal computer with Intel 2.13 GHz Pentium processor and 1.99 GB random access memory. The simulation conditions are listed as follows.

1) Baseline Profile is used.
2) All frames are intra coded.
3) RDO is enabled.
4) QP is set to 24, 28, 32, and 36.

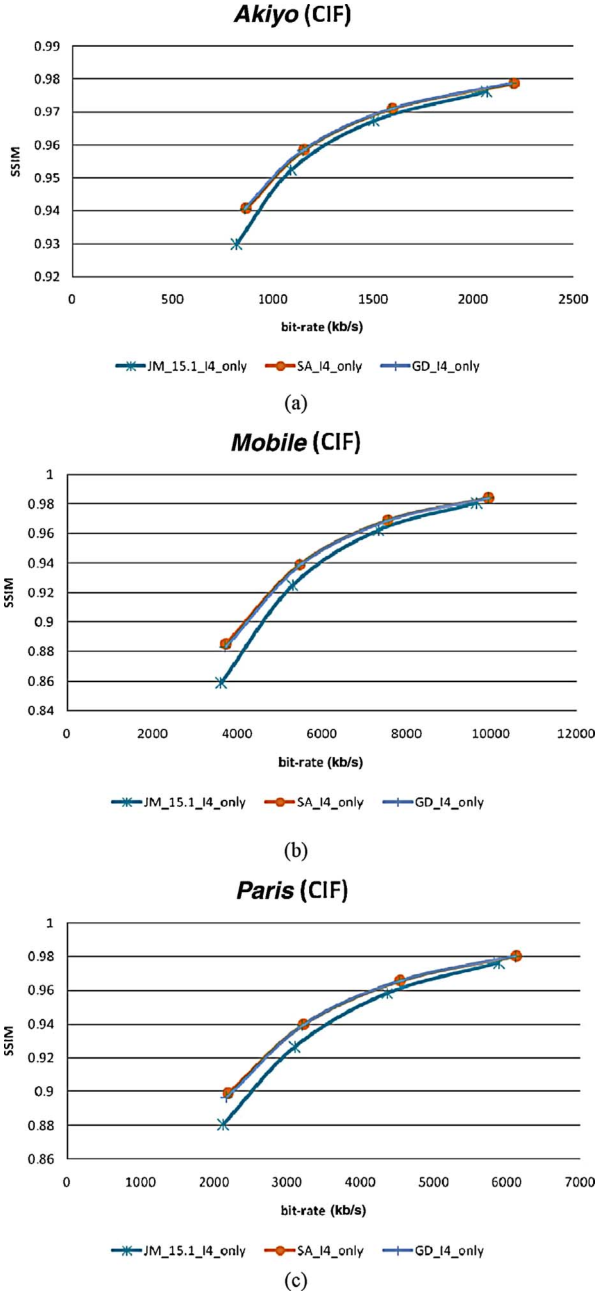
The performance of our method and the JM reference software is evaluated using R-D curves with the D measured by the SSIM index.

A. Prediction Mode Decision of I4 MB

The first experiment evaluates the performance improvement of perceptual-based RDO over the MSE-based RDO on the prediction mode decision of H.264 intra-frame coding. The I16 MB mode is disabled in the experiment (hence only I4 MB is enabled). The R-D curves generated by the two proposed approaches to Lagrange multiplier estimation, the gradient descent and the slope approximation, and the JM reference software for three test sequences are shown in Fig. 6, where SA_I4_only denotes the result of the slope approximation approach and GD_I4_only the result of the gradient descent approach. We can see that both approaches outperform the JM reference software (which uses MSE-based RDO) for every test sequence. In addition, the slope approximation approach has almost the same R-D performance as that of the gradient descent approach, strongly suggesting that this simplified, non-iterative algorithm can replace the gradient descent approach for practical purposes. A close look at Fig. 6 further indicates that the perceptual-based RDO has more performance gain at low bit-rates and that the gain is significant. The fact that SSIM image quality metric significantly outperforms MSE especially for highly compressed images is inspiring for low bit-rate video coding. The bit-rate reduction of our perceptual-based RDO algorithm for the mode decision of H.264 intra-frame coding is summarized in Table II. On the average, about 9% bit-rate reduction over the JM reference software is achieved at the same perceptual quality level for the case where only the I4 MB mode is used for mode prediction. Here, the bit-rate reduction is calculated using the method described in [26].

B. MB Mode Decision of Intra MB

This experiment tests the performance of the proposed perceptual-based RDO framework on the MB mode decision of H.264 intra-frame coding. The setting of the experiment is pretty much the same as that of the first experiment described above except that the I16 MB mode is turned on instead. The resulting R-D curves of two test sequences are shown in Fig. 7, where *SA* denotes the result of the slope approximation approach. Note that for comparison purpose, the JM_15.1_I4_only curve and the SA_I4_only curve are copied from Fig. 6. We can see that the perceptual-based RDO outperforms the MSE-based RDO (used in the JM reference software) again in this experiment, especially in the low bit-rate range. The bit-rate reduction of the perceptual RDO over the JM reference software at the same perceptual quality level is shown in the right column of Table II. An average of 9.6% gain is achieved. By comparing the SA curve with the SA_I4_only curve in Fig. 7, we can also see that the



(a)

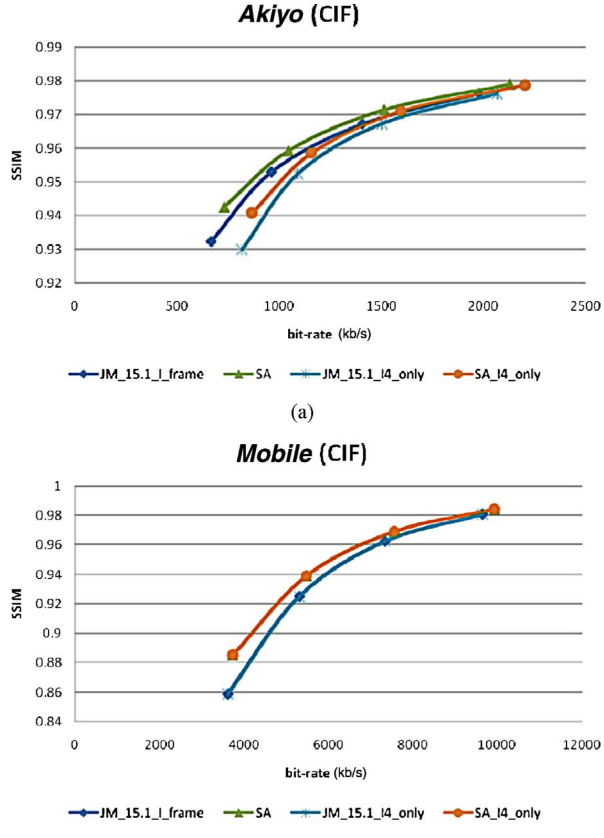
(b)

(c)

Fig. 6. R-D curves of (a) *Akiyo*, (b) *Mobile*, and (c) *Paris* sequences using only the I4 MB mode.

TABLE II
BIT-RATE REDUCTION (%) OF THE PROPOSED METHOD

Sequence	14 MB Only	Intra Frame
Akiyo (CIF)	9.74	11.65
Mobile (CIF)	11.93	11.90
Mother and Daughter (CIF)	10.28	11.90
Stefan (CIF)	6.12	6.15
Paris (CIF)	13.27	13.03
Character (D1)	10.59	10.23
Mobile (D1)	9.21	9.09
Night (D1)	7.80	7.76
Preakness (720p)	7.29	7.29
Spincalendar (720p)	7.77	7.63
Blue_sky (1080p)	10.13	10.71
Toys_and_calendar (1080p)	7.68	8.25
Average	9.32	9.63



(a)

(b)

Fig. 7. Evaluation of the inclusion of I16 MB for (a) *Akiyo* and (b) *Mobile* sequences.

inclusion of I16 MB mode leads to a higher coding gain for the *Akiyo* sequence (which has more homogenous regions) but not for the *Mobile* sequence (which has more texture regions), indicating that the outcome of MB mode decision in reference to texture complexity still holds well under the perceptual-based RDO framework.

C. Evaluation of the Estimated λ

Fig. 8 shows the λ estimated by our proposed Lagrange multiplier estimation method. Here only the values estimated by the slope approximation approach are shown since the result of the gradient descent approach is very similar to that of the slope approximation approach. We can see that the estimated λ varies with test sequences. Compared to the fixed λ-QP curve adopted in the MSE-based RDO for all sequences [15], what we just observed is an indication that the tradeoff between rate and distortion is content-dependent. Under the same QP, the sequences which require more bits to encode use smaller λs because the quality of these sequences can be improved with a relatively small percentage of bit-rate increase, resulting in better R-D performance. On the contrary, larger λs are chosen for sequences that are easy to be compressed so that we can save more percentage of bits at the cost of a relatively small increase of distortion. Similar relation is also observed in [15]. Note that the diversity of the estimated λ for various sequences decreases with QP. This is consistent with the well-known result that, under the high rate assumption, the Lagrange multiplier can be simply a function of QP. Therefore, our method is more adaptive to content than the existing methods that determine λ according to QP only.

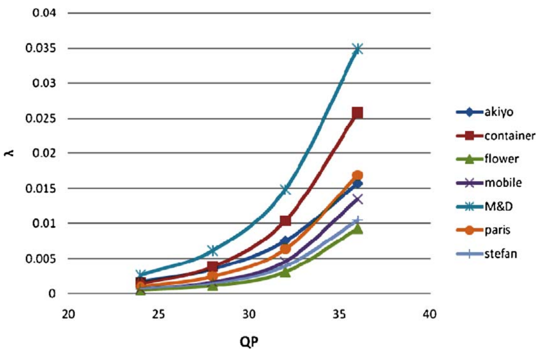


Fig. 8. Estimated λs of various test sequences.

D. Evaluation of Periodic Refreshment

In this experiment, the impact of the periodic refreshment on the overall R-D performance is discussed. The period of refreshment is set to 30 frames. Due to the same reason as described earlier, we only show the result for the slope approximation approach. As we can see from Fig. 9, the periodic refreshment achieves R-D performance improvement for sequences with frequent scene changes (e.g., the *Football* sequence) but not for sequences with high temporal correlation (the *Akiyo* sequence), because the R-D characteristic of the first frame well represents that of the subsequent frames for such sequences. However, the computational complexity incurred by periodic refreshment is relatively high since additional encoding pass is required periodically. Therefore, we suggest that this option be turned on only for sequences with time-varying R-D characteristic. Additional discussion of the computational complexity issue is given in Section V-E.

E. Complexity Overhead

In this experiment, the computational complexity overhead of the proposed method is evaluated. First, we compare the overhead per frame introduced by different λ estimation approaches with the JM reference software. As shown in Table III, the overhead of the slope approximation is about

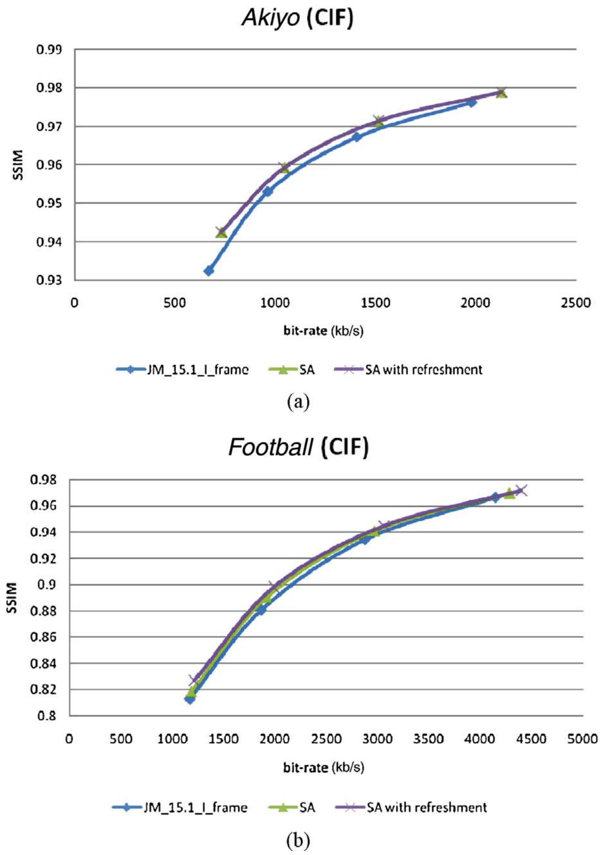


Fig. 9. Evaluation of periodic refreshment for (a) *Akiyo* and (b) *Football* sequences.

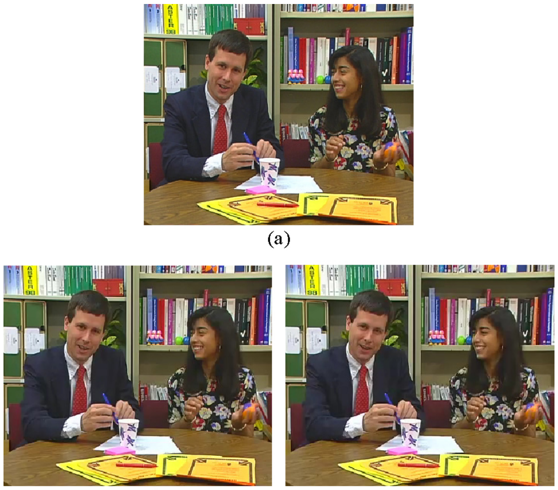


Fig. 10. 180-th frame of *Paris* sequence (CIF) coded using QP = 28. (a) Original image. (b) Reconstructed image using JM. (c) Reconstructed image using the proposed method. A bit-rate reduction of 10.23% is achieved by the proposed approach. Images shown here are resized to fit within the space.

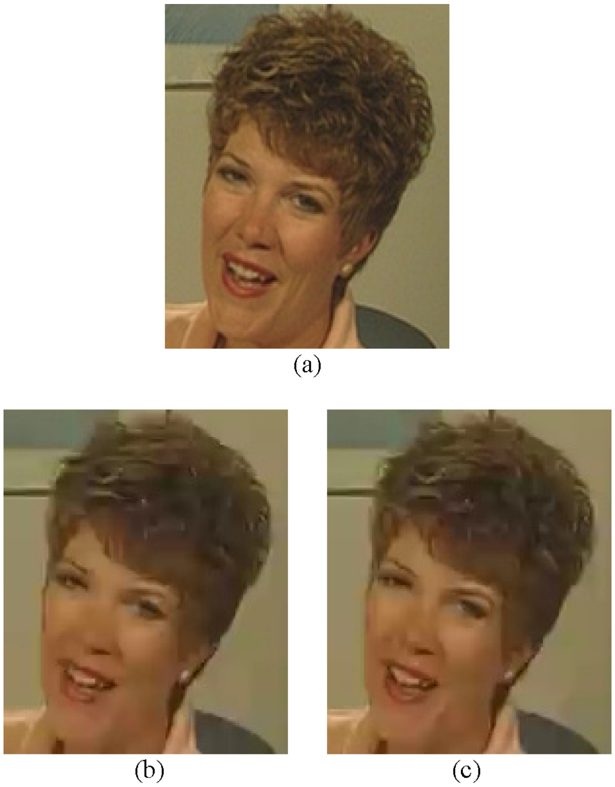


Fig. 11. Cropped image of the 213-th frame of *Mother and Daughter* sequence (CIF) coded using QP = 36. (a) Original image. (b) Reconstructed image using JM. (c) Reconstructed image using the proposed method.

TABLE III

Overhead (%) of Lagrange Multiplier Estimation Approaches

Sequence (CIF)	Gradient Descent	Slope Approximation
Flower	8.46	5.36
Mobile	12.64	5.82
Stefan	8.07	5.01
Mother and Daughter	4.71	4.41
Weather	3.58	3.49
Average	7.49	4.82

5%, almost all of which is due to the SSIM index computation. We can also see that the overhead of the slope approximation approach is lower and more stable for various sequences than the iterative gradient descent approach. We find that the number of iterations required for the iterative gradient descent approach is related to the temporal correlation of a sequence. For sequences with low temporal correlation, such as *Mobile* and *Stefan*, the R-D characteristics vary from frame to frame; the R-D curve of the current frame can be much different from that of the reference key frame. Hence, a poor initial point can take more iterations to converge.

Next, we discuss the computational overhead of the additional encoding passes required for different modes. The data are shown in Table IV, where N denotes the total number of encoding frames and T the period of refreshment. Without refreshment, only the first frame requires additional two encoding passes (see Section IV-C). Therefore, the overhead is fixed regardless of N. As N gets larger, the overhead percentage becomes smaller. For the periodic refreshment mode, the overhead depends on N and T. Specifically, it is proportional to N and inversely proportional to T. Therefore, the period should be carefully chosen.

F. Subjective Test

Subjective quality evaluation is performed to compare our method with the JM. In all cases, our method encodes the

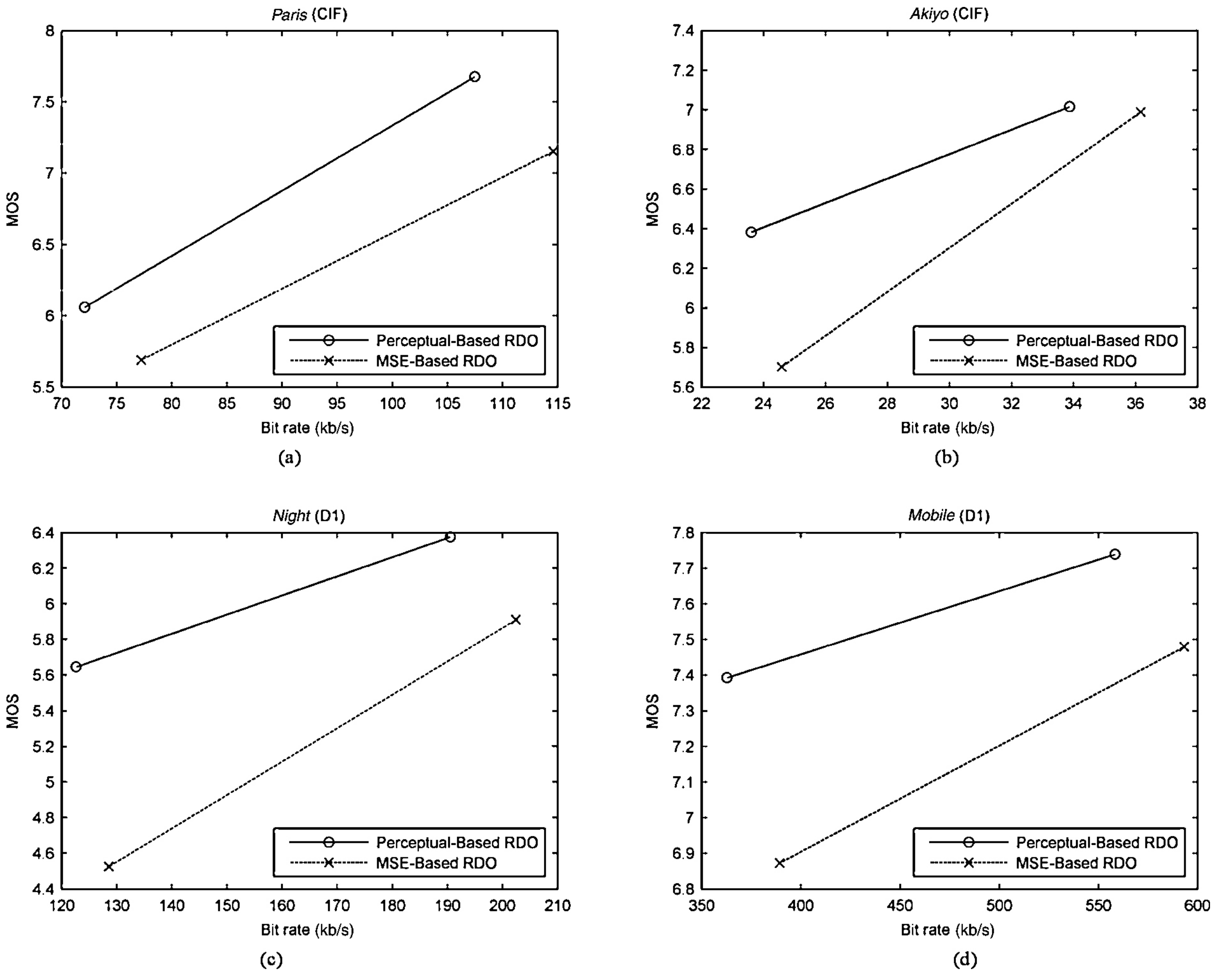


Fig. 12. Subjective performance comparison of perceptual-based RDO and MSE-based RDO for (a) *Paris* (CIF), (b) *Akiyo* (CIF), (c) *Night* (D1), and (d) *Mobile* (D1) sequences.

TABLE IV
Overhead Comparison

Method	No Refreshment	Periodic Refreshment	Low Complexity
Overhead (frames)	2	$2 \times \frac{N}{T}$	0

video at a lower bit-rate than the JM method. As the difference in video quality is too difficult to tell for the subjects when the video sequence is played at 30 f/s, we perform subjective evaluation of the video quality in a frame-by-frame manner. The first comparison is made in the middle bit-rate range by setting QP to 28 for our method. The QP for the JM is adjusted until the SSIM index of the coded sequence is close enough to that obtained by our method. Fig. 10 shows the reconstructed image using different methods. The image encoded by our method is indistinguishable from that encoded by the JM. However, our method achieves about 10% bit-rate reduction over the JM. On the contrary, as shown in Fig. 11, the perceptual quality of the image encoded by our method is better than that by JM at the same low bit-rate level. Specifically, our approach has less block artifact and preserves more structural information (check the hair, eyebrows, and lips of the mother in the image) since it takes the structural information into account and avoid selecting prediction modes that introduce the structural distortion to the reconstructed image. More images for subjective evaluation can be found in [27].

We also conduct a subjective evaluation based on the double-stimulus continuous quality-scale method [28]. A test session comprises a number of presentations. Each presentation shows a pair of frames encoded by our method and by the JM in random order. An assessor is free to switch between these two frames until the assessor has the mental measure of the quality associated with each frame. Since the bit-rates of the two frames in a presentation cannot be precisely controlled to be the same, we set QP of our method larger than that of the JM by one to ensure that our method does not consume more bits than the JM. Specifically, the QP value is set to: 1) 32 for our method and 31 for the JM, and 2) 36 for our method and 35 for the JM in the experiments. We encoded five test sequences with the two QP settings. For each QP setting, the frame with the most SSIM gain and the frame with the most bit-rate reduction among all frames of a sequence are selected as the samples frames to be evaluated subjectively. It should be noted that although our method uses less bits under each QP setting, it still achieves a higher SSIM value than the JM for almost all frames. There are total 23 presentations. The first

TABLE V

SPECIFICATION OF THE LCD MONITOR USED IN SUBJECTIVE EVALUATION

Display area	20.5" horiztonal × 11.5" vertical, 24" diagonal
Resolution	1920 × 1080
Contrast ratio	1000:1
Brightness	300 cd/m^2
Pixel pitch	0.2715 mm

TABLE VI

BIT-RATE REDUCTION (%) FOR INTER-FRAME CODING UNDER THE BASELINE PROFILE

Sequence	QP = 16−28	QP = 24−36
Stefan (CIF)	12.93	10.90
Flower (CIF)	11.85	15.01
Akiyo (CIF)	20.33	14.74
Mobile (CIF)	4.10	14.86
Weather (CIF)	15.90	13.21
Bridge-Close (CIF)	17.40	19.59
Bus (CIF)	6.86	11.21
Coastguard (CIF)	2.26	5.49
Container (CIF)	23.59	29.51
Hall_Monitor (CIF)	26.46	21.12
Paris (CIF)	11.30	13.13
MadCyclistL (CIF)	1.00	7.08
Mother and Daughter (CIF)	2.72	2.94
Silent (CIF)	4.79	5.52
Table (CIF)	7.48	5.89
Tempete (CIF)	4.21	9.88
Bus (D1)	3.21	11.12
Character (D1)	3.82	2.84
Mobile (D1)	6.20	16.60
Mobcal (720p)	13.91	13.38
Shields (720p)	11.50	7.01
Jets (720p)	12.48	3.38
Panslow (720p)	16.92	16.41
Spincalendar (720p)	12.82	12.76
Average	10.59	11.82

three presentations are dummy ones to stabilize the assessor's opinion. The remaining 20 presentations show the selected sample frames in random order. Table V lists the specification of the LCD monitors used in the test. The quality grade is recorded on the continuous rating scale, as suggested in [28]. The final opinion score is obtained by linearly mapping the recorded grade to the value between 0 and 10. Fig. 12 shows the mean opinion score (MOS) of 15 subjects. It can be seen that our method performs consistently better than the JM for different sequences.

G. MB Mode Decision of Inter MB

In this experiment, we evaluate the R-D performance of the perceptual-based RDO for H.264 inter-frame coding. The simulation conditions are as follows.

1) GOP structure is IPPP.
2) RDO is enabled.
3) QP is set to 16, 20, 24, 28, 32, and 36.
4) T_k is set to 0.5.

TABLE VII

BIT-RATE REDUCTION (%) FOR INTER-FRAME CODING UNDER THE HIGH PROFILE

Sequence (CIF)	QP = 16−28	QP = 24−36
Stefan	14.60	11.13
Flower	12.45	17.18
Akiyo	18.19	12.95
Mobile	6.15	14.76
Weather	13.57	13.06
Bridge-Close	16.26	17.99
Bus	6.49	12.28
Coastguard	1.23	4.46
Container	21.19	28.62
Hall_Monitor	24.81	27.69
MadCyclistL	1.91	8.95
Mother and Daughter	2.91	6.00
Paris	10.55	12.65
Silent	5.06	6.13
Table	8.54	6.38
Tempete	4.34	11.23
Average	10.52	13.22

Tables VI and VII, respectively, show the bit-rate reduction of the perceptual-based RDO over the MSE-based RDO for the Baseline Profile and the High Profile. We can see that, at the same SSIM index, more than 10% bit-rate reduction is achieved for both profiles. This demonstrates that the proposed perceptual-based RDO works well for both inter-frame coding and intra-frame coding.

VI. CONCLUSION

The distortion metric used for measuring video quality in the RDO process has a profound impact on video coding. Typically, objective metrics not well correlated with perceptual video quality were used for video coding. To set the ground for perceptual-based RDO of video coding, in this paper, we introduced an RDO using the SSIM as the quality metric and solved the optimization problem by the Lagrange technique. But because the R-D characteristic of the video content cannot be obtained until the video is encoded, and vice versa, the Lagrange multiplier cannot be directly determined. To resolve this dilemma, we developed a novel predictive Lagrange multiplier estimation approach and evaluated various methods for solving the Lagrange multiplier.

To evaluate the performance gain of the proposed perceptual-based RDO over the MSE-based RDO, we applied both frameworks to H.264 intra-frame and inter-frame coding, for which the perceptual-based RDO uses the R-D characteristics of a key frame to predict the R-D characteristics of the subsequent frames till the next key frame appears. The experimental results showed that, at the same SSIM value, the proposed approach achieves on the average 9% bit-rate reduction for intra-frame coding and 11% for inter-frame coding over the MSE-based RDO framework. Furthermore, the subjective test showed that the proposed perceptual-based RDO preserves edge and avoids block artifact better than the MSE-based RDO.

In our proposed estimation scheme, different video contents and QPs lead to different Lagrange multiplier values. So it can cope with the diversity of perceptual R-D characteristics of various video sequences. This feature made our method content-adaptive, which is shown to be very important for SSIM-based RDO.

References

[1] *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification*. ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC, Document JVT-G050.doc, Pattaya, Thailand, 2003.

[2] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of H.264 video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[4] Z. Y. Mai, C. L. Yang, K. Z. Kuang, and L. M. Po, "A novel motion estimation method based on structural similarity for H.264 inter prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2. May 2006, pp. 913–916.

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

[6] L. Teixeira and L. Corte-Real, "H.264 rate-distortion analysis using subjective quality metric," in *Proc. 2nd Int. Workshop Future Multimedia Netw.*, 2009, pp. 248–253.

[7] C.-J. Tsai, C.-W. Tang, C.-H. Chen, and Y.-H. Yu, "Adaptive rate-distortion optimization using perceptual hints," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 667–670.

[8] C. Sun, H.-J. Wang, T.-H. Kim, and H. Li, "Perceptually adaptive Lagrange multiplier for rate-distortion optimization in H.264," in *Proc. Future Generat. Commun. Netw.*, Dec. 2007, pp. 459–463.

[9] T.-S. Ou, Y.-H. Huang, and H. H. Chen, "A perceptual-based approach to bit allocation for H.264 encoder," in *Proc. SPIE Vis. Commun. Image Process.*, vol. 7744. Jul. 2010, pp. 1–10.

[10] P.-Y. Su, Y.-H. Huang, T.-S. Ou, and H. H. Chen, "Predictive Lagrange multiplier selection for perceptual-based rate-distortion optimization," in *Proc. 5th Int. Workshop Video Process. Qual. Metrics Consumer Electron.*, Jan. 2010.

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

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

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

[14] 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.

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

[16] 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.

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

[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] G.-H. Chen, C.-L. Yang, and S.-L. Xie, "Gradient-based structural similarity for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 2929–2932.

[20] B. Wang, Z. Wang, Y. Liao, and X. Lin, "HVS-based structural similarity for image quality assessment," in *Proc. Int. Conf. Signal Process.*, Oct. 2008, pp. 1194–1197.

[21] C. Li and A. C. Bovik, "Three-component weighted structural similarity index," in *Proc. SPIE Conf. Image Quality Syst. Performance*, Jan. 2009.

[22] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, Jun. 2008.

[23] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1196–1199.

[24] *JVT Reference Software* [Online]. Available: http://bs.hhi.de/∼suehring/tml/download

[25] J. L. Devore and N. R. Farnum, *Applied Statistics for Engineers and Scientists*. New York: Duxbury, 1999.

[26] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *Proc. 13th VCEG Meeting ITU-T Q.6/SG16*, Document VCEG-M33.doc., Austin, TX, Apr. 2001.

[27] *Performance Evaluation of Perceptual-Based RDO* [Online]. Available: http://mpac.ee.ntu.edu.tw/∼odeson/projects/ssimrdo

[28] ITU-R, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Rec. BT.500-11, 2002.

Yi-Hsin Huang received the B.S. degree in electrical engineering and the M.S. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2007 and 2009, respectively.

He is currently with MediaTek, Hsinchu, Taiwan. His current research interests include H.264/AVC video coding and perceptual-based image processing.

Tao-Sheng Ou received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2008. He is currently pursuing the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University.

His current research interests include H.264/AVC video coding and image processing.

Po-Yen Su received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2010. He is currently pursuing the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University.

His current research interests include H.264/AVC video coding and perceptual-based video processing.

Homer H. Chen (S'83–M'86–SM'01–F'03) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana.

Since 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, Taipei, Taiwan, where he is Irving T. Ho Chair Professor. Prior to that, he held various research and development management and engineering positions with U.S. companies over a period of 17 years, including AT/&T Bell Laboratories, Holmdel, NJ, Rockwell Science Center, Thousand Oaks, CA, iVast, Santa Clara, CA, and Digital Island, Thousand Oaks, CA. He was a U.S. delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. His professional interests lie in the broad area of multimedia signal processing and communications.

Dr. Chen is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology. He was an Associate Editor of IEEE Transactions on Image Processing from 1992 to 1994, a Guest Editor of IEEE Transactions on Circuits and Systems for Video Technology in 1999, and an Associate Editorial Member of *Pattern Recognition* from 1989 to 1999.