Joint Source-Channel Rate-Distortion Optimization for H.264 Video Coding Over Error-Prone Networks

Yuan Zhang, Wen Gao, Yan Lu, Qingming Huang, and Debin Zhao

***Abstract*—For a typical video distribution system, the video contents are first compressed and then stored in the local storage or transmitted to the end users through networks. While the compressed videos are transmitted through error-prone networks, error robustness becomes an important issue. In the past years, a number of rate-distortion (R-D) optimized coding mode selection schemes have been proposed for error-resilient video coding, including a recursive optimal per-pixel estimate (ROPE) method. However, the ROPE-related approaches assume integer-pixel motion-compensated prediction rather than subpixel prediction, whose extension to H.264 is not straightforward. Alternatively, an error-robust R-D optimization (ER-RDO) method has been included in H.264 test model, in which the estimate of pixel distortion is derived by simulating decoding process multiple times in the encoder. Obviously, the computing complexity is very high. To address this problem, we propose a new end-to-end distortion model for R-D optimized coding mode selection, in which the overall distortion is taken as the sum of several separable distortion items. Thus, it can suppress the approximation errors caused by pixel averaging operations such as subpixel prediction. Based on the proposed end-to-end distortion model, a new Lagrange multiplier is derived for R-D optimized coding mode selection in packet-loss environment by taking into account of the network conditions. The rate control and complexity issues are also discussed in this paper.**

***Index Terms*—Error resilience, H.264/MPEG-4 AVC, rate distortion optimization, video coding.**

I. Introduction

THE transmission of compressed video over the existing packet-switched networks presents many new challenges due to the problems caused by packet loss. As we know, most video coding standards are based on a hybrid coding method, which uses transform coding with motion-compensated prediction (MCP). In the packet-loss environment, transmitting the hybrid-coded video may suffer from error propagations and lead to the well-known drifting phenomenon [1]. These challenges have inspired several feasible solutions. One category of solutions focuses on the link-layer reliability, e.g., forward error correction (FEC) and/or automatic repeat request (ARQ). Another category of solutions is based on the error control strategy in source coding, e.g., error-resilient video coding [2]. Sometimes, error-resilient coding tools devised in the encoder and error concealment tools devised in the decoder are jointly employed in the transmission of hybrid-coded video over error-prone networks.

Adaptive intra/inter coding mode selection is a typical type of error-resilient video coding techniques [3]–[12]. In standard-compliant techniques, intra coding can suppress the error propagation at the cost of reduced coding efficiency. In other words, inserting more intra-coded macroblocks in the encoder can make the bitstream more resilient to potential errors, and meanwhile the bit rate is increased at the same visual quality. Therefore, one problem that has to be addressed is how to achieve the best quality of services while considering both the coding efficiency and the suppression of potential errors. The early intra refreshment algorithms have been developed to randomly insert intra blocks [3] or periodically intra-code contiguous blocks [4]. The intra refresh frequency is determined in a heuristic way. Later, the content-adaptive coding mode selection scheme has been proposed to intra-code macroblocks at regions with high activity [5]. The common disadvantage of these techniques is that they cannot always achieve satisfactory performances because of not taking into account of network condition and error concealment.

Alternatively, several rate-distortion (R-D) optimized techniques have been proposed for coding mode selection in error-prone environment [6]–[12]. In [8], a generalized end-to-end approach has been proposed for video communication over packet-switched networks, in which a set of global distortion metrics was derived for the first time. However, the distortion model derived in terms of mean absolute difference (MAD) is not suitable for R-D optimized mode selection. In [11], for the first time, a generalized framework for joint rate control and error control is proposed. However, it determines the intra refresh rate prior to the coding of each frame. The frame-level global optimization cannot consider some superior error concealment methods during the encoding process. In [12], a recursive optimal per-pixel estimate (ROPE) algorithm has been proposed to estimate the end-to-end distortion at pixel level by keeping track of the first and second moments of the reconstructed pixel value. However, ROPE assumes integer-pixel MCP rather than subpixel prediction, because it requires intensive computing and storage when pixel averaging operations (e.g., interpolation operations in subpixel prediction) involve [13].

Manuscript received June 23, 2005; revised July 20, 2006. This work was supported in part by the National Science Foundation of China (60333020) and opening project of Beijing Key Laboratory of Multimedia and Intelligent Software, Beijing University of Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chang Wen Chen.

Y. Zhang and Q. Huang are with the Graduate School, Chinese Academy of Sciences, Beijing 100080, China (e-mail: yzhang@jdl.ac.cn; qmhuang@jdl.ac.cn).

W. Gao is with the School of Electronic Engineering and Computer Science, Peking University, Beijing 100080, China (e-mail: wgao@jdl.ac.cn).

Y. Lu is with the Microsoft Research Asia, Beijing 100080, China (e-mail: yanlu@microsoft.com).

D. Zhao is with the Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China (e-mail: dbzhao@jdl.ac.cn).

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

Digital Object Identifier 10.1109/TMM.2006.887989

More recently, an error robust rate distortion optimization method, referred to as ER-RDO, has been developed for video coding in packet-loss environment [14], [15], which has been adopted in the H.264/AVC test model [16], [17]. ER-RDO estimates the expected overall end-to-end distortion in a manner of independently operating K copies of the random variable channel behavior and decoder pairs in the encoder. The expected decoder distortion can be estimated very accurately if K is chosen large enough. As shown in [14], ER-RDO compared to ROPE in general has lower bit rate and lower average PSNR for the same quantization parameter, but has higher overall R-D performance. However, ER-RDO also suffers from the main drawback that it requires high computational complexity and implementation cost, which makes it unsuitable for many practical applications. Therefore, it is desirable to develop a new end-to-end distortion estimation scheme that has both low complexity and high R-D performance.

Toward this goal, we first propose a concise and efficient end-to-end distortion model. In the proposed distortion model, the overall distortion is fine categorized into source, error-propagated and error-concealment *distortion* items. The basic idea of combining source and channel distortions has been presented in [11]. However, the proposed distortion model associates all distortion items with the error rate in theory, which is effective especially when the error rate is large. Moreover, each distortion item in the proposed model is further separated into several small *distortion* items that can be calculated either directly or recursively. The recursive calculation can trace the error propagation from all previous frames, which has also been used in [8] and [12]. However, ROPE in [12] recursively calculates the first and second moments of the reconstructed pixel value, which is very sensitive to the approximation errors caused by subpixel MCP and other pixel averaging operations [13]. Distinctively, the overall distortion in the proposed model is taken as the sum of several *distortion* items, which can suppress the approximation errors from subpixel MCP. The proposed model can be easily extended to the block-level implementation due to its robustness against the approximation errors involved in the pixel averaging operations.

Based on the above end-to-end distortion model, we further propose an R-D optimized mode selection scheme for error-resilient video coding. Considering the general utilization of Lagrange method in R-D optimization, we derive a new Lagrange multiplier for error-resilient video coding. Intuitively, the Lagrange multiplier should be related to the channel conditions. In [10], the Lagrange multiplier in packet-loss environment is taken as the multiplier in error-free environment added by a delta value. However, the delta value cannot be theoretically derived. To the best of our knowledge, the Lagrange multiplier has not been accurately derived before due to the lack of a proper distortion model. Instead, the proposed distortion model composed of a couple of distortion items can reveal the true R-D relationship in packet-loss environment. Moreover, we further discuss the rate control issue related to the coding mode selection. In particular, we employ the one-pass macroblock-level rate control scheme derived in [18]. The complexity issues related to the proposed error-resilient video coding are also discussed.

The rest of this paper is organized as follows. In Section II, the proposed end-to-end distortion estimation algorithm including the practical block-level implementation is presented in detail. In Section III, the R-D optimized coding mode selection scheme is presented, including the derivation of the optimized Lagrange multiplier in packet-loss environment and the discussion of rate control. Section IV shows the simulation results. Finally, Section V concludes this paper.

II. End-to-End Distortion Estimation

Above all, we define some notations used in the derivation of the proposed end-to-end distortion model. For pixel i in frame n that references pixel j in frame ref, let f_n^i be the original value, and let $\hat{f}_n^i$ and $\widetilde{f}_n^i$ be the reconstructed values in the encoder and decoder, respectively. Let $\hat{r}_n^i$ be the reconstructed residue in the encoder, i.e., $\hat{f}_n^i = \hat{f}_{ref}^j + \hat{r}_n^i$. When the current pixel is lost in the decoder, it copies from pixel k in frame $n-1$. Suppose the transmission error rate is known as p. Then, we can represent $\widetilde{f}_n^i$ as

$$\widetilde{f}_n^i = \begin{cases} \widetilde{f}_{ref}^j + \hat{r}_n^i & w.p. \quad 1-p \\ \widetilde{f}_{n-1}^k & w.p. \quad p. \end{cases} \tag{1}$$

Hence, we can derive the expectation of end-to-end distortion in the decoder to be

$$\begin{aligned} d(n,i) &= E\left\{\left(f_n^i - \widetilde{f}_n^i\right)^2\right\} \\ &= (1-p)E\left\{\left(f_n^i - \left(\widetilde{f}_{ref}^j + \hat{r}_n^i\right)\right)^2\right\} \\ &\quad + pE\left\{\left(f_n^i - \widetilde{f}_{n-1}^k\right)^2\right\} \\ &= (1-p)E\left\{\left(f_n^i - \hat{f}_n^i\right)^2\right\} \\ &\quad + (1-p)E\left\{\left(\hat{f}_{ref}^j - \widetilde{f}_{ref}^j\right)^2\right\} \\ &\quad + pE\left\{\left(f_n^i - \widetilde{f}_{n-1}^k\right)^2\right\} \\ &= (1-p)d_s(n,i) + (1-p)d_{ep}(ref,j) \\ &\quad + pd_{ec}(n,i) \end{aligned} \tag{2}$$

where $d_s(n,i)$ denotes the source distortion, $d_{ep}(ref,j)$ denotes the error-propagated distortion from the reference frame, and $d_{ec}(n,i)$ denotes the error-concealment distortion. The third equality in (2) bases on the assumption that effects of source distortion in the encoder and error-propagated distortion in the decoder are additive.

Since $d_s(n,i)$ can be readily calculated, the estimation of $d(n,i)$ in the encoder mainly relies on the calculation of $d_{ep}(ref,j)$ and $d_{ec}(n,i)$. Firstly, we derive the formula to calculate $d_{ec}(n,i)$ as

$$\begin{aligned} d_{ec}(n,i) &= E\left\{\left(f_n^i - \widetilde{f}_{n-1}^k\right)^2\right\} \\ &= E\left\{\left(f_n^i - \hat{f}_{n-1}^k + \hat{f}_{n-1}^k - \widetilde{f}_{n-1}^k\right)^2\right\} \\ &= E\left\{\left(f_n^i - \hat{f}_{n-1}^k\right)^2\right\} + E\left\{\left(\hat{f}_{n-1}^k - \widetilde{f}_{n-1}^k\right)^2\right\} \\ &= d_{ec\_o}(n,i) + d_{ep}(n-1,k) \end{aligned} \tag{3}$$

where $d_{ec\_o}(n,i)$ indicates the mean square error (MSE) between the original and error-concealment pixel values in the encoder, namely, the original error-concealment distortion. The third equality in (3) bases on the similar assumption that the effects of original error-concealment distortion in the encoder and error-propagated distortion in the decoder are additive. $d_{ec\_o}(n,i)$ can also be readily calculated. The remained problem is how to calculate $d_{ep}(n-1,k)$.

Note that $d_{ep}(n-1,k)$ in (3) and $d_{ep}(ref,j)$ in (2) are in the similar style. Without losing the generality, we derive the formula to calculate $d_{ep}(n,i)$ as

$$\begin{aligned} d_{ep}(n,i) &= E\left\{\left(\hat{f}_n^i - \tilde{f}_n^i\right)^2\right\} \\ &= (1-p)E\left\{\left(\hat{f}_n^i - \left(\tilde{f}_{ref}^j + \hat{r}_n^i\right)\right)^2\right\} \\ &\quad + pE\left\{\left(\hat{f}_n^i - \tilde{f}_{n-1}^k\right)^2\right\} \\ &= (1-p)E\left\{\left(\hat{f}_{ref}^j - \tilde{f}_{ref}^j\right)^2\right\} \\ &\quad + pE\left\{\left(\hat{f}_n^i - \hat{f}_{n-1}^k + \hat{f}_{n-1}^k - \tilde{f}_{n-1}^k\right)^2\right\} \\ &= (1-p)E\left\{\left(\hat{f}_{ref}^j - \tilde{f}_{ref}^j\right)^2\right\} \\ &\quad + pE\left\{\left(\hat{f}_n^i - \hat{f}_{n-1}^k\right)^2\right\} \\ &\quad + pE\left\{\left(\hat{f}_{n-1}^k - \tilde{f}_{n-1}^k\right)^2\right\} \\ &= (1-p)d_{ep}(ref,j) + pd_{ec\_r}(n,i) \\ &\quad + pd_{ep}(n-1,k) \end{aligned} \tag{4}$$

where $d_{ec\_r}(n,i)$ indicates the MSE between the reconstructed and error-concealment values in the encoder, namely, the reconstructed error-concealment distortion. Since $d_{ec\_r}(n,i)$ can also be readily calculated, the calculation of $d_{ep}(n,i)$ only depends on the availability of the error-propagated distortions from its previous frames. Note that d_{ep} of the first frame can be directly derived without considering the error propagation because it is typically coded as an intra frame. Hence d_{ep} of the following frames can also be recursively calculated frame by frame.

In the proposed end-to-end distortion model, the end-to-end distortion of the current frame is first calculated by referencing the error-propagated distortions of the previous frames, followed by the update of the error-propagated distortions of the current frame. The recursive calculation of some distortion items requires the definition of distortion map for the storage of error-propagated distortions at each frame. In terms of the definition of distortion map, the estimation of end-to-end distortion can be done at either pixel level or block level. Actually, the block-level solution owns some advantages as follows. On the one hand, it can reduce the computing complexity and memory cost. On the other hand, it can also increase the robustness against the effects of subpixel MCP and de-blocking filtering used in the video coding architectures.

Further, we propose that the element in the distortion map can correspond to the minimum block size in MCP. Suppose block m in frame n references block m_k in frame ref. The overall end-to-end distortion of block m can be taken as the sum of that from each pixel and calculated according to (2). In particular, since block m_k may not always correspond to a single element in the distortion map, we derive $D_{ep}(ref,m_k)$ by weight-averaging the error-propagated distortions of the overlapped blocks, i.e.,

$$D_{ep}(ref,m_k) = \sum_{l=1}^{4} w_l D_{ep}(ref,m_l) \tag{5}$$

where w_l indicates the ratio that the overlapped region between m_l and m_k.

III. Error-Resilient Video Encoding

A. General R-D Optimization

The hybrid video coding usually contains a number of coding modes in the macroblock coding. The coding mode in H.264 can vary from the block partition of 4 × 4 to the whole block of 16 × 16 with respect to the different prediction types. Besides, the multiple references structure in H.264 also increases the coding options in the macroblock coding. For the selection of coding option that is composed of coding mode and reference frame, the Lagrangian method is usually used due to the consideration of the joint rate and distortion optimization, leading to a number of R-D optimization technologies. Assume o denotes a candidate coding option that is the combination of coding mode and reference frame. The best coding option of macroblock m in frame n can be selected as the one having the minimum coding cost $J(n,m,o)$ throughout the candidate coding options. In particular,

$$J(n,m,o) = D(n,m,o) + \lambda R(n,m,o), \tag{6}$$

where the Lagrange multiplier λ reveals the trade-off between distortion $D(n,m,o)$ and rate $R(n,m,o)$.

Actually, the R-D optimization technology has been well studied for the source video coding in error-free environment [19]. When it is applied in the error-prone environment, it is also a common idea to jointly consider the source distortion and the potential channel distortion together so as to achieve the best tradeoff between the overall distortion and the rate. The rate can be accurately estimated or directly calculated, as used in the traditional video coding. The overall end-to-end distortion can also be estimated, as discussed in Section II. Therefore, the remained problem is the derivation of a proper Lagrange multiplier. Intuitively, the multiplier should be related to the channel conditions, which is unnecessary to be the same as that used in error-free environment.

B. Derivation of Lagrange Multiplier

In this subsection, we derive the new Lagrange multiplier in the packet-loss environment following the similar routine of the derivation in the error-free environment in [20]. Assuming high-resolution quantization, it is well known that source distortion $D_s(R)$ conforms to

$$D_s(R) = \beta \cdot 2^{-\alpha R} \tag{7}$$

where β is a constant depending on the variance of the source. Further assuming that the distortion-to-quantizer relation is at sufficiently high rates, the source probability distribution can be approximated as uniform within each quantization interval Δ

$$D_s(\Delta) = \frac{\Delta^2}{12}. \tag{8}$$

Combining (7) and (8), we obtain

$$R(\Delta) = \frac{1}{\alpha}\log_2\left(\frac{\beta}{D_s(\Delta)}\right) = \frac{1}{\alpha}\log_2\left(\frac{\beta}{\Delta^2/12}\right). \tag{9}$$

According to (2) and (8), we also obtain

$$D(\Delta) = (1-p)\frac{\Delta^2}{12} + (1-p)D_{ep} + pD_{ec}. \tag{10}$$

Note that both D_{ep} and D_{ec} are independent of the quantization interval Δ of the current frame. Further combining the derivatives for Δ in (9) and (10), we can derive the new Lagrange multiplier as

$$\lambda = -\frac{dD(R)}{dR} = -\frac{dD}{d\Delta}\frac{d\Delta}{dR} = (1-p)\frac{\alpha\ln 2}{12}\Delta^2 = (1-p)\lambda_0 \tag{11}$$

where

$$\lambda_0 = \frac{\alpha\ln 2}{12}\Delta^2. \tag{12}$$

Here, λ_0 indicates the Lagrange multiplier in the error-free environment, which is related to the quantization parameter Q. In this paper, we employ λ_0 defined in [20], i.e.,

$$\lambda_0 = \begin{cases} 0.85 \cdot Q^2 & \text{for H.263} \\ 0.85 \cdot 2^{Q/3} & \text{for H.264} \end{cases}. \tag{13}$$

Note further that the Lagrange multipliers for H.263 and for H.264 are different, because the relationship between Δ and Q is linear in H.263 but it changes to be exponential in H.264. The above derivation clearly indicates that the trade-off between the rate and the distortion should be revaluated due to the consideration of the increased distortion caused by potential channel errors. In other words, when the channel condition becomes worse (i.e., with a larger channel error rate p), the rate becomes less important in the overall coding cost and also in the mode selection.

C. Mode Selection in H.264 Encoder

Upon the availability of the end-to-end distortion and the Lagrange multiplier, the R-D optimized coding mode selection for H.264 encoder in packet-loss environment can be easily derived. Above all, we explain the problems associated with B frame coding. The B frame in the previous video coding architectures can be directly encoded without the consideration of the error propagation problem. However, the B frame in H.264 may also cause the error propagation because it can also serve as a reference frame. In other word, it also requires the R-D optimized coding mode selection including the definition of distortion map in the error-resilient video coding. Nevertheless, the mode selection in B frame can be the same as that in P frame coding.

In general, the overall end-to-end distortion of a macroblock can be defined as the sum of distortions of all contained 4 × 4 subblocks. Suppose REF lists the reference frames of all subblocks in macroblock m in frame n in terms of coding option o. In other words, REF is decided by o. Similarly, suppose m_J lists the motion vectors in the same macroblock. According to (2), we can derive the end-to-end distortion as

$$D(n,m,o) = (1-p)\left(D_s(n,m,o) + D_{ep}(REF,m_J)\right) + pD_{ec}(n,m) \tag{14}$$

where $D_s(n,m,o)$, $D_{ep}(REF,m_J)$ and $D_{ec}(n,m)$ denotes the macroblock-level source distortion, error-propagated distortion and error concealment distortion, respectively. Note that $D_{ec}(n,m)$ is independent of o.

Suppose the channel error rate p is known as *a priori* in the encoder. According to (6), (11), and (14), the coding option $o^*(n,m)$ can be selected with

$$\begin{aligned} o*(n,m) &= \arg\min_{o\in O}((1-p)(D_s(n,m,o)+D_{ep}(REF,m_J)) \\ &\qquad +pD_{ec}(n,m)+\lambda R) \\ &= \arg\min_{o\in O}(D_s(n,m,o)+D_{ep}(REF,m_J)+\lambda_0 R) \end{aligned} \tag{15}$$

where O denotes the set of all candidate coding options of a macroblock. Since $D_{ec}(n,m)$ is independent of coding option, it is unnecessary to be calculated in mode selection. Therefore, the two distortion items in the final formula are only parts of the overall distortion. In other words, we still use the derived Lagrange multiplier in the coding option selection. The optimal coding option can be selected by going through all candidates, as that used in source video coding without error control [19]. After the current frame is encoded, the corresponding distortion map of the current frame is derived according to (4) for the coding of future frames.

The proposed algorithm can easily handle the deblocking filtering. In the proposed distortion model, only the distortion items unrelated to deblocking filtering are involved in mode selection except for the source distortion item. On the other hand, the new distortion items related to deblocking filtering is only performed once after the encoding. Besides the deblocking filtering, rate control is also a practical issue in the error-resilient video coding. As discussed above, the selection of coding mode requires a pre-determined quantization parameter to decide the Lagrange multiplier. However, the decision of the quantization parameter in rate control instead requires a pre-selected coding mode. Actually, this problem also exists in the conventional video coding in error-free environment, which has been well studied in the past years.

Moreover, the proposed R-D optimized coding mode selection scheme can be used jointly with the rate control algorithm developed in [18], which includes the rate control scheme in H.264 reference software as a special case. In particular, we employ the one-pass macroblock level rate control scheme derived in [18]. The quantization parameter is first determined according to the allocated bits and macroblock activity based on a rate-quantization model. Then, the coding mode is selected in terms of this quantization parameter with the proposed algorithm, as used in mode selection in error-free environment. The efficiency of rate control mainly lies in the accuracy of the employed rate-quantization model that is determined by the model parameters. Since the model parameters are updated in statistics at frame level, as used in [18], it can also adaptively comply with the error-resilient video coding that includes the proposed mode selection.

IV. Experimental Results

A. End-to-End Distortion Estimation

In this subsection, we evaluate the accuracy of the proposed end-to-end distortion estimation method. Above all, we assume that it is sufficient to achieve the theoretical expectation of end-to-end distortion in the decoder by simulating the decoding process 500 times, as used in [14]. H.264/AVC JM7.5c is selected as the reference software [21], and ER-RDO is employed for the coding mode selection. Then, we estimate the end-to-end distortion in the encoder with the proposed method and ER-RDO, respectively. Two sequences including Foreman (300 frames, 30 fps, QCIF) and News (300 frames, 30 fps, QCIF) are used in the test. Only the first frame is encoded as I frame, and all the remaining frames are encoded as P frames. Each row of macroblocks composes a slice and is transmitted in a separate packet. Hence, each packet is independently decodable. For the Foreman sequence, bit rate $r = 64$ kbps, frame rate $f = 7.5$ fps, and packet loss rate $p = 10\%$. For the News sequence, $r = 64$ kbps, $f = 15$ fps, and $p = 10\%$.

The overall accuracy of end-to-end distortion estimation is evaluated first. Suppose the frame-level end-to-end distortion is defined as the average of macroblock-level distortions. In particular, $D(n)$ denotes the theoretical expectation of end-to-end distortion at frame n, i.e.,

$$D(n) = \frac{1}{M}\sum_{m=1}^{M} D(n,m), \tag{16}$$

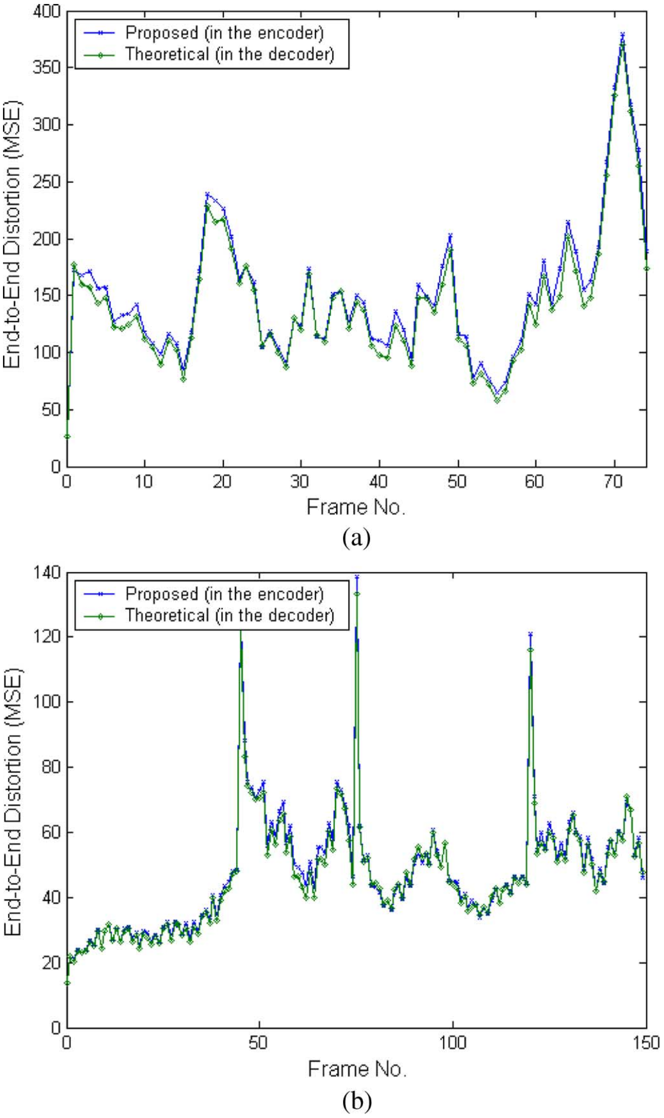
where M is the number of macroblocks in the frame. $D_P(n)$ and $D_K(n)$ denote the frame-level end-to-end distortions estimated by the proposed algorithm and ER-RDO with K decoders, respectively. We define the estimation error E_D between $D_J(n)$ and $D(n)$ as

$$E_D(J) = \sqrt{\frac{1}{T}\sum_{n=1}^{T}\left(D_J(n) - D(n)\right)^2} \tag{17}$$

where T is number of frames in the sequence. Note that $D_K(n)$ is identical to $D(n)$ when $K = 500$. We only test the cases that $K = 30$ and $K = 100$ for ER-RDO.

TABLE I
Statistical Results of Average Difference ED(J)

Sequence	E_D(K=30)	E_D(K=100)	E_D(P)
Foreman	17.49	11.84	9.44
News	5.37	2.05	1.74



(a)

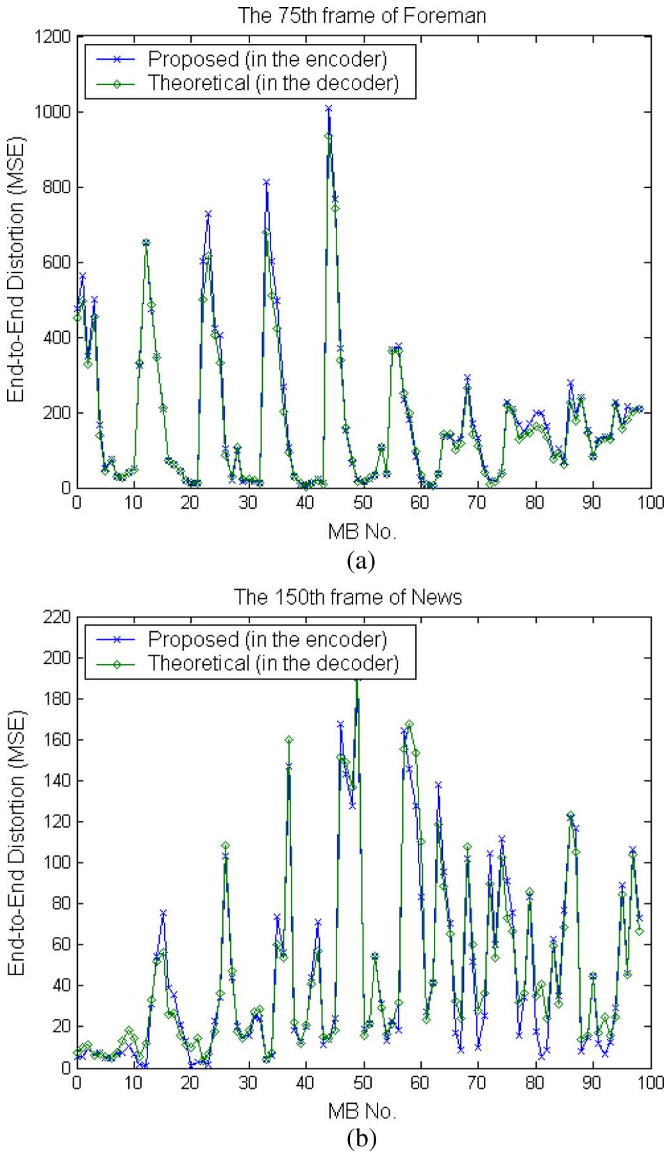
(b)

Fig. 1. Comparison between the actual and estimated end-to-end distortion at frame-level statistics. (a) *Foreman* at $r = 64\,\mathrm{kbps}$, $f = 7.5\,\mathrm{fps}$, and $p = 10\%$; and (b) *News* at $r = 64\ \mathrm{kbps}$, $f = 15\ \mathrm{fps}$, and $p = 10\%$.

Table I illustrates the estimation error between the estimated distortion in the encoder and the theoretical expectation in the decoder. Obviously, $E_D(P)$ is smaller than $E_D(K = 100)$ and $E_D(K = 30)$ for both sequences. In other words, the estimate of the proposed algorithm is more precise than ER-RDO with $K = 100$ and of course ER-RDO with $K = 30$. Further, Fig. 1 shows the end-to-end distortion estimated with the proposed method and the associated theoretical expectation at every frame. The plots indicate that the estimated distortion in the encoder is very close to its theoretical expectation in the decoder along the whole sequence. We also evaluate the accuracy of the end-to-end distortion estimated at macroblock level. Fig. 2 shows the distortion of every macroblock in the 75th frame of Foreman and in the 150th frame of News, respectively. According to these results, the estimated end-to-end distortion of every macroblock is also very close to its statistical expectation, which proves that the proposed method has successfully handled the local features in a frame.

B. Error-Resilient Video Coding

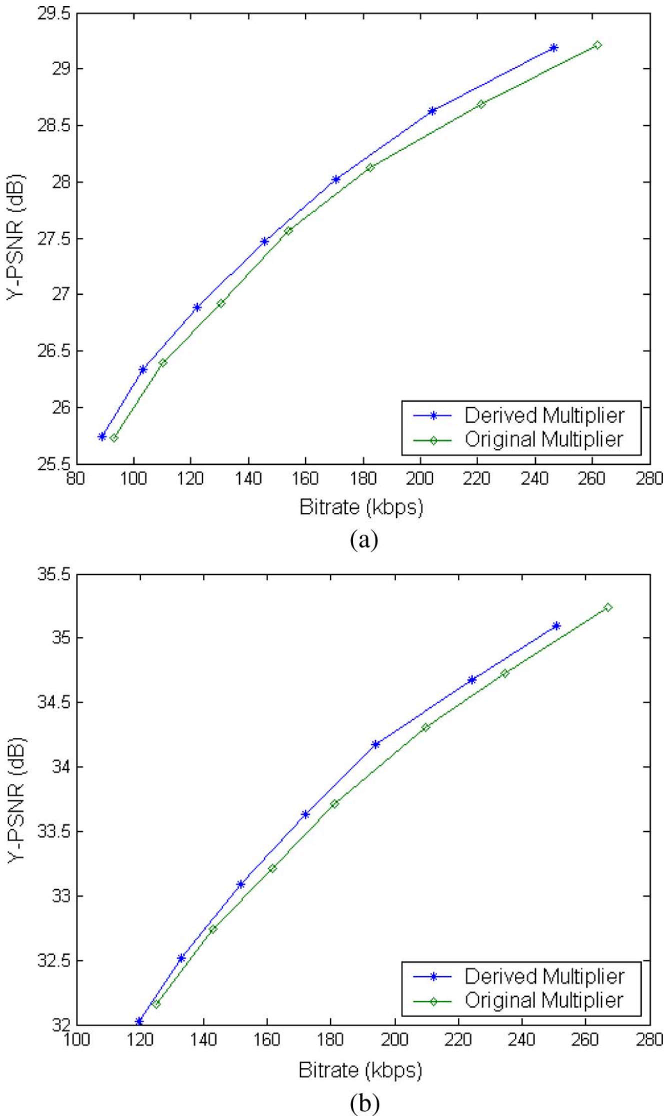
In this subsection, we evaluate the performance of the proposed R-D optimized coding mode selection scheme in H.264



(a)

(b)

Fig. 2. Comparison between the actual and estimated end-to-end distortion at MB-level statistics. (a) 75th frame of *Foreman* and (b) 150th frame of *News*.



(a)

(b)

Fig. 3. Evaluation of the derived Lagrange multiplier. (a) *Foreman* and (b) *News*.

error-resilient coding. Above all, we test the efficiency of the derived Lagrange multiplier. Foreman (30 fps, QCIF) and News (30 fps, QCIF) are selected as the test sequences. The reference software is still H.264/AVC JM7.5c. The coding mode is selected by the proposed method with the derived Lagrange multiplier and with the original multiplier, respectively. Each row of macroblocks composes a slice and is packed in a separate packet. As used in [12], a random packet loss generator is used to drop packets at a selected loss rate (i.e., 20% in the test). The temporal-replacement method is employed for error concealment in the decoder. The coded bitsteam is decoded 500 times under the generated packet loss patterns. Fig. 3 shows the R-D curves from the different Lagrange multipliers, which indicates that the proposed multiplier outperforms the original multiplier 0.2 dB. In the test, the overall distortions of News are smaller than that of Foreman, because the temporal-replacement error concealment scheme works better for News rather than for Foreman. It should be noted that the Lagrange multiplier typically has much larger magnitude than the rate. Considering the small variance of the rate, the effectiveness of the multiplier may be degraded sometimes.

Afterwards, we evaluate the overall performance of H.264 error-resilient coding using the test conditions specified in [22], which refers to the error patterns defined in [23]. The comparisons are among ER-RDO, ROPE, random update, and the proposed method. For ER-RDO, we set $K = 500$ and $K = 30$, respectively. In addition, we employ the rate control in the test so as to achieve the same bit-rates for the fair comparison among different error control schemes. We have done some experiments to evaluate the performance of rate control in ER-RDO, ROPE and the proposed algorithm, which show that the overall R-D performances from rate control and fixed quantization parameter coding are very close. Five bitstreams are generated in terms of each algorithm, including Hall Monitor (32 kbps, 10 fps, QCIF), Foreman (64 kbps, 7.5 fps, QCIF), Foreman (144 kbps, 7.5 fps, QCIF), Paris (144 kbps, 15 fps, CIF) and

TABLE II
COMPARISON RESULTS OF AVERAGE PSNR (IN dB)
AT DIFFERENT PACKET LOSS RATES

Sequence	Scheme	PSNR of Different Loss Rate (dB)				Running Time (s)
		3%	5%	10%	20%	
Foreman 64 kbps	ROPE	30.35	29.45	27.51	25.54	34.32
	Proposed	30.31	29.48	27.60	25.58	22.44
	ER-RDO (30)	30.04	28.99	27.28	25.35	88.70
	ER-RDO (500)	30.21	29.42	27.46	25.50	1317.38
	Random Update	29.40	28.33	26.47	24.83	15.73
Foreman 144 kbps	ROPE	34.30	33.19	30.78	27.94	39.76
	Proposed	34.36	33.22	30.85	28.17	27.28
	ER-RDO (30)	34.04	32.83	30.58	27.85	99.36
	ER-RDO (500)	34.23	33.15	30.60	28.11	1454.58
	Random Update	33.06	32.03	30.02	27.37	16.33
Hall 32 kbps	ROPE	33.64	33.41	32.21	31.20	37.43
	Proposed	33.58	33.44	32.29	31.19	25.38
	ER-RDO (30)	32.74	32.28	31.81	30.79	92.39
	ER-RDO (500)	33.25	33.09	31.87	30.93	1066.22
	Random Update	30.74	29.42	28.23	26.77	24.51
Paris 144 kbps	ROPE	27.47	26.91	25.88	24.79	582.63
	Proposed	27.51	27.01	25.93	24.84	387.88
	ER-RDO (30)	27.01	26.27	25.65	24.35	968.75
	ER-RDO (500)	26.71	26.15	25.59	24.54	7243.56
	Random Update	25.51	24.77	23.57	22.66	340.65
Paris 384 kbps	ROPE	33.09	32.26	30.92	29.19	653.54
	Proposed	33.08	32.29	30.98	29.33	445.31
	ER-RDO (30)	32.23	31.60	30.36	28.62	1006.51
	ER-RDO (500)	32.68	31.72	30.63	28.98	7512.04
	Random Update	30.29	29.23	27.70	26.14	358.57

Paris (384 kbps, 15 fps, CIF). Note that there are four packets per frame for QCIF and nine packets for CIF, and the 40 bytes of IP/UDP/RTP headers per packet have been taken into account. These bitstreams are decoded after simulating the packet loss rate of 3%, 5%, 10% and 20%, respectively. We assume that the packet containing the parameter set and packets of the first frame are conveyed reliably, which is possible out-of-band during the session setup. The bitstream is decoded multiple times, and the number of decoding runs is selected to have totally at least 8000 packets.

As shown in Table II, the proposed algorithm outperforms ER-RDO in all cases. In particular, it outperforms ER-RDO with $K = 30$ about 0.5 dB. The comparison between ROPE and ER-RDO has been reported in [14], in which ER-RDO outperforms ROPE in H.264 error resilience coding. However, in the test of ROPE in [14], the subpixel motion vector positions were rounded to the closest full-pixel positions and the deblocking filtering was ignored. In this test, we further extend ROPE in H.264 error resilience coding, in which the subpixel interpolation using 6-tap filters is performed on the first moment and the square root of the second moment of the reconstructed pixel value, respectively. Thus, in our test, the improved ROPE performs better than ER-RDO. Nevertheless, the R-D performance of the proposed method is still slightly better than ROPE. Moreover, it should be noted that the proposed method is implemented in block level and ROPE is implemented in pixel level, which leads to the different computational complexity and memory cost. Actually, it is almost impossible to have the block-level ROPE, because the distortion model in ROPE is very sensitive to the approximation errors caused by pixel averaging operations.

In addition, we present the tendencies of the intra update rate in terms of the optimal inter/intra coding mode selection as follows. Fig. 4 shows the percentage of intra blocks in terms of every frame selected with the proposed algorithm. Table III gives the average percentage of intra blocks in terms of all frames in a video. It can be observed that the intra update rate greatly depends on the video content. In general, the video with large motion usually corresponds to large number of intra blocks. Moreover, the intra rate also increases while the error rate increases. Based on the achieved number of intra blocks at each frame, we can further test the random intra update coding. It should be noted that the intra update rates are achieved from the fixed quantization parameter coding, because the rate control is closely related to the buffer conditions that may not be duplicated in the different coding runs. Table II shows the overall R-D performances in terms of random intra update coding. It can be observed that the R-D optimized coding mode selection schemes always outperform the random intra update, especially for the sequence with static and slow motion. It is because the R-D optimized coding mode selection can refresh with intra blocks at the regions that may cause significant drifting errors.

C. Video Transport Over IP Network

Besides the off-line testing, we also realize a live video streaming system for the on-line testing. At the server side, a real-time H.264 encoder is implemented, in which the error control and the rate control are jointly employed. At the receiver side, an elaborate error concealment method developed in [24] is employed. The described streaming system focuses on the unicast mode. The streaming session is setup with the real-time streaming protocol (RTSP). One UDP connection is immediately established between the server and the receiver once a streaming session is setup. The encoder real-time generates the bitstream, which is directly packed in the RTP format and delivered to the receiver. Meanwhile, the feedback data including the packet loss rate is packed in RTCP format and sent back to the network monitor module at the server side. Hence, the encoder can adjust the coding parameters based on the feedback information.

While the described video streaming system can be run under the real IP networks, we use WiNE, a network emulator for windows platform [25], to add the Gilbert-Elliot model-based packet loss behavior of a link. In the testing, the packet loss rate varies from 0.5% to 10%, and the average bit rate is 144 kbps. Note that both the packet loss rate can be estimated on-the-fly at the receiver. The Foreman sequence is repeatedly real-time encoded at the server side, with the proposed error control scheme switched on and off, respectively. The network conditions are recorded for the off-line testing of ER-RDO, and hence the subjective visual quality can be evaluated. Actually, it is difficult to realize a real-time H.264 encoder with ER-RDO due to its high complexity. Fig. 5 shows the decoded frames from three different schemes. Obviously, the visual quality without using error control is not acceptable. In the contrast, the visual quality

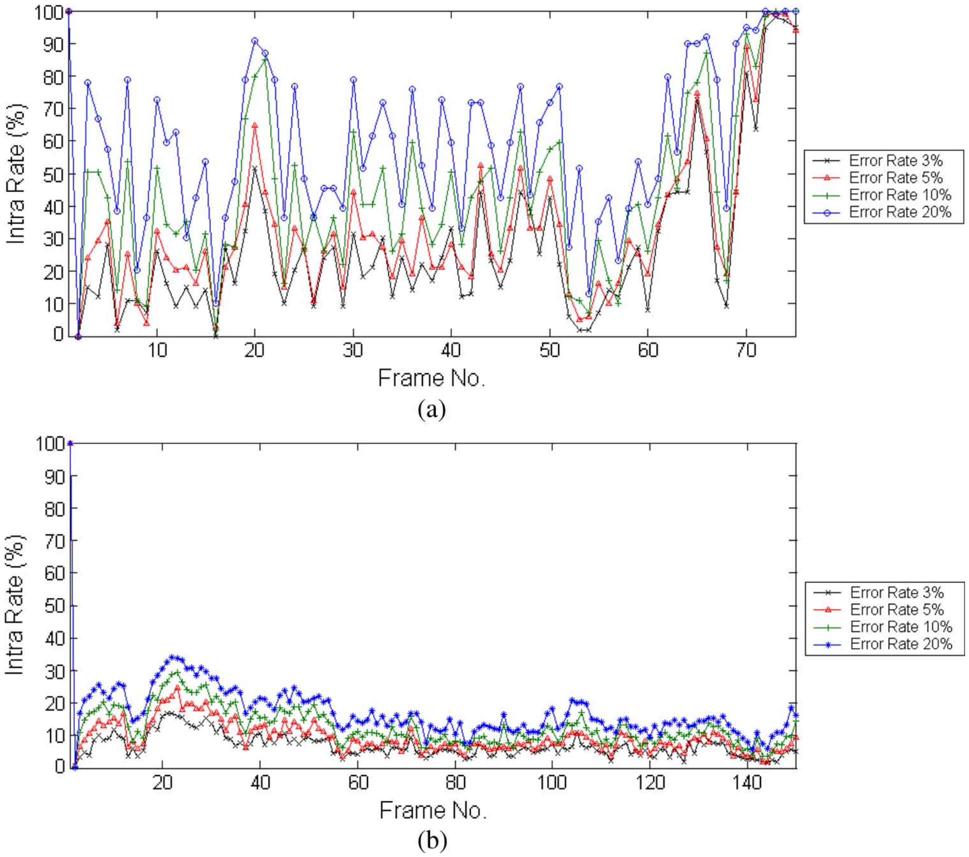


Fig. 4. The percentages of intra blocks from the proposed coding mode selection under the different error rates. (a) *Foreman* at 64 kbps and (b) Paris at 384 kbps.

TABLE III
PERCENTAGE OF INTRA-CODED BLOCKS AT DIFFERENT PACKET LOSS RATES

Sequence	Percentage of Intra Blocks (%)			
	Error Rate 3%	Error Rate 5%	Error Rate 10%	Error Rate 20%
Foreman (64 kbps)	27.52	33.05	43.84	59.06
Foreman (144 kbps)	46.98	54.33	65.90	80.47
Hall (32 kbps)	1.67	2.44	3.63	5.32
Paris (144 kbps)	2.27	3.39	5.02	7.92
Paris (384 kbps)	6.73	9.31	12.61	16.84

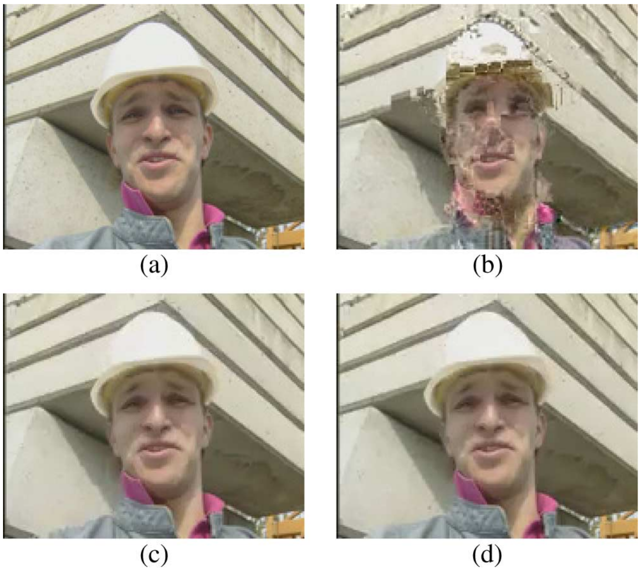


Fig. 5. Comparison of subjective visual quality from live streaming. (a) One original frame from *Foreman*, and decoded frames that are encoded (b) without error control, (c) with the proposed error control, and (d) with ER-RDO.

with either the proposed algorithm or the ER-RDO has shown significant improvements.

V. CONCLUSION

A concise and efficient end-to-end distortion model for R-D optimized coding mode selection in error-resilient video coding has been presented in this paper. Distinctively, the proposed model takes the end-to-end distortion as the sum of several separable distortion items. In particular, it keeps track of the error-propagated distortion through recursive calculation. Since the overall distortion is taken as the sum of several distortion items (i.e., positive values), it can suppress the approximation errors caused by pixel averaging operations such as subpixel prediction. Therefore, its extension to block-level implementation can be readily achieved. Further, a new Lagrange multiplier is derived based on the proposed end-to-end distortion model, which takes into account of the network conditions such as the packet loss rate. The rate control and complexity issues have also been analyzed in this paper.

Compared to the other R-D optimized coding mode selection schemes, the proposed algorithm owns the following advantages. Firstly, the physical explanation to the distortion items in the proposed distortion model is obvious, e.g., the distortion caused by error concealment, which is helpful in the derivation of a proper Lagrange multiplier for R-D optimized mode selection in packet-loss environment. Secondly, the proposed algorithm can easily tackle the problem of subpixel MCP, because the separated distortion items can suppress the estimation errors caused by pixel averaging operations. Thirdly, the proposed algorithm can also easily handle the deblocking filtering, because each distortion item in the proposed algorithm need be calculated only once either before or after coding mode selection (i.e., encoding with deblocking filtering).

REFERENCES

[1] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 1012–1032, Jun. 2000.

[2] Y. Wang and Q. F. Zhu, "Error control and concealment for video communication: a review," *Proc. IEEE*, vol. 86, pp. 974–997, May 1998.

[3] G. Cote and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the internet," *Signal Process.: Image Commun.*, vol. 15, pp. 25–34, Sep. 1999.

[4] Q. F. Zhu and L. Kerofsky, "Joint source coding, transport processing and error concealment for H.323-based packet video," in *Proc. SPIE VCIP'99*, San Jose, CA, Jan. 1999, vol. 3653, pp. 52–62.

[5] P. Haskell and D. Messerschmitt, "Resynchronization of motion-compensated video affected by ATM cell loss," in *Proc. IEEE ICASSP'92*, 1992, vol. 3, pp. 545–548.

[6] R. O. Hinds, "Robust Mode Selection for Block Motion-Compensated Video Encoding," Ph.D. dissertation, MIT, Cambridge, MA, Jun. 1999.

[7] A. Leontaris and P. C. Cosman, "Video compression with intra/inter mode switching and a dual frame buffer," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2003, pp. 63–72.

[8] D. Wu, Y. T. Hou, B. Li, W. Zhu, Y.-Q. Zhang, and H. J. Chao, "An end-to-end approach for optimal mode selection in Internet video communication: theory and application," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 977–995, Jun. 2000.

[9] T. Wiegand, N. Farber, K. Stuhlmuller, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 1050–1062, Jun. 2000.

[10] G. Cote, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 952–965, Jun. 2000.

[11] Z. H. He, J. F. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 511–523, Jun. 2002.

[12] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 966–976, Jun. 2000.

[13] H. Yang and K. Rose, "Recursive end-to-end distortion estimation with model-based cross-correlation approximation," in *Proc. IEEE ICIP'03*, Sep. 2003, vol. 3, pp. 469–472.

[14] T. Stockhammer, D. Kontopodis, and T. Wiegand, "Rate-distortion optimization for JVT/H.26L coding in packet loss environment," in *Proc. Packet Video Workshop*, Pittsburgh, PA, Apr. 2002.

[15] T. Stockhammer and S. Wenger, "Standard-compliant enhancement of JVT coded video for transmission over fixed and wireless IP," in *Proc. IWDC 2002*, Capri, Italy, Sep. 2002.

[16] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 560–576, Jul. 2003.

[17] MPEG Video Group, Text of ISO/IEC 14496-5: 2004/PDAM6 (AVC Reference Software) , ISO/IEC TC JTC 1/SC 29 N5821, Aug. 2003.

[18] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application to rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1533–1544, Dec. 2005.

[19] 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.

[20] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. ICIP2001*, Thessaloniki, Greece, Oct. 2001.

[21] H.264/MPEG-4 AVC Reference Software [Online]. Available: http://bs.hhi.de/~suehring/tml/download/jm75c.zip

[22] S. Wenger, Common Conditions for Wire-Line, Low Delay IP/UDP/RTP Packet Loss Resilient Testing ITU-T SG16 Doc. VCEG-N79r1, Sep. 2001.

[23] ——, Error Patterns for Internet Experiments ITU-T SG16 Doc. Q15-I-16r1, 1999.

[24] L. Su, Y. Zhang, W. Gao, Q. Huang, and Y. Lu, "Improved error concealment algorithms based on H.264/AVC non-normative decoder," in *Proc. ICME2004*, Taibei, Taiwan, R.O.C., Jun. 2004.

[25] Wireless/Wired Wide Area Network Emulator (WiNE) User Manual ver. Version 3.0, Microsoft Research Asia. Beijing, China, Jan. 6, 2003.

Yuan Zhang received the B.S. and M.S. degrees in electronic engineering from Communication University of China (CUC), Beijing, in 1995 and in 1998, respectively. Since 2001, she has been pursuing the Ph.D. degree at the Graduate School of the Chinese Academy of Sciences, Beijing.

In 1998, she joined the faculty of the TV Engineering Department, CUC, where she is currently an Associate Professor. Her research interests include video compression, joint source-network coding, and video streaming.

Wen Gao received the M.S. and the Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1985 and in 1988, respectively, and the Ph.D. degree in electronics engineering from University of Tokyo, Tokyo, Japan, in 1991.

He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, in 1992, and a Visiting Professor at Robotics Institute, Carnegie-Mellon University, Pittsburgh, PA, in 1993. From 1994 to 1995, he was a Visiting Professor with the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. Currently, he is a Professor with the School of Electronic Engineering and Computer Science, Peking University, Peking, China, and a Professor in computer science at Harbin Institute of Technology. He is also the Honor Professor in computer science at City University of Hong Kong, and the External Fellow of International Computer Science Institute, University of California, Berkeley. He has published seven books and over 200 scientific papers. His research interests are in the areas of signal processing, image and video communication, computer vision, and artificial intelligence. He is Editor-in-Chief of the *Chinese Journal of Computers*.

Dr. Gao chairs the Audio Video coding Standard (AVS) workgroup of China. He is the head of Chinese National Delegation to MPEG working group (ISO/SC29/WG11).

Yan Lu received the B.S., M.S., and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1997, 1999, and 2003, respectively.

From 1999 to 2000, he was a Research Assistant with the Computer Science Department, City University of Hong Kong, Hong Kong, China. From 2001 to 2004, he was with the Joint R&D Lab (JDL) for advanced computing and communication, Chinese Academy of Sciences, China. Since April 2004, he has been with Microsoft Research Asia, Beijing. His research interests include image and video coding, texture compression, and multimedia streaming.

Qingming Huang received the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China in 1994.

He was a Postdoctoral Fellow in National University of Singapore from 1995 to 1996, and worked in Institute for Infocomm Research, Singapore, as Member of Research Staff from 1996 to 2002. Currently, he is a Professor in Graduate School of Chinese Academy of Sciences. Beijing. His current research areas are image processing, video analysis, video coding, and pattern recognition.

Debin Zhao received the B.S., M.S., and Ph.D. degrees in computer science, all from the Harbin Institute of Technology, Harbin, China, in 1985, in 1988, and in 1998, respectively.

He was an Associate Professor in the Department of Computer Science, Harbin Institute of Technology, and a Research Fellow in the Department of Computer Science, City University of Hong Kong, Hong Kong, China, from 1989 to 1993. He is currently Professor with the Department of Computer Science, Harbin Institute of Technology. His research interests include data compression, image processing, and human–machine interface. He has coauthored over 70 publications.