EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Modeling of SSIM-based end-to-end distortion for error-resilient video coding

Qiang Peng[1], Lei Zhang[1,2*], Xiao Wu[1] and Qionghua Wang[3]

Abstract

Conventional end-to-end distortion models for videos measure the overall distortion based on independent estimations of the source distortion and the channel distortion. However, they are not correlating well with the perceptual characteristics where there is a strong inter-relationship among the source distortion, the channel distortion, and the video content. As most compressed videos are represented to human users, perception-based end-to-end distortion model should be developed for error-resilient video coding. In this paper, we propose a structural similarity (SSIM)-based end-to-end distortion model to optimally estimate the content-dependent perceptual distortion due to quantization, error concealment, and error propagation. Experiments show that the proposed model brings a better visual quality for H.264/AVC video coding over packet-switched networks.

Keywords: End-to-end distortion model; Structural similarity; Error resilience

1 Introduction

Most video coding standards achieve high compression using transform coding and motion-compensated prediction, which creates a strong spatial-temporal dependency in compressed videos. Thus, transmitting highly compressed video streams over packet-switched networks may suffer from spatial-temporal error propagation and may lead to severe quality degradation at the decoder side [1]. To protect compressed videos from packet loss, error-resilient video coding becomes a crucial requirement. Given transmission conditions, such as bit rate and packet loss ratio, the target of error resilient video coding is to minimize the distortion at the receiver [2]:

$$\min\{D\} \qquad \text{s.t.} \quad R \le R_T \quad \text{and} \ \rho \tag{1}$$

where D and R denote the distortion at the receiver and the bit rate, respectively. R_T is the target bit rate and ρ is the packet loss ratio. Note that we assume the transmission conditions are available at the encoder throughout this paper. This can be either specified as part of the initial negotiations or adaptively calculated from information provided by the transmission protocol [3].

Assume packet containing video data is lost in the channel and the decoder performs error concealment. Clearly, the resulting reconstruction at the decoder is different from the reconstruction at the encoder and the difference will propagate to the following frames due to the prediction chain. Therefore, the key challenge of the error-resilient video coding is to estimate at the encoder the reconstruction error and error propagation of the decoder, which is useful to optimize the coding options to solve the above minimization problem.

A number of end-to-end distortion models (also known as joint source-channel distortion models) for video transmission over lossy channels have been proposed in the literature. In [4,5], several low-complexity estimation models were presented for low error rate applications. For a more accurate distortion estimation model, the work in [2] developed a frame-level recursion distortion model, which relates to the channel-induced distortion due to bit errors. Another efficient approach is the well-known recursive optimal pixel estimation (ROPE) model [3] and its extensions [6-10], which estimate the overall distortion due to quantization, error concealment, and error propagation. Recently, several novel source-channel distortion models were developed for distributed video coding [11], generic multi-view video transmission [12], and error-resilient schemes based on forward error correction [13].

* Correspondence: zl.swjtu@gmail.com
[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
[2]Institution of Academia Sinica, Jiuzhou Electric Group, Sichuan, China
Full list of author information is available at the end of the article

However, these models are derived in terms of mean squared error (MSE), which has been criticized for weak correlation with perceptual characteristics. As most compressed videos are presented to humans, it is meaningful to incorporate visual features into the error-resilient video coding to protect important visual information of compressed videos from packet loss. Thus, several region-of-interest (ROI)-based approaches were presented to better evaluate the visual quality [14,15]. However, ROI-based approaches do not provide accurate distortion estimation, and ROI determination may be difficult for most videos, especially for videos with natural scenes. Therefore, it is expected that a perception-based end-to-end distortion model could provide a more general and accurate perceptual distortion estimation.

In [16], the structural similarity (SSIM)-based end-to-end distortion was predicted by several factors extracted from the encoder. Although the variation trend is very similar at the block level, the estimated SSIM cannot reach the peak points of the actual SSIM. In [17], a parametric model was proposed to accurately estimate the degradation of SSIM over error-prone networks, in which the content, encoding, and network parameters are considered. However, the encoding parameters only included the number of slices per frame and the GOP length. The proposed model cannot estimate the relative quality of a block given different coding modes. In our earlier work [18], we introduced a block-level SSIM-based distortion model into the error-resilient video coding to minimize the perceptual distortion. In [19], improved SSIM-based distortion model and Lagrange multiplier decision method are proposed for better coding performance. In [18] and [19], the expected SSIM scores were estimated by the expected decoded frames. Due to the nonlinear variation of SSIM, the estimated SSIM scores may be less accurate, especially at high bit rate.

In this paper, we develop an SSIM-based end-to-end distortion model to estimate the overall perceptual distortion for H.264/AVC coded video transmission over packet-switched networks. Unlike the traditional end-to-end distortion model, the perceptual quantization distortion and the perceptual error propagation distortion are dependent on the video content, which makes the end-to-end distortion become complex or difficult to estimate at the encoder. Therefore, this paper provides two major contributions: 1) a SSIM-based reconstruction quality model; 2) a SSIM-based error propagation model. Both models are useful to estimate the content-dependent perceptual distortion at the encoder. Our extensive experimental results demonstrate that the proposed end-to-end distortion model can bring visual quality improvement for H.264/AVC video coding over packet-switched networks. We would like to mention that the scheme presented in this paper is an enhanced approach based on our preliminary work in [20]. Different settings are considered in this paper, including additional descriptions of related works, technical and implementation details, and comparison experiment results to better evaluate the efficiency of the proposed scheme.

The rest of the paper is organized as follows. Section 2 states the problem and motivation. Section 3 describes the proposed SSIM-based end-to-end distortion model. Section 4 introduces the distortion model into the error-resilient video coding. Section 5 provides the simulation results and Section 6 concludes the paper.

2 Problem and motivation

For H.264/AVC coded video transmission over packet-switched networks, the general formulation of the widely used MSE-based end-to-end distortion can be defined as

$$D = (1-\rho)\cdot D_Q + \rho\cdot D_C + (1-\rho)\cdot D_{P\_f} + \rho\cdot D_{P\_c} \quad (2)$$

where ρ is the packet loss ratio. D is the estimated overall distortion. D_Q denotes the source distortion due to the quantization. D_c, D_{P_f}, and D_{P_c} represent the channel distortion due to the error concealment, error propagation from the reference frames, and error propagation from the concealment frames, respectively.

With such a model in Equation 2, the end-to-end distortion can be individually and independently estimated by the quantization distortion, error concealment distortion, and error propagation distortion. This model is appealing because it is easy to calculate and has clear physical meanings. However, since the perceptual distortion is dependent on the video content, the individual and independent objective distortion estimation does not correspond well with human perceptual characteristics. For instance, as shown in Figure 1, since ten compressed or lossy transmitted videos (Live video quality database [21,22]) have different perceptual characteristics, a similar objective distortion may result in different levels of perceptual quantization distortion or transmission distortion. Therefore, we aim to propose a perception-based end-to-end distortion model for more accurate estimation of the overall perceptual distortion in the following section.

3 SSIM-based end-to-end distortion model

To estimate the overall perceptual distortion of decoded videos, we adopt the SSIM index [23] as the perceptual distortion metric due to its best trade-off among simplicity and efficiency [24]. Three important perceptual components, luminance, contrast, and structure, are combined as an overall similarity measure. For two images x and y, the SSIM index is defined as follows:

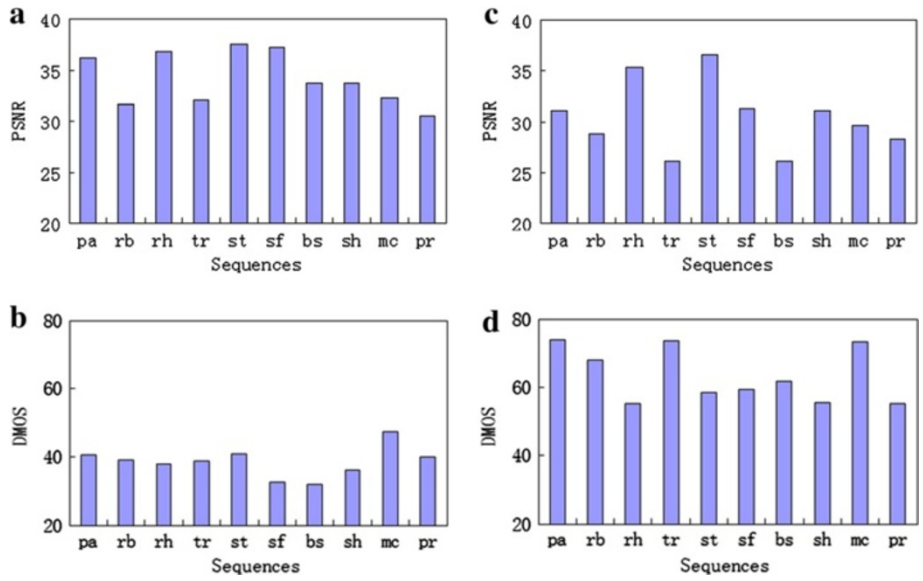

Figure 1 Quality comparison. (a) PSNR results for compressed videos; **(b)** DMOS results for compressed videos; **(c)** PSNR results for lossy transmitted videos; **(d)** DMOS results for lossy transmitted videos.

$$\begin{aligned}\mathrm{SSIM}(x,y) &= l(x,y)\cdot c(x,y)\cdot s(x,y)\\ &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2+\mu_y^2+c_1}\cdot\frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2}\end{aligned} \tag{3}$$

where $l(x,y)$, $c(x,y)$, and $s(x,y)$ represent the luminance, contrast and structure perceptual components, respectively. μ, σ^2, and σ_{xy} are the mean, variance, and cross covariance, respectively. c_1 and c_2 are used to avoid the instability when means or variances are close to zero.

Based on the perceptual distortion metric, we develop a novel end-to-end distortion model as follows. In Figure 2, b denotes the original block and $\tilde{b}$ is the corresponding reconstruction block at the decoder. $\hat{r}$ and $\tilde{r}$ represent the prediction block of b at the encoder and at the decoder, respectively. e denotes the prediction residual and its reconstruction value is $\hat{e}$. If the block is received correctly, $\tilde{b} = \tilde{r} + \hat{e}$. When the block is lost, an error concealment technique is used to estimate the missing content. Let $\hat{c}$ and $\tilde{c}$ represent the concealment block of b at the encoder and at the decoder, respectively. In this case, $\tilde{b} = \tilde{c}$. For a given packet loss ratio ρ, the general SSIM-based end-to-end distortion can be expressed as

$$\begin{aligned}D_{\mathrm{SSIM}}(b,\tilde{b}) &= (1-\rho)\cdot E\{1-\mathrm{SSIM}(b,\tilde{r}+\hat{e})\}\\ &\quad +\rho\cdot E\{1-\mathrm{SSIM}(b,\tilde{c})\}\end{aligned} \tag{4}$$

with

$$\begin{aligned}E\{1-\mathrm{SSIM}(b,\tilde{r}+\hat{e})\} &= 1-E\{\mathrm{SSIM}(b,\tilde{r}+\hat{e})\}\\ &= 1-\Phi_r\cdot E(b,\hat{r}+\hat{e})\end{aligned} \tag{5}$$

$$\begin{aligned}E\{1-\mathrm{SSIM}(b,\tilde{c})\} &= 1-E\{\mathrm{SSIM}(b,\tilde{c})\}\\ &= 1-\Phi_c\cdot\mathrm{SSIM}(b,\hat{c})\end{aligned} \tag{6}$$

where $E\{\ \}$ is the expectation operator. Φ is the error propagation factor. It indicates how the transmission errors from prediction block or concealment block influence the quality of current block. $\mathrm{SSIM}(b,\tilde{r}+\hat{e})$ and $\mathrm{SSIM}(b,\tilde{c})$ denote the quality of prediction coding and error concealment at the decoder, respectively. SSIM

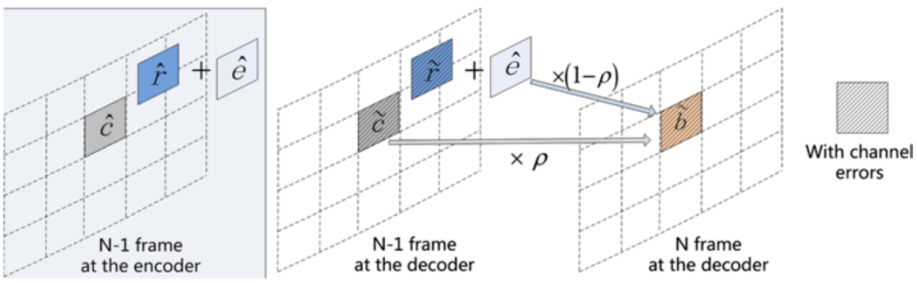

Figure 2 End-to-end distortion model in lossy transmission channel.

$(b, \tilde{r} + \hat{e})$ and $\mathrm{SSIM}(b, \hat{c})$ are the quality of prediction coding and error concealment at the encoder, respectively.

With this formula, the reconstruction quality SSIM $(b, \hat{r} + \hat{e})$ and error propagation factor Φ are the key terms of the SSIM-based end-to-end distortion model. In the following section, we will make a development of the two terms based on content dependency.

3.1 Development of reconstruction quality model

In this section, we aim to estimate the content-dependent reconstruction quality $\mathrm{SSIM}(b, \hat{r} + \hat{e})$ at the block level (the 4×4 transform and quantization unit is used throughout this paper). Since the accurate reconstruction quality can only be obtained after de-quantization, the proposed quality estimation reduces the computational complexity of de-quantization process for each candidate modes.

According to the SSIM index, the reconstruction quality is derived as

$$\mathrm{SSIM}(b, \hat{r} + \hat{e}) = l(b, \hat{r} + \hat{e}) \cdot c(b, \hat{r} + \hat{e}) \cdot s(b, \hat{r} + \hat{e}) \quad (7)$$

with

$$l(b, \hat{r} + \hat{e}) = \frac{2\mu_b \mu_{\hat{r}} + 2\mu_b \mu_{\hat{e}} + c_1}{\mu_b^2 + \mu_{\hat{r}}^2 + \mu_{\hat{e}}^2 + 2\mu_{\hat{r}} \mu_{\hat{e}} + c_1} \quad (8)$$

$$c(b, \hat{r} + \hat{e}) \cdot s(b, \hat{r} + \hat{e}) = \frac{2\sigma_{b\hat{r}} + 2\sigma_{b\hat{e}} + c_2}{\sigma_b^2 + \overset{2}{\sigma\hat{r}} + \overset{2}{\sigma\hat{e}} + 2\sigma_{\hat{r}\hat{e}} + c_2} \quad (9)$$

From Equations 8 and 9, we can see that the estimation of content-dependent reconstruction quality is converted to the estimation problem of three content-independent parameters: 1) the variance of reconstructed prediction residual; 2) the cross-covariance between the reconstructed prediction residual and current block; 3) the cross-covariance between the reconstructed prediction residual and prediction block.

It is reported that the DCT coefficients of prediction residual closely follows a zero-mean Laplacian distribution [25]. Based on this phenomenon, the work in [26] proved that the reconstruction distortion from the prediction residual can be estimated by the Laplacian parameter and the quantization step. Extending the derivation in [26] into pixel-domain, we establish the following two estimation models for above parameters:

$$\overset{2}{\sigma\hat{e}} = M_{\mathrm{var}}(\alpha, QP) \cdot \sigma_e^2, \quad \alpha = \sqrt{2/\sigma_e^2} \quad (10)$$

$$\sigma_{b\hat{e}} = M_{\mathrm{cov}}(\beta, QP) \cdot \sigma_{be}, \quad \beta = \sqrt{2/\sigma_{be}} \quad (11)$$

where α and β denote the Laplacian parameters. QP is the quantization parameter in H.264/AVC. M_{var} and M_{cov} indicate the scaling maps, which vary from 0 to 1.

The scaling maps M_{var} and M_{cov} are modeled based on four video sequences [27]: 'Crow_run', 'In_to_tree', 'Ducks_take_off', and 'Old_town_cross', which have abundant and various structural information. Each sequence is coded as intra-frame (I frame) and inter-frame (P frame), respectively. To cover various reconstruction variances, 11 different QP values are tested, ranging from 15 to 45 uniformly with the step size of 3.

Firstly, we calculate the variance of initial and reconstructed prediction residual with different QP values. Secondly, we obtain the scaling curve by doing statistics analysis for each test QP. Finally, we interpolate the eleven scaling curves to establish the scaling map. Based on the simulations, the fitted scaling map M_{var} and M_{cov} are shown in Figure 3, which can be constructed as look-up tables.

To demonstrate the accuracy of the proposed reconstruction quality models, 250 frames of each sequence are coded with constant quantization parameters: 20, 25, 30, and 35, respectively. Table 1 shows the average mean absolute deviation (MAD) between the actual and estimated variance, cross covariance, and reconstruction quality. The first two terms denote the accuracy of the fitted models (10) and (11), respectively. The following three terms show the accuracy of the final estimated SSIM scores. It can be seen that the proposed models are valid to predict the reconstruction quality.

3.2 Development of error propagation model

The error propagation is the key component of the end-to-end distortion model. Different from the independent estimation in conventional MSE-based end-to-end distortion model, the perceptual error propagation depends on the source distortion or the concealment distortion. In this section, our primary goal is to develop the error propagation models to estimate the overall perceptual quality for given transmission errors of prediction block or concealment block.

The error propagation models are motivated by three observations. The first observation is related to the impact of error propagation on the three components of SSIM. Let Q_{att} denote the quality attenuation of a given block b due to the error propagation.

$$Q_{\mathrm{att}}(b) = \mathrm{SSIM}\left(b, \tilde{b}\right) / \mathrm{SSIM}\left(b, \hat{b}\right) \quad (12)$$

To illustrate the fact, Q_{att} is measured by three different similarity metrics: 1) luminance component of SSIM, 2) the contrast and structure components of SSIM, 3) SSIM index. Figure 4 illustrates an example where the quality attenuation is calculated by Equation 12 for each frame suffering from random transmission errors. As shown, the contrast and structure components have the similar

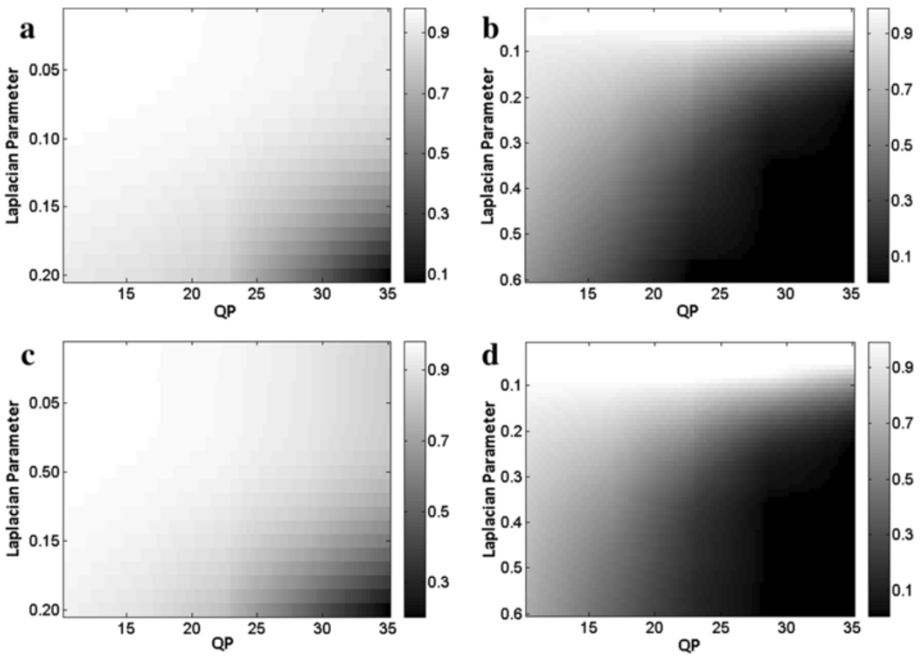

Figure 3 The scaling maps. (a) For residual variance by intra-coding; **(b)** for residual variance by inter-coding; **(c)** for residual cross covariance by intra-coding; **(d)** for residual cross covariance by inter-coding.

changes with SSIM. On the other hand, the impact of error propagation on the luminance component is limited.

The second observation is made on the relationship f_p between the quality attenuation of block b and the quality attenuation of its compensation block p. p indicates the compensation block of b, which may contain the transmission errors. Thus, p can be used to represent the prediction block r or concealment block c of b.

$$f_p = Q_{\text{att}}(b_p)/Q_{att}(p) = \frac{\text{SSIM}(b,\tilde{p})}{\text{SSIM}(b,\hat{p})} / \frac{\text{SSIM}(p,\tilde{p})}{\text{SSIM}(p,\hat{p})} \quad (13)$$

Usually, the quality attenuation of block b correlates with the quality attenuation of its compensation block p. In addition, the structural similarity between current block and its compensation block may be another factor in estimation of f_p. x_p is defined as follows to explore the effect of quality attenuation and structural similarity on f_p.

$$x_p = Q_{\text{att}}(p) \cdot \text{SSIM}(b,\hat{p}) \quad (14)$$

The simulation results are carried out on the same four sequences as Section 3.1. Each sequence is coded with four different QPs: 15, 25, 35, and 45. One I frame followed by all inter frames (IPPP). To cover various error propagation, each block is tested with random transmission errors propagated from prediction block and concealment block, respectively. Note that the prediction residuals of block b are not included in this observation.

Figure 5a displays the simulation results. The mean of each test frame is recorded as one blue sample, and the fitted curve of f_p (x_p) is shown as the red line. It shows that the quality attenuation of block, in terms of SSIM, is

Table 1 MAD between actual and estimated scores

Test terms	$\sigma^2_{\hat{e}}$	$\sigma_{b\hat{e}}$	I	$c \cdot s$	D_{SSIM}
MAD (Intra-coding)	5.8118	3.4852	0.0001	0.0067	0.0067
MAD (Inter-coding)	2.4700	3.0642	0.00004	0.0056	0.0056

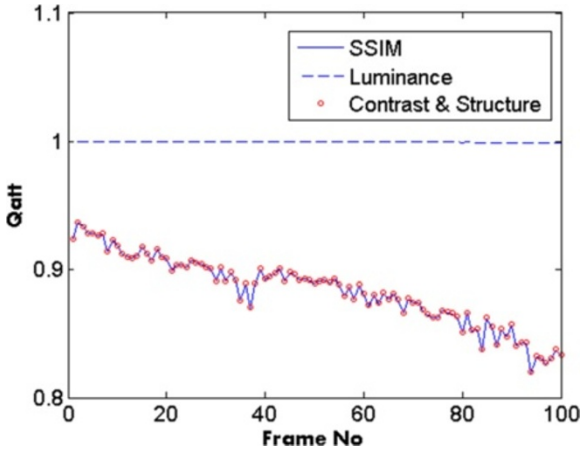

Figure 4 The quality attenuation due to transmission errors.

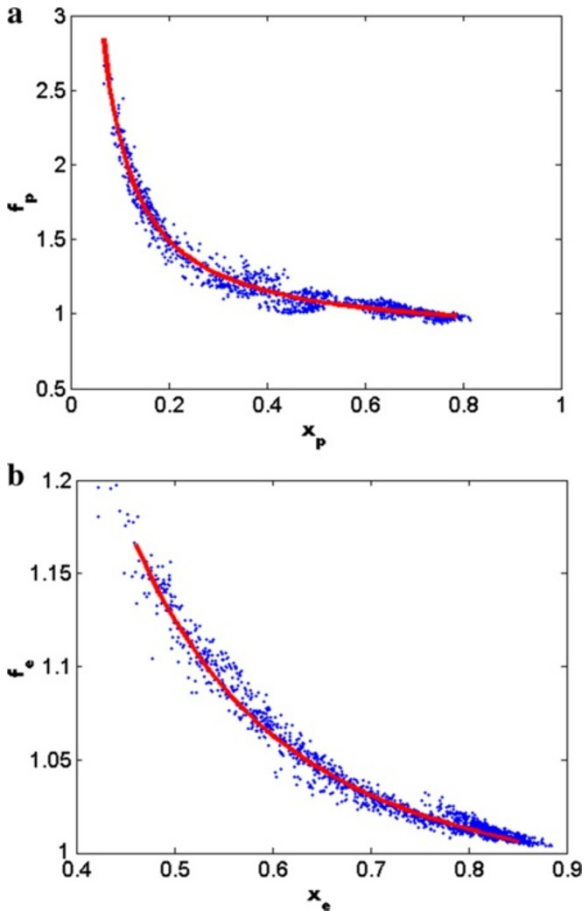

Figure 5 Quality attenuation due to error propagation. (a) The relationship between f_p and x_p; **(b)** the relationship between f_e and x_e.

related with the quality attenuation of its compensation block and the structural similarity. Moreover, it demonstrates that less quality attenuation of compensation block or less structural similarity between current block and compensation block leads to less quality attenuation.

The third observation is related to the impact of prediction residual on the decoded quality. Let Q_{enh} denote the quality enhancement of a given block b due to its prediction residuals *e.* f_e represents the relationship between the quality attenuation of block b with and without the prediction residual.

$$Q_{\text{enh}}(b_{p,e}) = \text{SSIM}(b, p+e)/\text{SSIM}(b,p) \tag{15}$$

$$f_e = Q_{\text{att}}(b_{p,e})/Q_{att}(b_p)^{att} = \frac{\text{SSIM}(b,\tilde{p}+\hat{e})}{\text{SSIM}(b,\hat{p}+\hat{e})} \Big/ \frac{\text{SSIM}(b,\tilde{p})}{\text{SSIM}(b,\hat{p})} \tag{16}$$

In this observation, the quality attenuation of block b may link with the quality attenuation of its compensation block p and the quality enhancement of its prediction residuals *e.* x_e is defined as follows to explore the effect of quality attenuation and quality enhancement on f_p:

$$x_e = Q_{\text{att}}(b_p)/Q_{\text{enh}}(b_{\hat{p},\hat{e}}) \tag{17}$$

The simulation set-up is the same as that in the second observation. In this observation, each block including the prediction residuals is tested with random transmission errors propagated from its prediction blocks. Figure 5b shows the simulation results. The mean of each test frame is recorded as one blue sample, and the fitted curve of f_e (x_e) is shown as the red line. The results show that a larger prediction residual leads to a better decoded quality of current block, and the influence of error propagation from prediction blocks will be smaller.

According to Equations 13 and 16, the effective approximation of $\text{SSIM}(b, \tilde{r}+\hat{e})$ and $\text{SSIM}(b,\tilde{c})$ can be developed as

$$\begin{aligned}\text{SSIM}(b,\tilde{r}+\hat{e}) &\approx f_e(x_e)\cdot\text{SSIM}(b,\hat{r}+\hat{e})\cdot\frac{\text{SSIM}(b,\tilde{r})}{\text{SSIM}(b,\hat{r})}\\ &\approx f_e(x_e)\cdot\text{SSIM}(b,\hat{r}+\hat{e})\cdot f_p(x_r)\cdot\frac{\text{SSIM}(r,\tilde{r})}{\text{SSIM}(r,\hat{r})}\\ &\approx Q_{\text{att}}(r)\cdot f_e(x_e)\cdot f_p(x_r)\cdot\text{SSIM}(b,\hat{r}+\hat{e})\end{aligned} \tag{18}$$

$$\begin{aligned}\text{SSIM}(b,\tilde{c}) &\approx f_p(x_c)\cdot\text{SSIM}(b,\hat{c})\cdot\frac{\text{SSIM}(c,\tilde{c})}{\text{SSIM}(c,\hat{c})}\\ &\approx Q_{\text{att}}(c)\cdot f_p(x_c)\cdot\text{SSIM}(b,\hat{c})\end{aligned} \tag{19}$$

Where $\text{SSIM}(r,\tilde{r})$ and $\text{SSIM}(c,\tilde{c})$ represent the end-to-end distortion of prediction block and concealment block, respectively. The approximations used in the equations represent the estimation of f_p and f_e.

Based on Equations 18 and 19, the error propagation factors in Equations 5 and 6 can be obtained by

$$\Phi_r = Q_{\text{att}}(r)\cdot f_p(x_r)\cdot f_e(x_e) \tag{20}$$

$$\Phi_c = Q_{\text{att}}(c)\cdot f_p(x_c) \tag{21}$$

To better demonstrate the accuracy of the proposed error propagation models, Table 2 shows the average MAD between the actual and estimated SSIM of the four test sequences. It indicates that video quality at the

Table 2 MAD between actual and estimated SSIM

Test sequences	Error propagation from concealment block	Error propagation from prediction block
Crow_run	0.0213	0.0238
In_to_tree	0.0112	0.0137
Ducks_take_off	0.0126	0.0140
Old_town_cross	0.0153	0.0173
Average	0.0151	0.0172

Table 3 MAD between actual and estimated end-to-end distortion

Test sequences	[16]	[19]	Proposed
Park_joy	0.15784	0.01182	0.00880
Blue sky	0.09282	0.00779	0.00780
Mobile calendar	0.12679	0.01444	0.01209
Pedestrian area	0.24684	0.02055	0.01561
Park run	0.19973	0.02865	0.04162
River bed	0.09601	0.00797	0.00560
Rush hour	0.14633	0.01719	0.00762
Sunflower	0.17936	0.01888	0.01078
Shields	0.10850	0.01320	0.00790
Station	0.18221	0.02256	0.01231
Average	0.153643	0.016305	0.013013

decoder, in terms of SSIM, can be approximately calculated by the fitted models (20) and (21) at the encoder.

4 Error-resilient video coding

It is widely recognized that intra-update is an effective approach for error-resilient video coding because decoding of an intra-coding block does not require information from its previous frames. To better evaluate the performance of our proposed model, we incorporate the proposed SSIM-based end-to-end distortion model into the mode selection to improve the RD performance over packet-switched networks. Thus, the optimization problem in Equation 1 can be converted to the problem of mode selection between intra-coding and inter-coding as follows:

$$\min\{J(\text{mode})\} = \min\{D_{\text{SSIM}}(\text{mode}|\rho,\text{QP}) + \lambda_{\text{SSIM}}\cdot R(\text{mode}|\text{QP})\} \quad (22)$$

where D_{SSIM} and R denote the end-to-end distortion and bit-rate of current coding block. ρ is the packet loss ratio. mode denotes the coding mode. QP is the quantization parameter, which is determined by the target bit rate. According to [8,28,29], the Lagrange multiplier λ_{SSIM} is determined as follows:

$$\lambda_{\text{SSIM}} = (1-\rho)\cdot\overline{D_{\text{SSIM}}}\cdot f(R_T) \quad (23)$$

where $\overline{D_{\text{SSIM}}}$ denotes the average distortion of previous coding units. $f(R_T)$ is an look-up experimental function [20], which is inversely proportional to the target bit rate R_T.

5 Experimental results

5.1 Evaluation of end-to-end distortion model

To validate the effectiveness of our proposed models, the end-to-end distortion models proposed in [16] and [19] are used as comparison. In [17], the SSIM-based

Table 4 Detail information of the test sequences

Test sequences	Resolution	Frame rate (fps)	Target bit rates (Kbps)			
Flower	352 × 288	30	2,000	1,300	700	400
Football	352 × 288	30	1,400	840	500	300
Mobile	352 × 288	30	2,800	1,600	800	380
Stefan	352 × 288	30	1,700	1,000	500	270
Bus	352 × 288	30	1,700	1,000	560	310
Crow_run	640 × 360	25	7,500	4,500	2,500	1,500
Park_joy	640 × 360	25	8,000	5,000	2,600	1,500
Ducks_take_off	640 × 360	25	7,000	4,500	2,400	1,200
In_to_tree	640 × 360	25	1,000	480	220	120
Old_town_cross	640 × 360	25	1,400	620	270	140
Blue sky	768 × 432	25	2,800	1,400	630	320
Mobile calendar	768 × 432	50	7,400	3,400	1,300	560
Pedestrian area	768 × 432	25	1,800	1,100	600	400
Park run	768 × 432	50	15,200	8,800	4,600	2,400
River bed	768 × 432	25	8,700	5,500	3,200	1,800
Rush hour	768 × 432	25	1,500	850	470	280
Sunflower	768 × 432	25	1,200	650	380	220
Shields	768 × 432	50	5,000	2,400	1,100	560
Station	768 × 432	25	850	440	250	160
Tractor	768 × 432	25	4,100	2,200	1,200	690

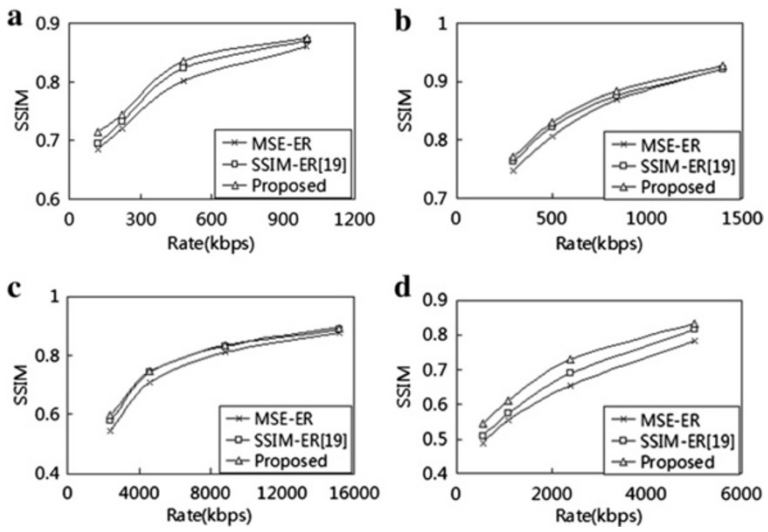

Figure 6 Comparison of Rate-SSIM performance. (a) 'Ducks_take_off' by with 5% packet loss. **(b)** 'Football' by with 10% packet loss. **(c)** 'Park run' by with 10% packet loss. **(d)** 'Shields' by with 20% packet loss.

estimation model is not suitable for a block given different coding modes. Thus, we did not compare the performance with the proposed model in [17].

We performed the simulation with ten LIVE sequences [21,22]. First 100 frames of each sequence are encoded by four different QPs: 24, 28, 32, and 36, respectively. Random packet losses (10% and 20%) are used. Each experiment is repeated 200 times and the results are averaged. Table 3 shows the average MAD between the actual and estimated end-to-end distortion, in terms of

Table 5 Simulation results of SSIM again with different packet loss ratios and bit rates

Test sequences	Packet loss ratio: 20%		Packet loss ratio: 10%		Packet loss ratio: 5%	
	[19]	Proposed	[19]	Proposed	[19]	Proposed
Flower	0.0234	0.0255	0.0143	0.0182	0.0054	0.0154
Football	0.0093	0.0171	0.0132	0.0167	0.0173	0.0174
Mobile	0.0185	0.0210	0.0128	0.0279	0.0112	0.0198
Stefan	0.0192	0.0294	0.0233	0.0334	0.0220	0.0299
Bus	0.0312	0.0319	0.0387	0.0428	0.0384	0.0447
Crow_run	0.0134	0.0174	0.0109	0.0168	0.0091	0.0141
Park_joy	0.0229	0.0241	0.0212	0.0240	0.0148	0.0213
Ducks_take_off	0.0056	0.0042	0.0001	0.0039	−0.0038	0.0026
In_to_tree	0.0200	0.0280	0.0163	0.0266	0.0129	0.0248
Old_town_cross	0.0072	0.0170	0.0038	0.0142	0.0026	0.0120
Blue sky	0.0131	0.0309	0.0197	0.0366	0.0121	0.0311
Mobile calendar	0.0272	0.0239	0.0222	0.0138	0.0231	0.0125
Pedestrian area	0.0040	−0.0029	0.0055	−0.0025	0.0060	0.0013
Park run	0.0318	0.0375	0.0242	0.0332	0.0129	0.0213
River bed	0.0069	0.0062	0.0058	0.0065	0.0049	0.0061
Rush hour	0.0083	0.0043	0.0062	0.0046	0.0050	0.0045
Sunflower	0.0363	0.0554	0.0258	0.0560	0.0292	0.0566
Shields	0.0262	0.0589	0.0295	0.0512	0.0278	0.0313
Station	0.0039	0.0097	−0.0004	0.0076	0.0007	0.0198
Tractor	0.0147	0.0179	0.0088	0.0166	0.0037	0.0157
Average	0.01715	0.02287	0.01509	0.02240	0.01276	0.02011

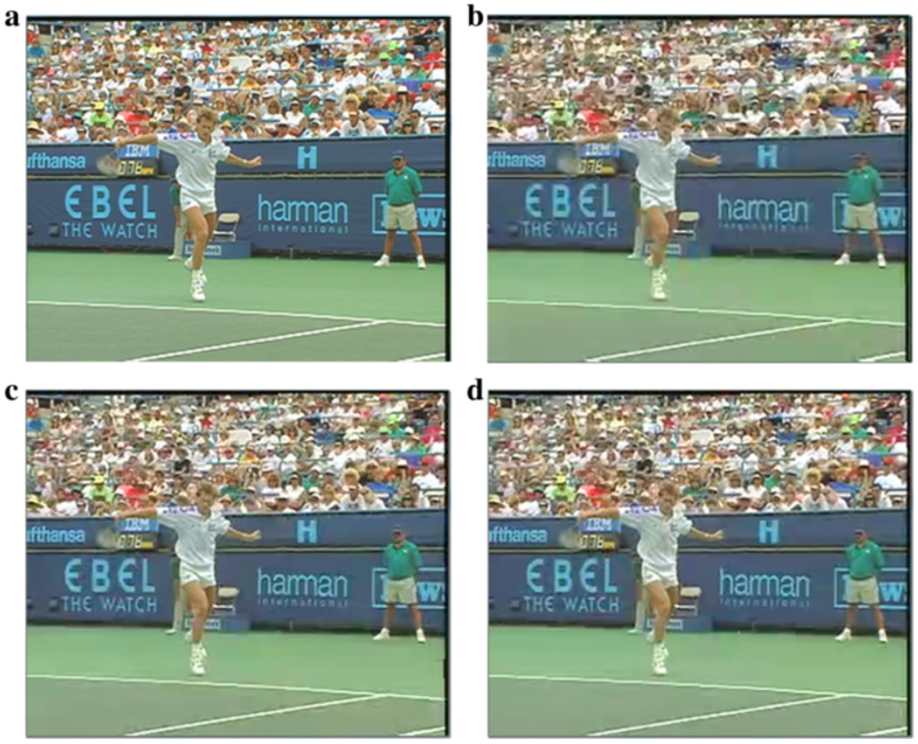

Figure 7 Subjective quality comparison of one CIF sequence. (a) Original frame of 'Stefan'; **(b)** 'Stefan' by 'MSE-ER' (SSIM 0.913); **(c)** 'Stefan' by SSIM-ER [19] (SSIM 0.917); **(d)** 'Stefan' by our proposed (SSIM 0.929).

SSIM. It is obvious that our proposed model achieves better performance for most sequences.

5.2 Evaluation of RD performance

To validate the RD performance of error-resilient video coding, the MSE-based error resilient video coding scheme (MSE-ER) and the SSIM-based error-resilient video coding scheme (SSIM-ER) in [19] are used as the comparison schemes. For MSE-ER, the end-to-end distortion is estimated by the ROPE model [3], which is well studied and regarded as an advanced MSE-based distortion model, and the

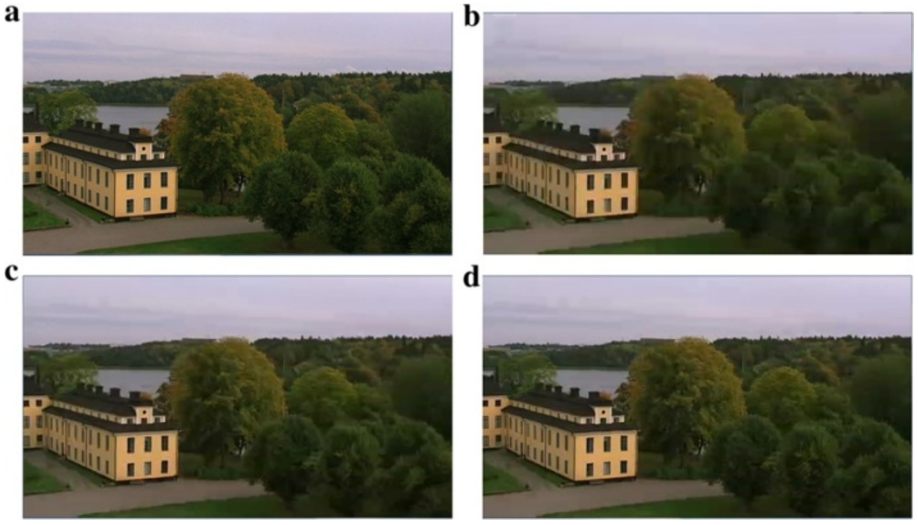

Figure 8 Subjective quality comparison of one 640 × 360 sequence. (a) Original frame of 'In_to_tree'; **(b)** 'In_to_tree' by MSE-ER (SSIM 0.833); **(c)** 'In_to_tree' by SSIM-ER [19] (SSIM 0.860); **(d)** 'In_to_tree' by our proposed (SSIM 0.873).

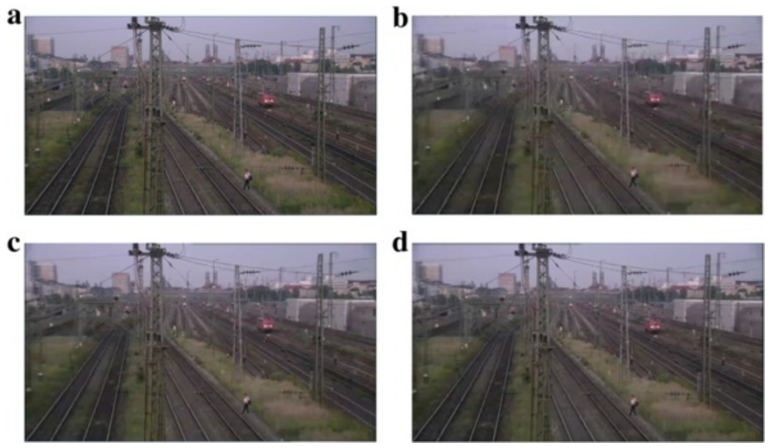

Figure 9 Subjective quality comparison of one LIVE sequence. (a) Original frame of 'Station'; **(b)** 'Station' by MSE-ER (SSIM 0.826); **(c)** 'Station' by SSIM-ER [19] (SSIM 0.851); **(d)** 'Station' by our proposed (SSIM 0.869).

Lagrange multiplier is calculated by the model presented in [8].

We evaluate the performance on the platform of JM 15.1 [30], in which the SSIM index is adopted as an optimal quality metric. Five CIF sequences [27], five 640 × 360 sequences [27] and ten LIVE sequences [21,22] are tested in the experiments. The first frame is coded as I frame and the rest are coded as P frames. The rate control is turned on. Table 4 shows the detail information of target bit rates for test sequences. Corresponding to four different target bit rates, the initial QP is equal to 24, 28, 32, and 36 for the first I frame and P frame. Frames are partitioned into one or more slices (each slice contains no more than 1,200 bytes), and the slices are organized in packets for transmission where each slice is packed into one packet. The test sequences are encoded with 5%, 10%, and 20% random packet loss ratio, respectively. For each packet loss ratio, four different target bit rates are tested in the experiments. Each experiment is repeated by 200 times, and the results are averaged, which are used as the final result.

Figure 6 illustrates the results of Rate-SSIM performance comparison for four test sequences. Moreover, we choose the MSE-ER scheme as the baseline and calculate all the simulation results of average SSIM gain with different bit rates and packet loss ratios, which are tabulated in Table 5.

It can be seen that the proposed model yields consistent gains over the MSE-ER for all sequences except 'Pedestrian area'. Our proposed scheme achieves an average SSIM gain of 0.0218 or equivalently a bit rate saving of 15.7%. Comparing to the SSIM-ER [19], our proposed scheme has better performance of most sequences and obtains an average gain of 0.0068. For some sequences, such as 'Mobile calendar' and 'Pedestrian area', although our proposed scheme cannot achieve the best performance, the quality of the two SSIM-ER schemes is similar.

Table 6 Average encoding time ratio of SSIM-ER [19] and proposed scheme to MSE-ER, respectively

Test sequences	[19]	Proposed	Test sequences	[19]	Proposed
Flower	2.77%	4.63%	Blue sky	4.09%	6.97%
Football	3.70%	4.99%	Mobile calendar	6.23%	9.77%
Mobile	4.37%	6.95%	Pedestrian area	3.89%	3.66%
Stefan	3.60%	4.78%	Park run	2.72%	4.85%
Bus	5.33%	7.13%	River bed	2.64%	2.21%
Crow_run	1.39%	3.98%	Rush hour	3.99%	4.46%
Park_joy	2.24%	3.22%	Sunflower	6.33%	7.54%
Ducks_take_off	2.40%	2.76%	Shields	5.19%	7.43%
In_to_tree	1.91%	6.57%	Station	3.70%	7.70%
Old_town_cross	3.45%	6.90%	Tractor	2.98%	5.23%
Average	3.12%	5.19%	-	4.18%	5.98%

5.3 Evaluation of subjective quality

Finally, we show the visual quality comparison of reconstructed images by different error-resilient video coding schemes. Figure 7 compares the subjective quality of the 25th frame of 'Stefan' encoded at 1.7 Mbps with 10% packet loss. Figure 8 shows the comparison on visual quality of the 38th frame of 'In_to_tree' encoded at 1 Mbps with 20% packet loss. Figure 9 represents the visual quality of the 29th frame of 'Station' encoded at 0.85 Mbps with 5% packet loss.

For the similar bit rate, the reconstructed images based on the SSIM-based error-resilient video coding can provide a better visual quality due to more image details being protected from transmission errors. On the other side, the reconstructed images based on the conventional MSE-based error-resilient video coding suffer from larger perceptual distortion. Compared to SSIM-ER [19], our proposed scheme obtains similar or better visual quality.

5.4 Evaluation of coding complexity

Our proposed SSIM-based error resilient video coding scheme improves the RD performance for lossy transmission over packet-switched network. However, the computational complexity of codec is increased due to SSIM-based distortion calculation and mode selection.

We compare the coding efficiency with different bit rates and packet loss ratios. Table 6 shows the average encoding time ratio of SSIM-ER [19] and proposed scheme to MSE-ER, respectively. The experiments are performed on a laptop with 3.4 GHz Intel Core i7-3770 CPU and 4G memory running on Microsoft Windows 7 professional platform. Each experiment is repeated 100 times and the results are averaged.

Comparing to MSE-ER, the average computation of SSIM-ER [19] and proposed scheme increase by 3.65% and 5.58%, respectively. In addition, different sequences have inconsistent degree of encoding time, as can be seen in Table 6. That is because the computation complexity is also affected by the characteristics of video content and the results of mode selection. The SSIM-based schemes may take more time to code the image details, such as 'Sunflower' and 'Mobile'.

6 Conclusions

In this paper, we propose an SSIM-based end-to-end distortion model for H.264/AVC video coding over packet-switched networks. This model is useful to estimate the content-dependent perceptual distortion of quantization, error concealment, and error propagation. We integrate the proposed end-to-end distortion model into the error-resilient video coding framework to optimally select the coding mode. Simulation results show that the proposed scheme outperforms the state-of-the-art schemes in terms of SSIM.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The work described in this paper was supported by the National Natural Science Foundation of China (No. 60972111, 61036008, 61071184, 61373121), Research Funds for the Doctoral Program of Higher Education of China (No. 20100184120009, 20120184110001), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029, 13QNJJ0149), the Fundamental Research Funds for the Central Universities (Project no. SWJTU09CX032, SWJTU10CX08, SWJTU11ZT08), and Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Author details
[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. [2]Institution of Academia Sinica, Jiuzhou Electric Group, Sichuan, China. [3]School of Electronics and Information Engineering, Sichuan University, Chengdu, China.

Received: 11 December 2013 Accepted: 25 August 2014
Published: 13 September 2014

References

1. S Wenger, H.264/AVC over IP. IEEE Transact Circ Syst Video Technol **13**, 645–656 (2003)
2. ZH He, JF Cai, CW Chen, Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding. IEEE Transact Circ Syst Video Technol **12**, 511–523 (2002)
3. R Zhang, SL Regunathan, K Rose, Video coding with optimal inter/intra-mode switching for packet loss resilience. IEEE J Selected Areas Commun **18**, 966–976 (2000)
4. XK Yang, C Zhu, ZG Li, X Lin, GN Feng, S Wu, N Ling, Unequal loss protection for robust transmission of motion compensated video over the Internet. Signal Process. Image Commun. **18**, 157–167 (2003)
5. CY Zhang, H Yang, SY Yu, XK Yang, GOP-level transmission distortion modeling for mobile streaming video. Signal Process. Image Commun. **23**, 116–126 (2008)
6. Y Wang, ZY Wu, JM Boyce, Modeling of transmission-loss-induced distortion in decoded video. IEEE Transact Circ Syst Video Technol **16**, 716–732 (2006)
7. H Yang, K Rose, Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H.264/AVC. IEEE Transact Circ Syst Video Technol **17**, 845–856 (2007)
8. Y Zhang, W Gao, Y Lu, Q Huang, D Zhao, Joint source-channel rate-distortion optimization for H.264 video coding over error-prone networks. IEEE Transact Multimed **9**, 445–454 (2007)
9. H Yang, K Rose, Optimizing motion compensated prediction for error resilient video coding. EURASIP J Image Video Process **19**, 108–118 (2010)
10. JM Xiao, T Tillo, CY Lin, Y Zhao, Joint redundant motion vector and intra macroblock refreshment for video transmission. EURASIP J Image Video Process **12**, (2011). doi: 10.1186/1687-5281-2011-12
11. YX Zhang, C Zhu, KH Yap, A joint source-channel video coding scheme based on distributed source coding. IEEE Transact Multimed **10**, 1648–1656 (2008)
12. Y Zhou, CP Hou, W Xiang, F Wu, Channel distortion modeling for multi-view video transmission over packet-switched networks. IEEE Transact Circ Syst Video Technol **21**, 1679–1692 (2011)
13. JM Xiao, T Tillo, CY Lin, Y Zhao, Dynamic sub-GOP forward error correction code for real-time video applications. IEEE Transact Multimed **14**, 1298–1308 (2012)
14. Z Xue, KK Loo, J Cosmas, M Tun, PY Yip, Error-resilient scheme for wavelet video coding using automatic ROI detection and Wyner-Ziv coding over packet erasure channel. IEEE Transact Broadcast **56**, 481–493 (2010)
15. MB Dissanayake, S Worrall, WAC Fernando, Error resilience for multi-view video using redundant macroblock coding, in *Proceedings of the IEEE International Conference on Industrial Information Systems (ICIIS)* (University of Peradeniya, Kandy, 2011), pp. 472–476
16. YX Wang, Y Zhang, R Lu, PC Cosman, SSIM-Based End-to-End Distortion Modeling for H.264 Video Coding, in *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)* (Singapore, 2012), pp. 117–128

17. YJ Kwon, J-S Lee, Parametric estimation of structural similarity degradation for video transmission over error-prone networks. Electron. Lett. **49**, 1147–1148 (2013)
18. L Zhang, Q Peng, X Wu, SSIM-based Error-resilient video coding over packet-switched, in *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)* (Singapore, 2012), pp. 263–272
19. PH Zhao, YW Liu, JX Liu, S Li, RX Yao, SSIM-based error-resilient rate-distortion optimization of H.264/AVC video coding for wireless streaming. Signal Process. Image Commun. **29**, 303–315 (2014)
20. L Zhang, Q Peng, X Wu, SSIM-based end-to-end distortion model for error resilient video coding over packet-switched networks, in *Proceedings of the International Conference on Multimedia Modeling (MMM)* (Huangshan, 2013), pp. 307–317
21. K Seshadrinathan, R Soundararajan, AC Bovik, LK Cormack, Study of subjective and objective quality assessment of video. IEEE Transact Image Process **19**, 1427–1441 (2010)
22. *Live Video Quality Database*, 2012. http://www.utexas.edu/ece/research/live/vqdatabase/
23. Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Transact Image Process **13**, 600–612 (2004)
24. WS Lin, CCJ Kuo, Perceptual visual quality metrics: a survey. J Visual Commun Image Represent **22**, 297–312 (2011)
25. E Lam, J Goodman, A mathematic analysis of the DCT coefficient distribution for images. IEEE Transact Image Process **9**, 1661–1666 (2000)
26. TW Yang, C Zhu, XJ Fan, Q Peng, Source distortion temporal propagation model for motion compensated video coding optimization, in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (Melbourne, 2012), pp. 85–90
27. *Xiph.org Video Test Media*, 2010. http://media.xiph.org/video/derf/
28. TS Ou, YH Huang, HH Chen, SSIM-based perceptual rate control for video coding. IEEE Transact Circ Syst Video Technol **21**, 682–691 (2011)
29. T Wiegand, B Girod, Lagrange multiplier selection in hybrid video coder control, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (Thessaloniki, Thessaloniki, 2001), pp. 542–545
30. *JVT Reference Software*, 2011. http://iphome.hhi.de/suehring/tml/

doi:10.1186/1687-5281-2014-45
Cite this article as: Peng *et al.*: **Modeling of SSIM-based end-to-end distortion for error-resilient video coding.** *EURASIP Journal on Image and Video Processing* 2014 **2014**:45.