Signal Processing: *Image Communication* 26 (2011) 24–38

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Evaluation of temporal variation of video quality in packet loss networks

Changhoon Yim [a,*], Alan C. Bovik [b]

[a] *Department of Internet and Multimedia Engineering, Konkuk University, Seoul, Republic of Korea*
[b] *Department of Electrical and Computer Engineering, The University of Texas at Austin, TX 78712, USA*

ARTICLE INFO

Article history:
Received 9 July 2010
Accepted 3 November 2010

Keywords:
Video communication
Wireless network
Channel error
Quality assessment
Quality variation

ABSTRACT

We examine the effect that variations in the temporal quality of videos have on global video quality. We also propose a general framework for constructing temporal video quality assessment (QA) algorithms that seek to assess transient temporal errors, such as packet losses. The proposed framework modifies simple frame-based quality assessment algorithms by incorporating a temporal quality variance factor. We use packet loss from channel errors as a specific study of practical significance. Using the PSNR and the SSIM index as exemplars, we are able to show that the new video QA algorithms are highly responsive to packet loss errors.

1. Introduction

Next-generation wireless networks promise to provide increased bandwidth that will greatly enhance video transmission applications, yet they will present new challenges due to packet losses from channel errors [1,2]. Most video coding standards, such as MPEG-2 and H.264, are based on motion-compensated prediction coding, which causes severe error propagation effects from packet losses. Since even single bit errors may cause packet losses and since bit errors are quite common in wireless networks, compressed video transmissions may be expected to be subject to severe degradation in quality.

Joint source-channel rate-distortion (R-D) models of video transmission over packet-lossy networks are a topic of lively investigation [1–5]. Most research in this context has been directed towards reducing the effects of packet loss. Several R-D optimized methods have been proposed for coding mode selection in packet-lossy networks [1,6,7]. A variety of unequal error protection methods have also been investigated for video transmission over packet-lossy networks using an R-D framework [8–11].

Video is a sequence of images. In most existing R-D optimized methods for video transmission, the video distortion is assumed to be the sample mean of the image (frame) distortions. Such methods implicitly assume that the video quality is the mean of the image qualities.

Recently, image and video quality assessment has received a great deal of attention [12–28]. Some video quality indices modify still image quality indices by the addition of temporal filtering, as in [21]. Lowpass filtering along the temporal dimension was utilized in the digital video quality (DVQ) metric [22] and in a scalable wavelet based video distortion index [23]. A general framework for measuring spatial and temporal video distortions along motion trajectories was proposed in the MOtion-based Video Integrity Evaluation (MOVIE) index [24]. The MOVIE index integrates explicit motion information with the measurement of spatial artifacts in the video quality assessment (VQA) process. In [26], a foveated signal-to-noise ratio was defined that assigns different weights to foveated parts of images. A quality evaluation approach was proposed for blocking, blurring, ringing, and motion-compensated edge artifacts in [27]. A harmonic strength analysis model computed from edge-detected features was

* Corresponding author. Tel.: +82 2 450 4016; fax: +82 2 458 1997.
E-mail addresses: cyim@konkuk.ac.kr (C. Yim), bovik@ece.utexas.edu (A.C. Bovik).

0923-5965/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2010.11.002

proposed for reduced-reference video quality assessment in [28]. All these methods define the video quality index to be the mean of image quality metrics across image frames.

Quality *variations* in the temporal direction have been largely ignored in previous video quality assessment work. Here, we argue that the temporal variation of quality needs to be explicitly included. Variations in quality in a video sequence can be very annoying, possibly worse than a constant quality video with lower mean value.

Packet losses arising from channel errors in wireless networks result in significant distortions. Indeed, channel distortions can be much more annoying than source distortions arising from quantization errors. Since packet losses occur randomly in a channel, the temporal variation of the induced channel distortion can be quite large, and so the temporal quality variations arising from channel errors can also be large. Based on these observations, we begin by studying the degree of distortion and quality variation that occurs in video from packet losses. We subsequently propose a new framework for video quality measurement that seeks to account for quality variations arising from channel errors.

The H.264 video coding standard has demonstrated an excellent compression efficiency compared to previous video coding standards [29], and has features that are suitable for wireless video applications [30,31]. We utilize the H.264 video coding standard in our investigation of temporal distortions and quality variations that occur in video transmission over wireless channels.

The organization of this paper is as follows. Section 2 investigates the nature of temporal distortion and quality variation that occurs in video. In Section 3, we propose a new video quality index that accounts for these variations. Section 4 presents simulation results for video transmission over wireless networks. In Section 5, we present conclusion and future work.

2. Distortion and quality variation of video

A video is a sequence of frames (images), $\mathcal{V}=\{I_n, n=0,1,\ldots,K-1\}$ where K is the number of frames. Given an original frame I_n, let $\hat{I}_n$ be the reconstructed frame in the encoder feedback loop, and $\tilde{I}_n$ be the reconstructed frame at the decoder.

Let $D_s(n)$ and $D_c(n)$ be the source and channel distortion, respectively. The most common distortion measure is the mean square error (MSE). The source distortion $D_s(n)$ represents the distortion between I_n and $\hat{I}_n$, while the channel distortion $D_c(n)$ represents the distortion between $\hat{I}_n$ and $\tilde{I}_n$. If there is no channel error between the encoder and the decoder, then $\tilde{I}_n=\hat{I}_n$ and the channel distortion $D_c(n)=0$. Let $D(n)$ be the end-to-end distortion between the original frame I_n and the reconstructed frame $\tilde{I}_n$ at the decoder. Assume that the source distortion $D_s(n)$ and the channel distortion $D_c(n)$ are uncorrelated as in [1,2]

$$D(n)=D_s(n)+D_c(n). \tag{1}$$

In [1,2], the channel distortions are modeled as statistical expectation in terms of packet loss rate (PLR). Since the purpose of these distortion models are for improving rate-distortion (R-D) behavior in the context of joint source-channel coding, the distortion models also contain many source and channel coding parameters including source/channel code ratio and intra-mode macroblock ratio [1,2]. In previous distortion models [1–11], the distortion values are estimated *a priori* in terms of PLR and many source-channel coding parameters. In the distortion model used in this paper, the source and channel distortion values are obtained *a posteriori* with known packet loss frame indices. Hence the purpose of the distortion model in this paper is not to propose any source, channel, or joint source-channel coding scheme that considers a specific statistically variable channel, but to investigate source and channel distortion behavior posteriorly as a case study for the analysis of its relationship with video quality, especially with temporal quality variation.

The channel distortion can be modeled as the sum of two distortions [11]: the error concealment distortion $D_{ec}(n)$ and the error propagation distortion $D_{ep}(n)$

$$D_c(n)=D_{ec}(n)+D_{ep}(n). \tag{2}$$

The error concealment distortion $D_{ec}(n)$ is generated by a packet loss in frame n. The error concealment operation restores the lost information using the received information, but would result in restoration error and distortion. The error propagation distortion $D_{ep}(n)$ results from packet loss in any preceding frame within the current group-of-pictures (GOP), and is propagated through a motion-compensated prediction path in inter-frame coding. Let $D_{ep}(m\to n)$ represent the error propagation distortion in frame n from the packet loss in frame m ($m<n$).

$$D_{ep}(m\to n)=\alpha^{n-m}D_{ec}(m), \tag{3}$$

where α is an error propagation factor, which is generally less than 1, but might be greater than 1 in some cases. In [10], the error propagation distortion in a block-of-packets (BOP) is modeled as the average value of length of error propagation (LEP) and corresponds to $\alpha=1$. In [11], the factor α is modeled as a function of the intra-mode macroblock rate for the prior estimation of error propagation distortion. Here, the α values are calculated *a posteriori*, given the channel distortion values and packet loss frame indices.

The error would not be propagated through an I-frame, and the first frame of a GOP is an I-frame. Hence the error would not be propagated from the previous GOP into the current GOP. Let $o(n)$ be the I-frame index of frame n, which can be obtained as $o(n)=\lfloor n/N\rfloor\cdot N$ where N is the GOP size. Then $D_{ep}(n)$ would be the sum of the error propagation distortions from packet losses from the I-frame $o(n)$ through the previous frame indexed $(n-1)$:

$$D_{ep}(n)=\sum_{m=o(n)}^{n-1}D_{ep}(m\to n). \tag{4}$$

If there is no packet loss in frame $o(n)$ through frame n, then $D_{ec}(o(n))=\cdots=D_{ec}(n)=0$, $D_c(n)=0$, and $D(n)=D_s(n)$.

Let $f(a_i)$ be the frame number of packet a_i. Then

$$D_{ep}(f(a_i)\to n)=\alpha_{f(a_i)}^{n-f(a_i)}D_{ec}(a_i) \quad \text{for } o(n)\le f(a_i)<n<o(n)+N. \tag{5}$$

The term $D_{ec}(a_i)$ is the error concealment distortion in frame $f(a_i)$ resulting from the loss of packet a_i and $\alpha_{f(a_i)}$ is the

error propagation factor from frame $f(a_i)$. The error concealment distortion $D_{ec}(a_i)$ is dependent on many factors, such as spatial complexity and amount of motion, which can be quite different for each packet loss.

If packet a_i is lost, the channel distortions would be increased in frame $f(a_i)$ through frame $o(f(a_i))+N-1$. Let $D_c(n,a_i)$ be the channel distortion in frame n from the loss of packet a_i. Then

$$D_c(n,a_i) = D_{ep}(f(a_i)\rightarrow n), \tag{6}$$

which can be obtained as (5)

$$D_c(n,a_i) = \begin{cases} \alpha_{f(a_i)}^{n-f(a_i)} D_{ec}(a_i) & \text{if } o(n) \le f(a_i) < n < o(n)+N, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

If packet a_i is lost, the channel distortions are increased in frames $f(a_i)$ through $o(n)+N-1$ as in (7). Since the location $f(a_i)$ of a packet loss is random, and $D_{ec}(a_i)$ can be quite variable, the $D_{ec}(n, a_i)$ can take wide range of values.

Let $\mathcal{L}$ be the set of indices of lost packets and $\mathcal{C}(\mathcal{L})$ be the cardinality of $\mathcal{L}$. If the number of packets is M, the packet loss rate (PLR) is $\mathcal{C}(\mathcal{L})/M$. The channel distortion in frame n would be affected by any packet loss in frame $o(n)$ through n. Let $\mathcal{L}_n$ be the subset of $\mathcal{L}$ with indices i, such that $o(n) \le f(a_i) \le n$. The channel distortion in frame n from packet losses can then be modeled as the sum of the distortions from packet losses in $\mathcal{L}_n$

$$D_c(n) = \sum_{i\in\mathcal{L}_n} D_c(n,a_i) = \sum_{i\in\mathcal{L}_n} \alpha_{f(a_i)}^{n-f(a_i)} D_{ec}(a_i). \tag{8}$$

In this channel distortion model, the channel distortion $D_c(n)$ in (8) would be dependent on packet loss locations, error propagation distances, and error concealment distortions. Hence $D_c(n)$ can be quite variable, and can be much larger than the source distortion $D_s(n)$.

Actually the error propagation factor α is changing and can be different at each frame. Let α_n be the error propagation factor at frame n. Typically α_n values are similar for error propagation distortions from the same packet losses. If frame m is the last previous packet loss frame of frame n, then we can approximate α_n value as

$$\alpha_n \approx \alpha_{m+1}. \tag{9}$$

The α_{m+1} value can be calculated as

$$\alpha_{m+1} = \frac{D_c(m+1)}{D_c(m)}. \tag{10}$$

When there are packet losses in frame n and errors are propagated from previous frames, error concealment distortion and error propagation distortion are mixed in frame n as in (2). From (9) and (10), the error propagation distortion in frame n can be approximately calculated as

$$\begin{aligned} D_{ep}(n) &= \alpha_n D_c(n-1) \\ &\approx \alpha_{m+1} D_c(n-1), \end{aligned} \tag{11}$$

where frame m is the last previous packet loss frame of frame n. The error concealment distortion at frame n can be approximately calculated as

$$\begin{aligned} D_{ec}(n) &= D_c(n) - D_{ep}(n) \\ &\approx D_c(n) - \alpha_{m+1} D_c(n-1). \end{aligned} \tag{12}$$

For video sequence $\mathcal{V} = \{I_n, n = 0,1,\ldots,K-1\}$, let $\mathcal{S}(D_c)$, $\mathcal{S}(D_{ec})$, and $\mathcal{S}(D_{ep})$ be the sum of channel distortion, error concealment distortion, and error propagation distortion, respectively. The $\mathcal{S}(D_c)$ can be calculated as

$$\mathcal{S}(D_c) = \sum_{n=0}^{K-1} D_c(n). \tag{13}$$

The error concealment distortion occurs only in frames with lost packets. Let $\mathcal{F}$ be the set of frame indices that contain lost packets. Then

$$\mathcal{S}(D_{ec}) = \sum_{n\in\mathcal{F}} D_{ec}(n). \tag{14}$$

The $\mathcal{S}(D_{ep})$ can be obtained as

$$\mathcal{S}(D_{ep}) = \mathcal{S}(D_c) - \mathcal{S}(D_{ec}). \tag{15}$$

We define the error propagation distortion ratio ρ_{ep} as

$$\rho_{ep} = \frac{\mathcal{S}(D_{ep})}{\mathcal{S}(D_c)}. \tag{16}$$

We consider temporal distortion statistics of source distortion ($D_s(n)$), channel distortion ($D_c(n)$), and end-to-end distortion ($D(n)$) through frames (images) in video sequence. The most basic temporal distortion statistics are the mean and standard deviation of the frame distortions. Let $\mu(D_s)$ and $\sigma(D_s)$ be the mean and standard deviation of source distortions:

$$\mu(D_s) = \frac{1}{K}\mathcal{S}(D_s), \tag{17}$$

$$\sigma(D_s) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(D_s(n)-\mu(D_s))^2}. \tag{18}$$

Similarly, let $\mu(D_c)$ and $\sigma(D_c)$ be the mean and standard deviation of channel distortions:

$$\mu(D_c) = \frac{1}{K}\mathcal{S}(D_c), \tag{19}$$

$$\sigma(D_c) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(D_c(n)-\mu(D_c))^2}. \tag{20}$$

Let $\mu(D)$ and $\sigma(D)$ be the mean and standard deviation of end-to-end distortions:

$$\mu(D) = \frac{1}{K}\mathcal{S}(D), \tag{21}$$

$$\sigma(D) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(D(n)-\mu(D))^2}. \tag{22}$$

We define the channel distortion ratio ρ_c as

$$\rho_c = \frac{\mu(D_c)}{\mu(D)}. \tag{23}$$

Generally, image quality is a function Q of the image distortion: $Q(D(n))$. The well-known image quality index peak signal-to-noise ratio (PSNR) is a monotonically decreasing function of the MSE:

$$\mathrm{PSNR(MSE)} = 10\log_{10}\frac{255^2}{\mathrm{MSE}}. \tag{24}$$

Other quality indices such as SSIM [12], VIF [14], and VSNR [17] are more complex functions of the distortion, and are

designed to correlate well with visual perception of the distortion.

Let $Q_s(n)$ represent the quality of an image (frame) n suffering only from source distortions. $Q_s(n)$ could be the result of computing the frame PSNR, SSIM, or other quality index value

$$Q_s(n) = Q(D_s(n)). \tag{25}$$

Let $Q_{s+c}(n)$ represent the quality of frame n containing both source and channel distortions

$$Q_{s+c}(n) = Q(D_s(n)+D_c(n)) = Q\left(D_s(n)+\sum_{i\in\mathcal{L}_n}\alpha_{f(a_i)}^{n-f(a_i)}D_{ec}(a_i)\right). \tag{26}$$

The most basic temporal quality statistics are the mean and standard deviation of the frame (image) qualities. Let $\mu(Q_s)$ and $\sigma(Q_s)$ be the mean and standard deviation of the image frame qualities assuming only source distortion:

$$\mu(Q_s) = \frac{1}{K}\sum_{n=0}^{K-1} Q_s(n) = \frac{1}{K}\sum_{n=0}^{K-1} Q(D_s(n)), \tag{27}$$

$$\sigma(Q_s) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(Q(D_s(n))-\mu(Q_s))^2}. \tag{28}$$

If the source distortion $D_s(n)$ arises only from quantization, it depends on the quantization parameter (QP). If the QP is held constant over time, $D_s(n)$ may be expected to take similar values across frames, while the standard deviation of the frame qualities (28) will generally be small, although there will be some variation with content.

Likewise, let $\mu(Q_{s+c})$ and $\sigma(Q_{s+c})$ be the mean and standard deviation of frame quality assuming both source and channel distortions:

$$\mu(Q_{s+c}) = \frac{1}{K}\sum_{n=0}^{K-1} Q_{s+c}(n) = \frac{1}{K}\sum_{n=0}^{K-1}(Q(D_s(n)+D_c(n)), \tag{29}$$

$$\begin{aligned}\sigma(Q_{s+c}) &= \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(Q_{s+c}(n)-\mu(Q_{s+c}))^2} \\ &= \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}\left(Q\left(D_s(n)+\sum_{i\in\mathcal{L}_n}\alpha_{f(a_i)}^{n-f(a_i)}D_{ec}(a_i)\right)-\mu(Q_{s+c})\right)^2}.\end{aligned} \tag{30}$$

In (26), $Q_{s+c}(n)$ is a complex function of packet loss locations ($\mathcal{L}_n$), error propagation distances ($n-f(a_i)$), and error concealment distortions ($D_{ec}(a_i)$). Generally, $Q_{s+c}(n)$ can be expected to vary much more than $Q_s(n)$, hence $\sigma(Q_{s+c})$ will take much larger values than $\sigma(Q_s)$ if packet losses occur.

3. Video quality index accounting for variation

For video sequence $\mathcal{V}=\{I_n, n=0,1,\ldots,K-1\}$, let $Q(I_n)$ be the quality of frame (image) I_n, where Q is an image quality index such as PSNR or SSIM. If there are channel distortions resulting from packet losses as well as source distortions, then $Q(I_n)=Q_{s+c}(n)$.

Let $\mu(Q)$ be the mean of the frame qualities

$$\mu(Q) = \frac{1}{K}\sum_{n=0}^{K-1} Q(I_n), \tag{31}$$

which is a given frame-based approach to pooling video quality over time:

$$Q(\mathcal{V}) = \mu(Q). \tag{32}$$

In conventional applications such as broadcasting, where the bandwidth is guaranteed, the packet loss rate (PLR) is very low and the channel distortion is close to 0. Since variations in the source distortion would be relatively small in this application, the quality variance would also be small. However, in modern networked applications, such as video over the Internet and video over wireless, the bandwidth is not guaranteed. In wireless applications, the bandwidth fluctuates a lot, and with a large PLR when the bandwidth becomes small. In this case, the channel distortions can become quite large for some frames, resulting in a large quality variance.

Let $\sigma(Q)$ be the standard deviation of frame quality:

$$\sigma(Q) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(Q(I_n)-\mu(Q))^2}. \tag{33}$$

To capture the frame quality variation in videos, we propose a generalized video quality index that incorporates the frame quality standard deviation:

$$Q(\mathcal{V}) = \mu(Q)-w\sigma(Q), \tag{34}$$

where w weights the quality variation. This weight determines the decrement in objective quality as $\sigma(Q)$ increases.

If packet losses arise from channel errors, then $D_c(n)$ can take very large values, significantly decreasing $Q_{s+c}(n)$ over some frames, and increasing $\sigma(Q_{s+c})$ substantially. The proposed generalized temporal video quality index (34) efficiently represents both quality and quality variation.

If the image quality index for frame n is PSNR(n), we may define the temporal variance sensitive video quality index

$$\text{PSNR–TV} = \mu(\text{PSNR})-w\sigma(\text{PSNR}), \tag{35}$$

where

$$\mu(\text{PSNR}) = \frac{1}{K}\sum_{n=0}^{K-1}\text{PSNR}(n), \tag{36}$$

and

$$\sigma(\text{PSNR}) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(\text{PSNR}(n)-\mu(\text{PSNR}))^2}. \tag{37}$$

Similarly, if the image quality index for frame n is SSIM(n), we may also define the temporal variance sensitive video quality index

$$\text{SSIM-TV} = \mu(\text{SSIM})-w\sigma(\text{SSIM}), \tag{38}$$

where

$$\mu(\text{SSIM}) = \frac{1}{K}\sum_{n=0}^{K-1}\text{SSIM}(n), \tag{39}$$

and

$$\sigma(\mathrm{SSIM}) = \sqrt{\frac{1}{K}\sum_{n=0}^{K-1}(\mathrm{SSIM}(n)-\mu(\mathrm{SSIM}))^2}. \quad (40)$$

As w becomes larger, the impact of the temporal variance is increased. Since large values of $\sigma(\mathrm{PSNR})$ negatively impact video quality, $w>0$. Since quality index needs to be positive, $\mu(\mathrm{PSNR})-w\sigma(\mathrm{PSNR})>0$. Hence we require $0<w<\min(\mu(\mathrm{PSNR})/\sigma(\mathrm{PSNR}))$ in (35). Similarly, $0<w<\min(\mu(\mathrm{SSIM})/\sigma(\mathrm{SSIM}))$ in (38). In the dataset used in the simulations in Section 4, $\min(\mu(\mathrm{PSNR})/\sigma(\mathrm{PSNR}))$ and $\min(\mu(\mathrm{SSIM})/\sigma(\mathrm{SSIM}))$ are 5.8 and 29.4, respectively. Hence w was confined to the range (0, 5.8) and (0, 29.4) for PSNR-TV and SSIM-TV, respectively, in the simulations. It is possible that $\min(\mu(\mathrm{PSNR})/\sigma(\mathrm{PSNR}))$ and $\min(\mu(\mathrm{SSIM})/\sigma(\mathrm{SSIM}))$ may be smaller in other practical dataset. In our simulations, values of w ranging from 0.5 to 3 and from 2 to 10 were used for PSNR-TV and SSIM-TV, respectively.

4. Simulation results

We simulated H.264 video transmission over wireless channels. H.264 video encoding and decoding was performed using the H.264 reference software in [32]. Wireless channels were simulated using the software provided by the Video Coding Expert Group (VCEG) in [33]. The real-time transport protocol (RTP) packet mode was selected in the H.264 encoding, and robust header compression (RoHC) mode was assumed for 40 byte IP/UDP/RTP headers to be compressed into three bytes in the packetization. The link layer packet[1] size was set to 80 bytes, and the link layer packet header size was set to 4 bytes. The QCIF *Foreman*, *Mother and Daughter*, and *Bus* videos with 75 frames were used for simulations. The first frame was encoded as an I-frame and the rest frames were encoded as P-frames. Each row of macroblocks composed a slice and was packed into a packet: 11 macroblocks in a slice/packet and nine slices in a QCIF frame.

Packet losses increase the variability of video quality. To capture and quantify this phenomenon, we fixed the H.264 quantization parameter (QP), which results in variable bit rate compressed video. If a constant bit rate were imposed using rate control, the quality would vary from the source coding, complicating our attempt to isolate and analyze variations arising from packet losses. In the following we examine the performance of PSNR-TV with w = 1, and also the performance of SSIM-TV with w = 4.

It is well known that channel coding can efficiently reduce packet loss. In conventional rate-distortion (RD) schemes for video transmission, only the mean distortion or mean quality, using perceptually questionable measures (such as PSNR), is used for optimization. We argue that the channel coding can also greatly reduce the temporal variation of video quality. To investigate this possibility, we simulated channel coding in the link layer. The scope of the simulation is not to propose a new channel coding or joint source-channel coding scheme, but to investigate the *possibility* of reducing quality variations by *existing* error control schemes as a case study. Hence multiple runs with random starting positions of errors for statistical performance experiment were not attempted in our simulations.

We used Reed–Solomon coding in the link layer packets as an error control scheme. The symbol size for the Reed–Solomon code was set to eight bits (one byte). Denoting the link layer packet size and source code size as n and k symbols, respectively, then the error correction capability is $(n-k)/2$. The source codes are mainly composed of RTP payload, which is the H.264 bitstream. The source codes also include the protocol data unit (PDU) header and the link layer packet header in the packetization as in [33]. The simulation software in [33] was modified to include error correction by channel coding in the link layer packets.

We considered three cases in our simulations:

- Case 1: QP=24 without packet loss (no wireless channel).
- Case 2: QP=24 without channel coding (in wireless channel).
- Case 3: QP=30 with channel coding (in wireless channel).

In Case 1, the wireless channel is not included and there is no packet loss. In Cases 2 and 3, wireless channels are simulated. In Case 2, channel coding is not applied in the link layer packets. Case 2 would result in the loss of link layer packets if any symbol errors occur in a packet. If a link layer packet is lost, then the upper layer RTP packet, including the lost link layer packet, would be lost. In Case 3, the source and channel code size is set to 44 and 36 bytes, respectively, in a link layer packet with 80 bytes.

The total number of bytes used for source and channel codes in Case 3 is similar to the number of bytes used only for source codes in Case 2, for the *Foreman* video sequence. This set-up is also used in other video simulations. Table 1 shows the number of bytes allocated for source codes (including RTP payload, PDU header, and radio link packet header) and channel codes for the simulated distorted videos. Actually the sum of the number of bytes for the source and channel codes in Case 3 is smaller than the number of bytes for the source codes only in Case 2, for all videos.

Table 1
Number of bytes for source codes and channel codes.

		Case 2 (QP=24)	Case 3 (QP=30)
Foreman	Source code	144880	78056
	Channel code	0	63864
Mother and Daughter	Source code	60160	31812
	Channel code	0	26028
Bus	Source code	228720	121528
	Channel code	0	99432

[1] The terminology in [33] is frame. We use the terminology *packet* to avoid confusion with the *frame* in video sequence.

4.1. Distortion analysis

From simulation, source ($D_s(n)$), channel ($D_c(n)$), and end-to-end ($D(n)$) relationship values are obtained for n=0,1,…,74.

Fig. 1 shows the source, channel, and end-to-end distortion for the *Foreman* video sequence. The source distortion D_s in Cases 1 and 2 with QP=24 (represented by $*$) was smaller than the source distortion in Case 3 with QP=30 (represented by $\times$).

However, the channel distortion D_c in Case 2 (represented by +) without channel coding was quite large in many frames. In this case, packet losses occurred in the simulated wireless channel: seven RTP packets were lost among 677 RTP packets. The packet loss rate (PLR) was about 1%, which is rather small, but the impact on the distortion and video quality was quite large. More specifically, 2, 2, and 3 RTP packets were lost in the 32nd, 35th, and 58th frames. There was a significant error concealment distortion arising from packet loss in the 32nd and 35th frames. These error concealment distortions propagated through subsequent frames. The error was reduced over time, and was stopped near the 50th frame by the introduction of intra-mode encoding. There was also an error concealment distortion in the 58th frame, and this error concealment distortion propagated through the end of the video.

Since there was no packet loss before the 32nd frame and hence no error propagation distortion in the 32nd frame, $D_{ec}(32)=D_c(32)=65.2$. Since $D_c(33)=D_{ep}(33)=60.0$, the α_{33} can be calculated from (10) as $\alpha_{33}=D_c(33)/D_c(32)=0.92$. Since there was packet loss in the 35th frame, error concealment distortion and error propagation distortion were mixed in the 35th frame. The error propagation distortion in the 35th frame can be calculated from (11) as $D_{ep}(35)\approx\alpha_{33}D_c(34)=0.92\cdot 56.9=52.3$. The error concealment distortion in 35th frame can be calculated from (12) as $D_{ec}(35)=D_c(35)-D_{ep}(35)=171.1-52.3=118.8$. Since $D_c(57)$ was almost 0, $D_{ec}(58)\approx D_c(58)=65.5$. The α_{36} can be calculated as $\alpha_{36}=D_c(36)/D_c(35)=104.8/171.1=0.61$. Similarly, $\alpha_{59}=D_c(59)/D_c(58)=61.1/65.5=0.93$. The sum of channel distortion $\mathcal{S}(D_c)$ can be calculated from (13), $\mathcal{S}(D_c)=1752.9$. The sum of error concealment distortion $\mathcal{S}(D_{ec})$ can be calculated from (14) as $\mathcal{S}(D_{ec})=D_{ec}(32)+D_{ec}(35)+D_{ec}(58)=249.8$. The sum of error propagation distortion $\mathcal{S}(D_{ep})$ can be obtained from (15) as $\mathcal{S}(D_{ep})=\mathcal{S}(D_c)-\mathcal{S}(D_{ec})=1503.1$. We can see that the sum of error propagation distortion was much larger than the sum of error concealment distortion.

Since Case 3 uses encoding with a larger QP, the number of source code bits was much smaller than for Case 2. Owing to the error correction capability in the link layer, no RTP packet was lost in the simulated wireless channel. Since there was no packet loss, the channel distortion in Case 3 (represented by $\diamond$) was 0 for all frames in this case. In Fig. 1, the channel distortion was much larger than the source distortion when packet losses occurred in the wireless channel in Case 2. Indeed, the variance of D_c was quite large in Case 2. Hence the variation of the PSNR in Case 2 was large (Fig. 4).

When there was no packet loss, the end-to-end distortions in Case 1 (represented by $*$) and in Case 3 (represented by $\bigcirc$) were the same as the source distortions in both cases. When there were packet losses in Case 2, the end-to-end distortion (represented by $\square$) was mostly determined by channel distortion (represented by +), which was much larger than the source distortion (represented by $*$).

Fig. 2 shows the source, the channel, and the end-to-end distortion for the *Mother and Daughter* video. The source distortion D_s in Cases 1 and 2 with QP=24 (represented by $*$) was much smaller than the source distortion in Case 3 with QP=30 (represented by $\circ$). In Case 2, eight RTP packets were lost among 677 RTP packets in the simulated wireless channel: 2, 1, and 5 packets were lost in the 16th, 73rd, and 74th frames. The channel distortion D_c in Case 2

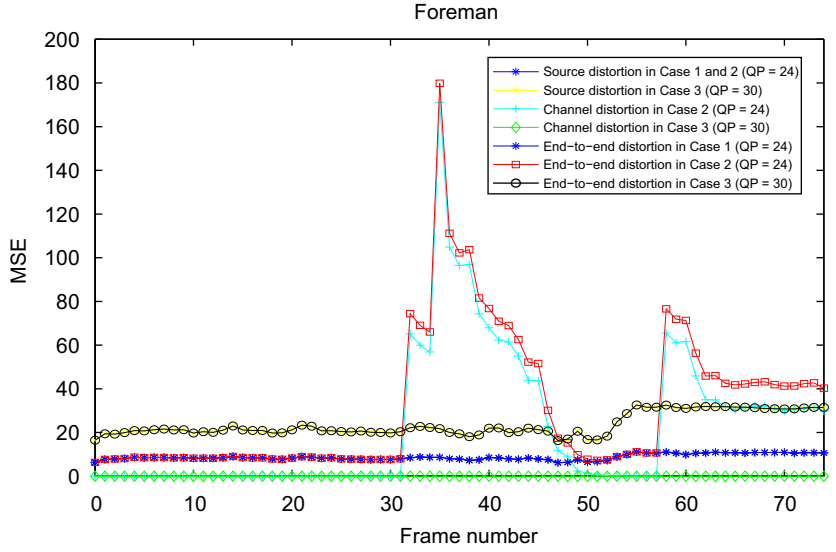

Fig. 1. Source, channel, and end-to-end distortion (MSE) for *Foreman* video.

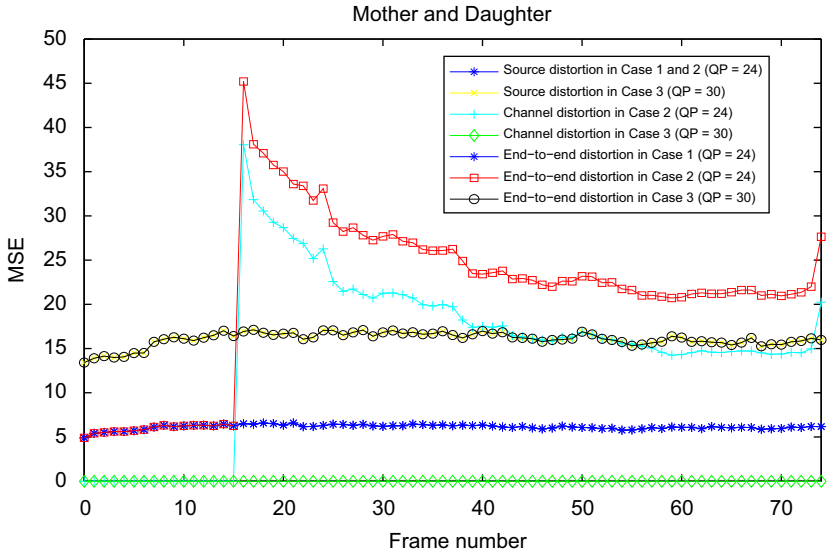

Fig. 2. Source, channel, and end-to-end distortion (MSE) for *Mother and Daughter* video.

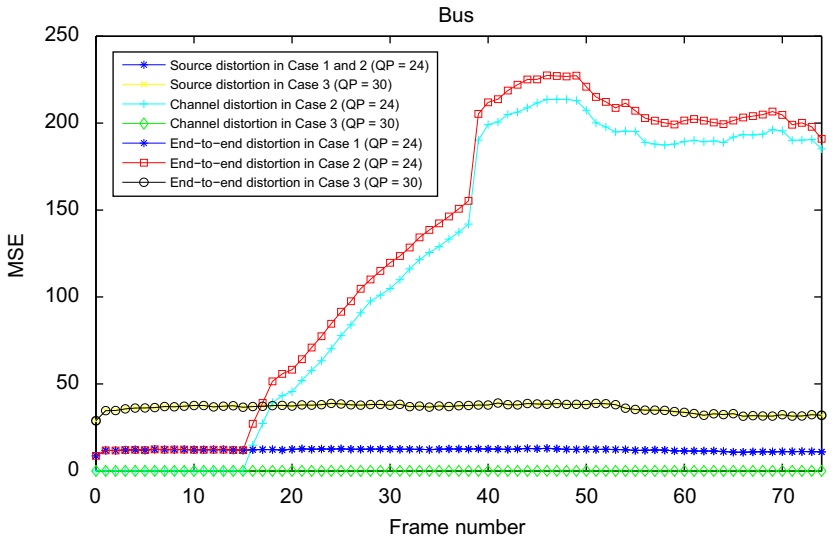

Fig. 3. Source, channel, and end-to-end distortion (MSE) for *Bus* video.

(represented by +) was rather large in some frames. Since there was no packet loss before the 16th frame, $D_{ec}(16) = D_c(16)$=38.0. Since $D_c(17)$=$D_{ep}(17)$=31.8, the α_{17} can be calculated as $\alpha_{17} = D_c(17)/D_c(16) = 0.84$. The $D_{ep}(73)$ can be approximated as $D_{ep}(73) \approx \alpha_{17} D_c(72) = 12.2$. Hence $D_{ec}(73) = D_c(73) - D_{ep}(73) \approx 2.8$. Similarly, $D_{ec}(74) = D_c(74) - D_{ep}(74) \approx 7.6$. The sum of channel distortion can be obtained as (13), $\mathcal{S}(D_c) = 1117.3$. The sum of error concealment distortion can be obtained as $\mathcal{S}(D_{ec}) = D_{ec}(16) + D_{ec}(73) + D_{ec}(74) = 48.4$. Hence $\mathcal{S}(D_{ep}) = \mathcal{S}(D_c) - \mathcal{S}(D_{ec}) = 1068.9$, and the $\mathcal{S}(D_{ep})$ is much larger than $\mathcal{S}(D_{ec})$. The channel distortion D_c for *Mother and Daughter* was not as large as for *Foreman* (Fig. 1), but D_c was still much larger than D_s once the packet losses occurred in the 16th frame.

Fig. 3 shows the source, the channel, and the end-to-end distortion for the *Bus* video sequence. The source distortion D_s in Cases 1 and 2 with QP=24 (represented by ∗) was much smaller than the source distortion in Case 3 with QP=30 (represented by ○).

In Case 2, without channel coding, one packet was lost in each of the 16th, 17th, 18th, 19th, and 39th frames. Since there was no packet loss before the 16th frame, $D_{ec}(16)$=$D_c(16)$=14.8. Since there was packet loss in the 17th frame, $D_{ec}(17)$ was approximately calculated as $D_{ec}(17) = D_c(17) - D_{ep}(17) \approx D_c(17) - D_c(16) = 12.5$. Similarly, $D_{ec}(18) \approx 12.0$ and $D_{ec}(19) \approx 3.9$. The α_{20} can be calculated as $\alpha_{20} = D_c(20)/D_c(19) = 1.06$. In this case, the error propagation factor $\alpha_{20} > 1$, which is a special characteristic of the *Bus* video. Since there was another packet

loss in the 39th frame, $D_{ep}(39) \approx \alpha_{20} D_c(38) = 150.4$ and $D_{ec}(39) = D_c(39) - D_{ep}(39) \approx 39.8$. The sum of channel distortion was $\mathcal{S}(D_c) = 9070.0$. The sum of error concealment distortion can be calculated as $\mathcal{S}(D_{ec}) = D_{ec}(16) + D_{ec}(17) + D_{ec}(18) + D_{ec}(19) + D_{ec}(39) = 83.0$. Then the sum of error propagation distortion was $\mathcal{S}(D_{ep}) = \mathcal{S}(D_c) - \mathcal{S}(D_{ec}) = 8987.0$. The maximum channel distortion $D_c(n)$ occurred around the 47th frame, which is about five times larger than the source distortion D_s with QP=24. The variation of the channel distortion in Case 2 was also large for the *Bus* video, which would result in a large variation in quality (Fig. 10).

4.2. Quality index analysis

Fig. 4 plots the PSNR of compressed versions of the *Foreman* video with and without packet loss. Case 1, plotted using the symbol $*$, shows the PSNR variation of the decoded video encoded with QP=24 and without packet loss. In Case 1, $\mu(\text{PSNR}) = 38.8$ dB and $\sigma(\text{PSNR}) = 0.66$ dB. The mean quality was rather high and the quality variation was small.

Case 2, plotted using the symbol □, shows the PSNR variation of the decoded video sequence with QP=24 and without channel coding. There was significant quality degradation arising from packet loss in the 32nd frame and in the 35th frame. These degradations propagated through subsequent frames. The error was reduced over time, and was stopped near the 50th frame by the introduction of intra-mode encoding. There was also quality degradation in the 58th frame, and this degradation propagated through the end of the video. These quality degradations not only reduced the mean video quality, but also greatly increased the variance of the video quality.

Case 3, plotted using the symbol ○, shows the PSNR variation of the decoded video sequence with QP=30 and with channel coding. The value of $\mu(\text{PSNR})$ is slightly larger (by 0.7 dB) in Case 2 than in Case 3. However, the value of PSNR-TV was significantly larger (by 2.7 dB) in Case 3 than in Case 2. According to our proposed quality index PSNR-TV, the video quality in Case 3 is much better than the quality in Case 2, even though it is worse according to the conventional video quality index $\mu(\text{PSNR})$.

Fig. 5 shows the SSIM plot of the *Foreman* video. In Case 1 (represented by $*$), $\mu(\text{SSIM}) = 0.9684$ and $\sigma(\text{SSIM}) = 0.0028$. The mean quality was rather high and the quality variation was small.

In Case 2 (represented by □), there was significant quality degradation arising from packet loss in the 32nd and 35th frames, and these degradations propagated through subsequent frames. These quality degradations by packet loss not only reduced the mean SSIM, but also increased the variance of the SSIM values.

Case 3 (represented by ○) shows the SSIM variation of the decoded video sequence with QP=30 and with channel coding. The value of $\mu(\text{SSIM})$ is slightly larger (by 0.0245) in Case 2 than in Case 3. Conversely, the value of SSIM-TV was larger (by 0.0085) in Case 3 than in Case 2. The video quality in Case 3 is better than the quality in Case 2 according to our proposed quality index SSIM-TV, even though it is worse according to the conventional video quality index $\mu(\text{SSIM})$.

Fig. 6 shows some of the decoded frames in Cases 2 and 3 of Fig. 4. Fig. 6(a) shows the first frame encoded with QP=24 without channel coding (Case 2). The quality was very good with PSNR=40.1 dB and SSIM=0.9731, since the QP was low and there was no packet loss in the first frame. However, the quality was seriously degraded in later frames by packet loss. Fig. 6(b) shows the 35th frame in Case 2. This frame was seriously degraded by the packet loss in the 35th frame and the error propagation from the packet loss in the 32nd frame. This frame would be very annoying for human viewing. The PSNR change in the 35th frame from packet losses was about 14 dB, which is a huge

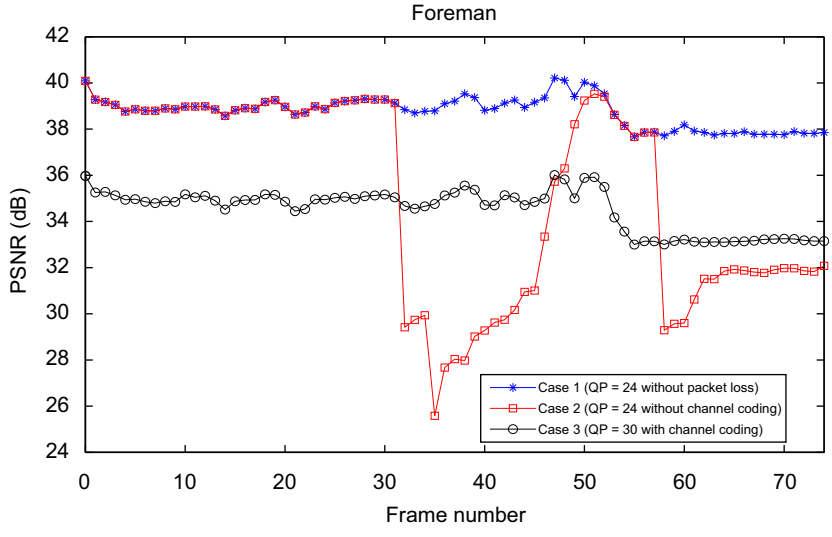

Fig. 4. PSNR of compressed *Foreman* video. Case 1 (QP=24 without packet loss): $\mu(\text{PSNR}) = 38.8$ dB, $\sigma(\text{PSNR}) = 0.66$ dB, PSNR-TV=38.1 dB. Case 2 (QP=24 without channel coding): $\mu(\text{PSNR}) = 35.2$ dB, $\sigma(\text{PSNR}) = 4.3$ dB, PSNR-TV=30.9 dB. Case 3 (QP=30 with channel coding): $\mu(\text{PSNR}) = 34.5$ dB, $\sigma(\text{PSNR}) = 0.91$ dB, PSNR-TV=33.6 dB.

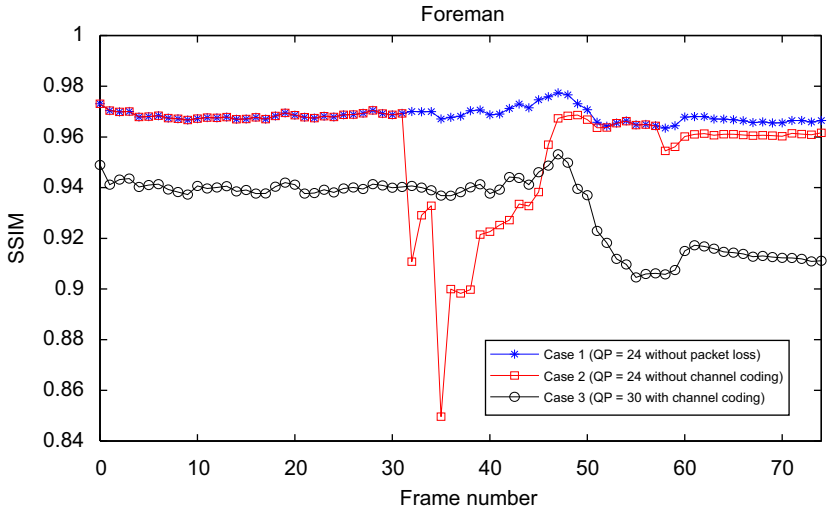

Fig. 5. SSIM of compressed *Foreman* video. Case 1 (QP=24 without packet loss): $\mu(\text{SSIM}) = 0.9684$, $\sigma(\text{SSIM}) = 0.0028$, SSIM-TV=0.9574. Case 2 (QP=24 without channel coding): $\mu(\text{SSIM}) = 0.9562$, $\sigma(\text{SSIM}) = 0.0221$ dB, SSIM-TV=0.8678. Case 3 (QP=30 with channel coding): $\mu(\text{SSIM}) = 0.9317$, $\sigma(\text{SSIM}) = 0.0138$, SSIM-TV=0.8763.

Fig. 6. (a) First frame in Case 2 (QP=24 without channel coding), PSNR=40.1 dB, SSIM=0.9731. (b) 35th frame in Case 2, PSNR=25.6 dB, SSIM=0.8497. (c) First frame in Case 3 (QP=30 with channel coding), PSNR=36.0 dB, SSIM=0.9489. (d) 35th frame in Case 3, PSNR=34.8 dB, SSIM=0.9369.

degradation. Fig. 6(c) shows the first frame encoded with QP=30 using channel coding (Case 3). The image quality is worse than the first frame in Case 2 (Fig. 6(a)) because the QP is larger in the source coding. This quality degradation is quite acceptable as compared to the quality degradation by the packet loss (Fig. 6(b)). Fig. 6(d) shows the 35th frame. Since there was no packet loss, there is no serious degradation, and is of much better quality than the 35th frame in Case 2. Although the PSNR was lower than the 35th frame with QP=24 and without packet loss (Case 1), it was much

higher than the 35th frame with QP=24 and with packet loss (Case 2). The variances of the PSNR and SSIM values in Case 3 were small, and the quality level remained nearly constant through the video, which is desirable.

Fig. 7 shows the PSNR values of *Mother and Daughter* video with and without packet loss. Case 1 (represented by $*$) shows the PSNR variation of the decoded video sequence, encoded with QP=24 without packet loss. Here $\mu(\text{PSNR}) = 40.3$ dB and $\sigma(\text{PSNR}) = 0.21$ dB.

Case 2 (represented by $\square$) shows the PSNR variation of the decoded video with QP=24 and without channel coding. There was significant quality degradation from packet loss in the 16th frame. This quality degradation was propagated through the end of video. Since the error propagation from the packet loss in the 16th frame affected the rest of the videos, the quality variation was large. The quality degradation from packet loss in the 73rd and 74th frames was, in this case, not large.

Case 3 (represented by $\bigcirc$) shows the PSNR variation of decoded video with QP=30 and with channel coding. Since in Case 3 used encoding with a larger QP, the number of allocated source code bits was much smaller than in Case 2. The error correction capability of channel codes ensured that all the erroneous radio link packets were corrected in the simulated wireless channel. Indeed, $\mu(\text{PSNR})$ was just 0.6 dB larger in Case 3 than in Case 2. Conversely, the value of PSNR-TV in Case 3 greatly exceeds the value in Case 2 (by 3.0 dB).

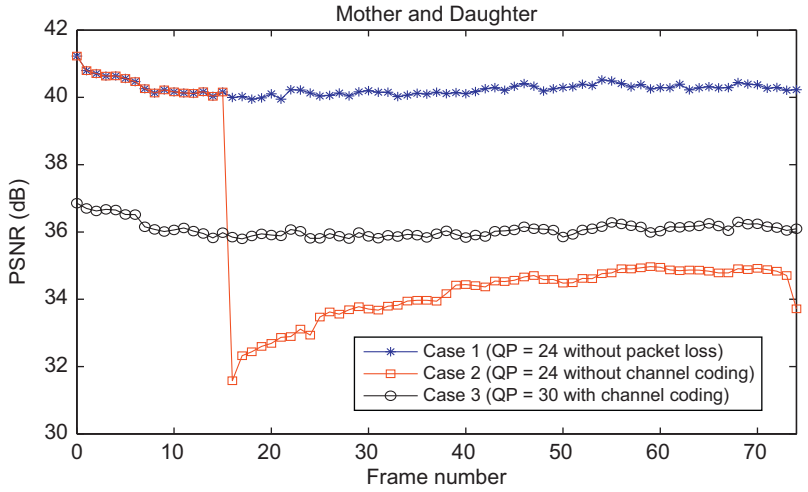

Fig. 7. PSNR values of *Mother and Daughter* video. Case 1 (QP=24 without packet loss): $\mu(\text{PSNR}) = 40.3$ dB, $\sigma(\text{PSNR}) = 0.21$ dB, PSNR-TV=40.1 dB. Case 2 (QP=24 without channel coding): $\mu(\text{PSNR}) = 35.5$ dB, $\sigma(\text{PSNR}) = 2.7$ dB, PSNR-TV=32.8 dB. Case 3 (QP=30 with channel coding): $\mu(\text{PSNR}) = 36.1$ dB, $\sigma(\text{PSNR}) = 0.23$ dB, PSNR-TV=35.8 dB.

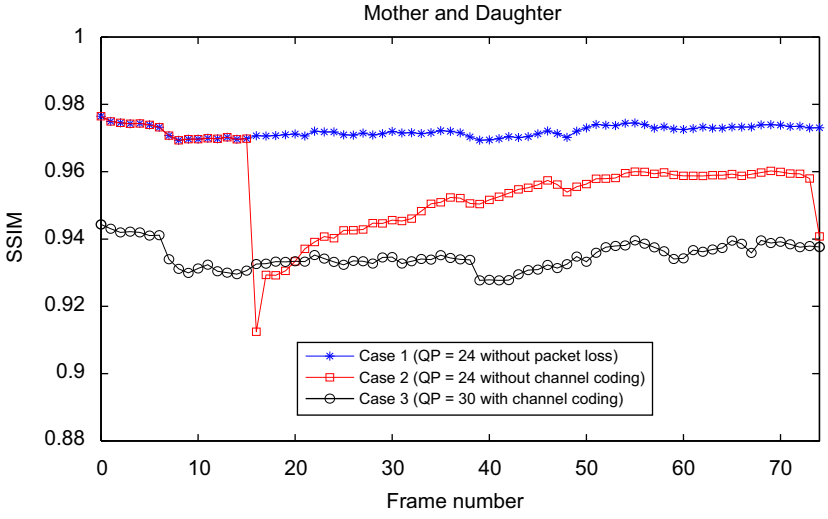

Fig. 8. SSIM of compressed *Mother and Daughter* video. Case 1 (QP=24 without packet loss): $\mu(\text{SSIM}) = 0.9720$, $\sigma(\text{SSIM}) = 0.0016$, SSIM-TV=0.9655. Case 2 (QP=24 without channel coding): $\mu(\text{SSIM}) = 0.9554$, $\sigma(\text{SSIM}) = 0.0125$ dB, SSIM-TV=0.9054. Case 3 (QP=30 with channel coding): $\mu(\text{SSIM}) = 0.9349$, $\sigma(\text{SSIM}) = 0.0038$, SSIM-TV=0.9195.

Fig. 8 shows the SSIM values of *Mother and Daughter* video with and without packet loss. In Case 1 (represented by $*$), encoded with QP=24 without packet loss, $\mu(\text{SSIM}) = 0.9720$ and $\sigma(\text{SSIM}) = 0.0016$. In Case 2 (represented by $\square$), encoded with QP=24 and without channel coding, there was significant quality degradation from packet loss in the 16th frame, and this quality degradation was propagated through the end of video. Case 3 (represented by $\bigcirc$) shows the SSIM variation of decoded video with QP=30 and with channel coding. Here, $\mu(\text{SSIM})$ was slightly larger in Case 2 than in Case 3 (by 0.0205). Conversely, the value of SSIM-TV in Case 3 was larger than in Case 2 (by 0.0141).

Fig. 9 shows some of the decoded frames in Cases 2 and 3 of Fig. 7. Fig. 9(a) shows the first frame encoded in Case 2 with QP=24 and without channel coding. The quality was very good (PSNR=41.2 dB, SSIM=0.9765), since the QP was low and there was no packet loss in the first frame. Fig. 9(b) shows the 16th frame. This frame was degraded by the packet loss in the 16th frame. Since part of the child's face, which is of particular perceptual significance, was degraded, this frame is annoying. Fig. 9(c) shows the first frame encoded with QP=30 and with channel coding in Case 3. The quality was comparably worse (PSNR=36.9 dB, SSIM=0.9443) than the first frame in Case 2 (Fig. 9(a)) because the QP was large. Fig. 9(d) shows the 16th frame. Since there was no packet loss, there was no serious quality degradation. It shows much higher quality (PSNR=35.8 dB, SSIM=0.9326) than the 16th frame in Case 2 (PSNR=31.6 dB, SSIM=0.9125). The quality variation in Case 3 was much smaller than the quality variation in Case 2.

Fig. 10 shows the PSNR of *Bus* video. In Case 1, encoded with QP=24 and without packet loss, the mean PSNR was high and the PSNR variation was small. In Case 2, without channel coding, five packets were lost in the simulated wireless channel. One packet was lost in each of the 16th, 17th, 18th, 19th, and 39th frames. The PSNR reduction from packet loss was propagated through the end of the video. In this case, the PSNR decreased from the 20th through the 38th frame from error propagation. The mean PSNR decreased greatly and the variation of the PSNR values was much larger from packet loss, as compared to Case 1 without packet loss. In Case 3, the variance of PSNR values was small, yielding stable video quality. Comparing the PSNR-TV values, Case 3 resulted in a 8.6 dB higher value than Case 2.

Fig. 11 shows the SSIM values of *Bus* video. In Case 1, the mean SSIM was high ($\mu(\text{SSIM})$) and the SSIM variation was small ($\sigma(\text{SSIM})$). In Case 2, five packets were lost (one packet in each of the 16th, 17th, 18th, 19th, and 39th frames), and the SSIM reduction from packet loss was propagated through the end of the video. In Case 3, the variance of SSIM values was relatively small. The SSIM-TV value was much larger in Case 3 than in Case 2 (by 0.101), while the mean SSIM values were very close in Cases 2 and 3.

Table 2 shows the statistical value comparison of distortion and quality indices for video sequences in Cases

Fig. 9. (a) First frame in Case 2 (QP=24 without channel coding), PSNR=41.2 dB, SSIM=0.9765. (b) 16th frame in Case 2, PSNR=31.6 dB, SSIM=0.9125. (c) First frame in Case 3 (QP=30 with channel coding), PSNR=36.9 dB, SSIM=0.9443. (d) 16th frame in Case 3, PSNR=35.8 dB, SSIM=0.9326.

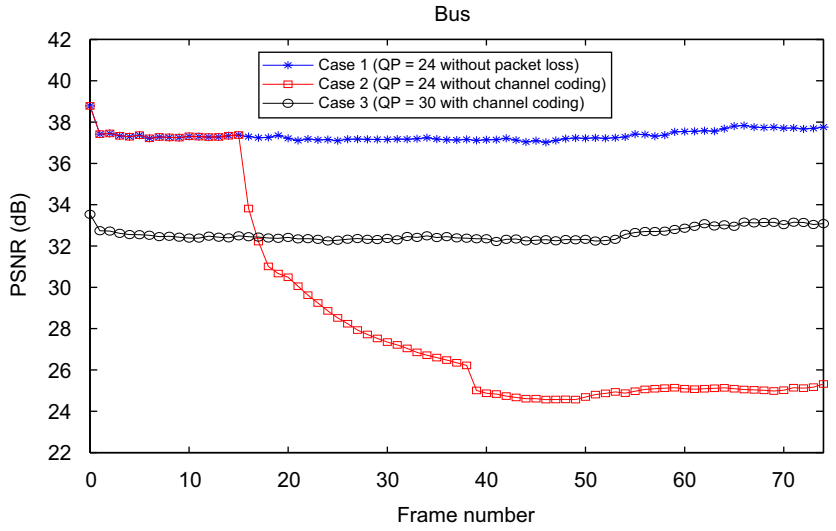

Fig. 10. PSNR values of compressed *Bus* video. Case 1 (QP=24 without packet loss): $\mu(\text{PSNR}) = 37.3$ dB, $\sigma(\text{PSNR}) = 0.27$ dB, PSNR-TV=37.0 dB. Case 2 (QP=24 without channel coding): $\mu(\text{PSNR}) = 28.7$ dB, $\sigma(\text{PSNR}) = 5.0$ dB, PSNR-TV=23.7 dB. Case 3 (QP=30 with channel coding): $\mu(\text{PSNR}) = 32.6$ dB, $\sigma(\text{PSNR}) = 0.31$ dB, PSNR-TV=32.3 dB.

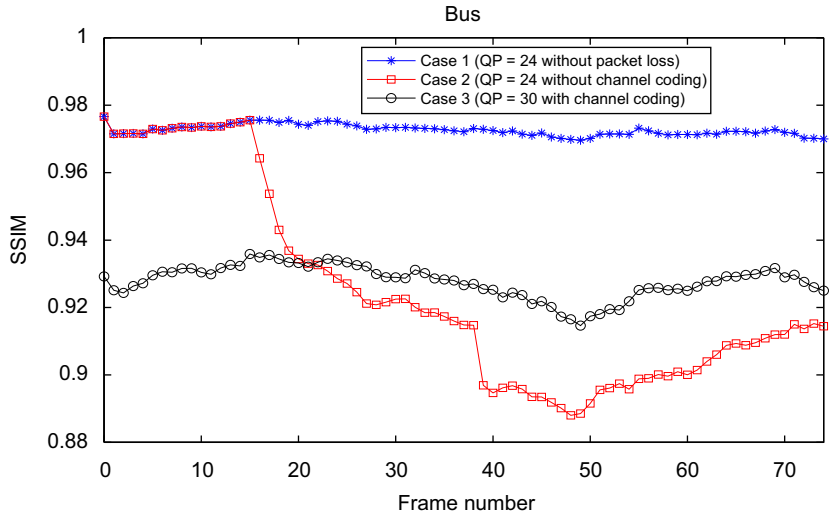

Fig. 11. SSIM of compressed *Bus* video. Case 1 (QP=24 without packet loss): $\mu(\text{SSIM}) = 0.9726$, $\sigma(\text{SSIM}) = 0.0016$, SSIM-TV = 0.9663. Case 2 (QP=24 without channel coding): $\mu(\text{SSIM}) = 0.9247$, $\sigma(\text{SSIM}) = 0.0294$ dB, SSIM-TV=0.8072. Case 3 (QP=30 with channel coding): $\mu(\text{SSIM}) = 0.9278$, $\sigma(\text{SSIM}) = 0.0048$, SSIM-TV=0.9084.

2 and 3. This comparison shows that the statistical values of quality indices are closely related to statistical values of distortion. Especially, the variation of quality indices is closely related to the variation of distortions.

When there is no packet loss with channel coding in Case 3, the channel distortion values are zero and distortion values are determined by source distortion values. However, when there are packet losses without channel coding in Case 2, the channel distortion ratio ρ_c values in (23) are around 0.7–0.9 and the distortion values are dominantly determined by channel distortion even though the PLR values are relatively small (around 1%) in wireless channel. The error propagation distortion ratio ρ_{ep} values in (16) are larger than 0.85 and error propagation distortions are much larger than error concealment distortions.

When there is no packet loss in Case 3, the $\sigma(D)$ values are small, since $\sigma(D_c)$ values are zero and $\sigma(D_c)$ values are small. In source coding, a constant quantization step value results in a smaller variation of distortion and hence small variations of the quality indices $\sigma(\text{PSNR})$ and $\sigma(\text{SSIM})$ values as shown in Case 3. When there are packet losses in Case 2, the $\sigma(D)$ values are large, since $\sigma(D_c)$ values are large. The channel errors result in much larger variations of distortion than the source coding errors. The $\sigma(\text{PSNR})$ and $\sigma(\text{SSIM})$ values are much larger in Case 2 than in Case 3, since the $\sigma(D)$ values are much larger in Case 2. Hence the temporal

Table 2
Statistical value comparison of distortion and quality indices for video sequences in Case 2 (QP=24 without channel coding) and in Case 3 (QP=30 with channel coding).

	Case 2			Case 3		
	Foreman	*Mother and Daughter*	*Bus*	*Foreman*	*Mother and Daughter*	*Bus*
$\mu(D_s)$	8.7	6.1	12.0	23.5	16.1	36.1
$\sigma(D_s)$	1.3	0.3	0.7	5.2	0.8	2.5
$\mu(D_c)$	23.4	14.9	120.9	0	0	0
$\sigma(D_c)$	33.2	9.1	82.5	0	0	0
$\mu(D)$	32.1	21.3	133.4	23.5	16.1	36.1
$\sigma(D)$	33.3	9.3	82.7	5.2	0.8	2.5
PLR(%)	1.03	1.18	0.74	0	0	0
ρ_c	0.729	0.700	0.906	0	0	0
ρ_{ep}	0.857	0.956	0.991	–	–	–
μ(PSNR)	35.2	35.5	28.7	34.5	36.1	32.6
σ(PSNR)	4.3	2.7	5.0	0.91	0.23	0.31
PSNR-TV	30.9	32.8	23.7	33.6	35.8	32.3
μ(SSIM)	0.9562	0.9554	0.9247	0.9317	0.9349	0.9278
σ(SSIM)	0.0221	0.0125	0.0294	0.0138	0.0038	0.0048
SSIM-TV	0.8678	0.9054	0.8072	0.8763	0.9159	0.9084

quality variations become very large when there are noticeable packet losses (around 1% in these video sequences). The temporal quality variations negatively impact subjective video quality, as will be shown in the next subsection.

4.3. Subjective video quality experiment

We also performed a subjective video quality assessment experiment. The experiment was based on the double-stimulus impairment scale (DSIS) Variant II in ITU-R BT.500-11 [19]. The double-stimulus method is cyclic, wherein the reference (original) video is first presented, then the same impaired test video is presented. In DSIS Variant II, the reference video and the test video are presented twice as shown in the presentation timing in Fig. 12.

The observer was asked to vote the subjective impairment score (SIS) using the five-grade impairment scale in [19]:

5 imperceptible
4 perceptible, but not annoying
3 slightly annoying
2 annoying
1 very annoying

The subjective impairment score ranges from 1 to 5, and we used a continuous scale for improved accuracy. The *Mobile*, *Suzie*, *Silent*, *Coastguard*, *Hall Monitor*, and *Football* QCIF video sequences were also included with the three video sequences in subjective experiment. Hence 27 test videos were used in the experiment: nine video sequences with three cases. The observers were recruited among undergraduate junior, senior, and graduate students at Konkuk University, Seoul (17 observers participated), and their scores were averaged for each test video.

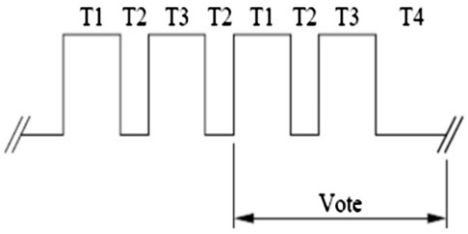

Fig. 12. DSIS Variant II presentation timing (T1=10 s reference video sequence, T2=3 s mid-gray video, T3=10 s test video sequence, T4=5–11 s mid-gray video) [19].

Fig. 13 shows the scatter plots of subjective impairment score (SIS) vs. quality indices. In Fig. 13, the line indicates the fitted curve by regression. We calculated three statistical values for performance comparison. The first value is the linear correlation coefficient (CC) between SIS and quality indices. The second and third values are root-mean-squared error (RMSE) and mean-absolute error (MAE) between the SIS and the fitted curve after regression. The PSNR-TV with w=3 resulted in much better performance compared to the conventional video quality index μ(PSNR): the CC was much larger, and the RMSE and MAE were smaller. Similarly, SSIM-TV with w=8 resulted in much better performance compared to μ(SSIM). The temporal quality variance factor in the new video quality indices (PSNR-TV, SSIM-TV) also yielded a notable performance improvement. Humans are sensitive to quality variations over time, apparently preferring a constant video quality to a variable video quality, even though the average quality indices (μ(PSNR), μ(SSIM)) might be lower.

Table 3 shows the statistical performance comparison for the subjective video quality experiment. We also tried other values of w. The PSNR-TV and SSIM-TV for all values of w that were tried resulted in better performance than μ(PSNR) and μ(SSIM), respectively. In our experiment, as

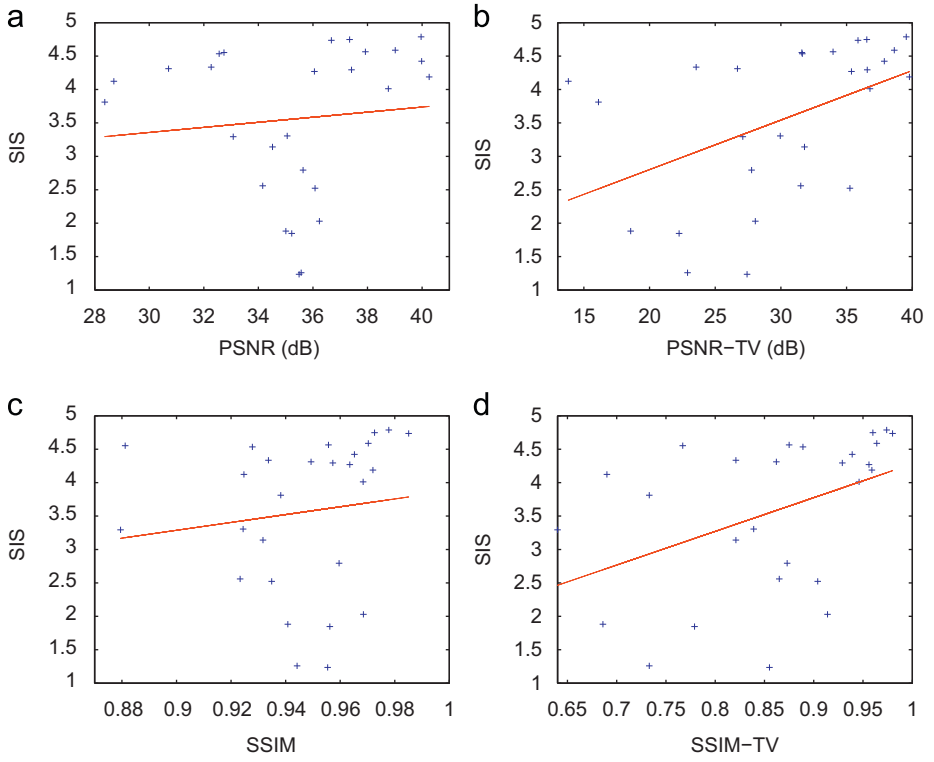

Fig. 13. Scatter plots of subjective impairment score (SIS) vs. video quality indices. (a) SIS vs. μ(PSNR): CC=0.1044, RMSE=1.1161, MAE=0.9970. (b) SIS vs. PSNR-TV (w=3): CC=0.4658, RMSE=0.9931, MAE=0.8276. (c) SIS vs. μ(SSIM): CC=0.1339, RMSE=1.1121, MAE=0.9534. (d) SIS vs. SSIM-TV (w=8): CC=0.4324, RMSE=1.0119, MAE=0.8513.

Table 3
Statistical performance comparison of subjective video quality experiment.

	CC	RMSE	MAE
μ(PSNR)	0.1044	1.1161	0.997
PSNR-TV (w=0.5)	0.2275	1.0928	0.9459
PSNR-TV (w=1)	0.3131	1.0658	0.9104
PSNR-TV (w=2)	0.4133	1.0219	0.8580
PSNR-TV (w=3)	0.4658	0.9931	0.8276
μ(SSIM)	0.1339	1.1121	0.9534
SSIM-TV (w=2)	0.3217	1.0626	0.9008
SSIM-TV (w=4)	0.3855	1.0355	0.8740
SSIM-TV (w=6)	0.4153	1.0209	0.8600
SSIM-TV (w=8)	0.4324	1.0119	0.8513
SSIM-TV (w=10)	0.4418	1.0067	0.8467

the w value was increased, the performance improved. Methods for optimally choosing w remain an important, if difficult, topic. The subjective video quality experiment suggests that PSNR-TV and SSIM-TV predict subjective video quality much better than the conventional μ(PSNR) and μ(SSIM), respectively.

5. Conclusion and future work

We examined the effect that temporal quality variation has on global video quality, using packet losses as a case study. We also proposed a simple method by which a frame-based quality assessment algorithm, such as PSNR or SSIM, can be adopted to account for temporal quality variation. In simulations on H.264 video transmission in typical wireless channel deployments, channel distortions are generally more severe than source distortions, and impact quality more severely when packet losses occur in error-prone networks. It was also shown that channel coding can greatly reduce the distortion and quality variation as well as the mean distortion and quality in distortion analysis and quality index analysis.

We showed that modifying video quality assessment to include temporal variance is more effective than conventional video quality assessment. We validated the approach by simulations and human study. The proposed PSNR-TV and SSIM-TV were found to predict subjective video quality much better than the conventional μ(PSNR) and μ(SSIM), respectively. The inclusion of the temporal quality variance factor in PSNR-TV and SSIM-TV appears to contribute significantly to the performance improvement of video quality assessment in a packet loss environment.

Significant research has been done on joint source-channel coding for video transmission using R-D frameworks. Such approaches implicitly assume that video quality is the sample mean of the frame qualities in a video as in (32). We argue that the video quality index can benefit by inclusion of the temporal variation as in (34) for R-D frameworks in video transmission applications. We believe that future work might profitably

be directed towards including the temporal variation of video quality in joint source-channel coding algorithms for video transmission over packet-lossy networks. The temporal variation of distortion or video quality can also be included in R-D frameworks for rate control that can achieve constant bit-rate source coding having variable quality levels.

Acknowledgements

The work of the first author (C. Yim) was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2008-013-D00083) and by the Konkuk University. This work was also supported in part by the MKE under the ITRC support program supervised by the NIPA (NIPA-2010-C1090-1031-0003). The work of the second author (A. C. Bovik) was supported in part by Intel Corp. and Cisco Inc. under the VAWN program.

References

[1] Z. He, J. Cai, 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. 12 (6) (2002) 511–523.
[2] 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 Trans. Multimedia 9 (3) (2007) 445–454.
[3] M.F. Sabir, H.R. Sheikh, R.W. Heath Jr., A.C. Bovik, Joint source-channel distortion model for JPEG compressed images, IEEE Trans. Image Process. 15 (6) (2006) 1349–1364.
[4] M.F. Sabir, R.W. Heath Jr., A.C. Bovik, Joint source-channel distortion modeling for MPEG-4 video, IEEE Trans. Image Process. 18 (11) (2009) 90–105.
[5] K. Stuhmüller, N. Fäber, M. Link, B. Girod, Analysis of video transmission over lossy channels, IEEE J. Sel. Areas Commun. 1 (6) (2000) 1012–1032.
[6] R. Zhang, S.L. Regunathan, K. Roth, Video coding with optimal inter/intra-mode switching for packet loss resilience, IEEE J. Sel. Areas Commun. 18 (6) (2000) 966–976.
[7] D. Wu, Y.T. Hou, B. Li, W. Zhu, Y.-Q. Zhang, H.J. Chao, An end-to-end approach for optimal mode selection in Internet video communication: theory and application, IEEE J. Sel. Areas Commun. 18 (6) (2000) 977–995.
[8] A.E. Mohr, E.A. Riskin, R.E. Ladner, Unequal loss protection: graceful degradation of image quality over packet erasure channels through forward error correction, IEEE J. Sel. Areas Commun. 18 (6) (2000) 819–828.
[9] M. Gallent, F. Kossentini, Rate-distortion optimized layered coding with unequal error protection for robust Internet video, IEEE Trans. Circuits Syst. Video Technol. 11 (3) (2001) 357–372.
[10] X. Yang, C. Zhu, Z.G. Li, X. Lin, N. Ling, An unequal packet loss resilience scheme for video over the Internet, IEEE Trans. Multimedia 7 (4) (2005) 753–765.
[11] H. Ha, C. Yim, Y.Y. Kim, Packet loss resilience using unequal forward error correction assignment for video transmission over communication networks, Comput. Commun. 30 (12) (2007) 3676–3689.
[12] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[13] Z. Wang, L. Lu, A.C. Bovik, Video quality assessment based on structural distortion measurement, Signal Process. Image Commun. 19 (2) (2004) 121–132.
[14] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[15] C. Li, A.C. Bovik, Content-partitioned structural similarity index for image quality assessment, Signal Process. Image Commun. 25 (7) (2010) 517–526.
[16] A.C. Bovik (Ed.), The Essential Guide to Image Processing, second ed., Academic Press, 2009.
[17] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[18] ITU-T, Contribution COM 9-80-E, Final report from video quality experts group on the validation of objective models of video quality assessment, [Online] 〈http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI〉, 2000.
[19] ITU-R, Recommendation BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002.
[20] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[21] S. Winkler, Perceptual distortion metric for digital color video, Proc. SPIE 3644 (1) (1999) 175–184.
[22] A.B. Watson, J. Hu, J.F. McGowan III, Digital video quality metric based on human vision, J. Electron. Imaging 10 (1) (2001) 20–29.
[23] M. Masry, S.S. Hemami, Y. Sermadevi, A scalable wavelet-based video distortion metric and applications, IEEE Trans. Circuits Syst. Video Technol. 16 (2) (2006) 260–273.
[24] K. Seshadrinathan, A.C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Trans. Image Process. 19 (2) (2010) 335–350.
[25] A.C. Bovik (Ed.), The Essential Guide to Video Processing, second ed., Academic Press, 2009.
[26] S. Lee, M.S. Pattichis, A.C. Bovik, Foveated video quality assessment, IEEE Trans. Multimedia 4 (1) (2002) 129–132.
[27] A. Leontaris, P.C. Cosman, A.R. Reibman, Quality evaluation of motion-compensated edge artifacts in compressed video, IEEE Trans. Image Process. 16 (4) (2007) 943–956.
[28] I.P. Gunawan, M. Ghanbari, Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration, IEEE Trans. Circuits Syst. Video Technol. 18 (1) (2008) 511–523.
[29] T. Wiegand, G. Sullivan, J. Bjøntegaard, G.A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 560–576.
[30] T. Stockhammer, M.M. Hannuksela, T. Wiegand, H.264/AVC in wireless environments, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 657–673.
[31] T. Stockhammer, M.M. Hannuksela, H.264/AVC video for wireless transmission, IEEE Wireless Commun. (2005) 6–13.
[32] 〈http://iphome.hhi.de/suehring/tml/index.htm〉, H.264/AVC Software Coordination.
[33] V. Varsa, M. Karczewicz, G. Roth, R. Sjöberg, T. Stockhammer, G. Liebl, Common test conditions for RTP/IP over 3GPP/3GPP2, ITU-T SG16, Doc. VCEG-N80, September 2001.