SUBJECTIVE QUALITY EVALUATION OF DECODED VIDEO IN THE PRESENCE OF PACKET LOSSES

Tao Liu, Yao Wang,
Polytechnic University, Brooklyn, NY

Jill M. Boyce, Zhenyu Wu, Hua Yang
Thomson Corporate Research, Princeton, NJ

ABSTRACT

This paper investigates the perceptual quality of decoded video bitstreams after packet losses. We focus on low-resolution and low-bit rate video coded by the H.264/AVC encoder, and packet loss patterns likely in 3G wireless networks. We examine the impact of several factors on the perceptual quality, including the error length (the error propagation duration after a loss), the loss severity (measured by the drop in PSNR due to a loss), and loss location (forgiveness effect). We further propose an objective quality measure based on the findings of our study.

Index Terms— Perceptual Quality, H.264/AVC, 3G Networks, PSNR, error propagation

1. INTRODUCTION

In video transmission, the decoder may not receive all the encoded video data because of the losses occurred in various layers of the underlying transmission network. In this case, the perceived video quality depends on many factors, including channel loss characteristics, video encoder configuration, and video decoder error concealment methods, etc.

Although, generally speaking, the average of frame PSNR has been found to correlate reasonably well with perceived quality of decoded video in the absence of transmission losses, it is not at all clear, especially for the newly developed codec H.264/AVC and low-bit rate networks like third generation mobile telecommunication system (3G), whether such a measure bears much resemblance to the perceived quality in the presence of transmission losses [1].

We hypothesize that several factors may have significant impact on the perceptual quality of transmitted videos, including the number of erroneous frames caused by a packet loss and subsequent error propagation (the error length), the severity of a loss (which conceptually refers to the difficulty in concealing a lost frame, and can be measured by the PSNR drop after a loss), the loss position (measured by the time to the end of the sequence, known as the "forgiveness effect"), the number of losses in a sequence, and the loss pattern (clustered or spread), and so on.

In this paper, we describe the subjective tests carried out in our lab, with test sequences constructed to allow us examine the effect of the aforementioned factors. Based on the test results and our analysis, we report the effect of each factor on the perceptual quality. We also propose an objective measure integrating the effects of all factors. Due to limited space, we only describe results pertaining to the case where a test sequence contains a single loss.

2. TEST SET-UP

2.1. Testing method

Two subjective rating methods are recommended by ITU-R BT.500 [2]: the Double Stimulus Continuous Quality Scale (DSCQS) and Single Stimulus Continuous Quality Evaluation (SSCQE). Because we are interested in knowing the quality rating by a user when viewing a video sequence without seeing an error-free version, we choose to display a test video sequence without reference. Although we are primarily interested in the overall rating by a viewer for a sequence, we would like to investigate the immediate reaction of a viewer to a loss, and the impact of this reaction to the overall rating. Therefore, we choose to record both continuous rating and overall rating. Specifically, a viewer is asked to give both continuous-time quality rating (SSCQE) while the sequence is being displayed and an overall quality rating (overall score) (DSCQS, but without references) at the end of the sequence. The viewer uses a mouse to drag a scaling bar to give scores from 0 to 100. ("100" means best quality) both for continuous-time rating and overall rating. A viewer is asked to view several test sequences in each viewing session. The entire procedure is controlled by an interactive rating software that we developed, and a session lasts about 30 minutes without a break. The viewing conditions are set up as described in [2].

2.2. Testing materials

All the testing sequences are cuts (either 20 sec or 40 sec in duration) from one 60 second movie clip, which primarily consists of indoor people interactions, standing，walking and talking, with many face and body movements, some sections with high motion. We choose to use this clip because there are many scene changes and camera panning as well as different types of motion.

All the testing sequences are encoded and decoded following the H.264 standard, using JM10.0 encoder/decoder (baseline profile, level 3, and IDRPP...P GOP structure). The coding frame rate is 12 fps, GOP

lengths are either 2 sec or 4 sec, QCIF resolution, and bit rate is about 128kbps (QP=31 with 2s GOP, average PSNR=35.22 dB or QP=30 with 4s GOP, PSNR=35.72 dB). Each frame is coded into a slice, which is then converted to a RTP packet. We assume each transmission loss leads to the loss of two consecutive frames. This is to take into account the fact that the loss of a single PDU at the link layer of the 3G network typically leads to the loss of two RTP/IP packets, because variable RTP/IP packets are mapped to fixed length PDUs without padding.

In a preliminary study involving 3 viewers, we evaluated decoded videos using three error concealment methods: frame copy, motion copy, and frame freeze [3]. It was concluded that frame copy gives overall more consistent results across viewers and loss patterns, and it gives highest score of the three. Therefore, in the formal tests reported here, we only used the frame-copy method.

The length of a sequence, the encoding parameters (GOP length, the QP), the number of losses and the loss positions within a sequence are varied to produce test sequences with different error characteristics (in terms of length, location, pattern, etc.). A total of 28 sequences are produced. Test1 includes 15 sequences and Test2 includes 13 sequences. Some of the 28 sequences contained multiple losses. Discussion of results pertaining to these sequences is omitted in the following for lack of space.

2.3. Viewers and Viewing Orders

During a viewing session, a viewer rates all sequences in either Test1 or Test2 in a certain order, and then rates them again in the same order, so that he or she gave scores twice for each video sequence. The viewer is not informed about this repetition. The viewing orders by different viewers are randomized so that the same sequence is viewed at different times in a session by different viewers.

The viewers are chosen from Polytechnic students, mostly majoring in electrical and computer engineering or computer science. Totally 30 students participated in the subjective test, with most of the viewers viewed either Test1 or Test2, and a few took part in both.

3. DATA ANALYSIS

3.1. Post-Test Viewer Screening

A viewer may be inconsistent with the majority of the viewers in its rating for the same testing sequence, or he or she may be inconsistent at different times when rating similar sequences. A viewer screening is conducted to eliminate the ratings by these viewers from further data analysis. We conducted the inter-viewer consistency test following [2], and all the viewers passed this test. To test the self-consistency of each viewer, we calculated the correlation coefficient between the two sets of overall scores given by this viewer for the same set of sequences during a viewing session. A total of 6 viewers with low self-consistency are screened out. There are 15 viewers considered valid in each test, and only their viewing scores are included for data analysis.

To perform data analysis on the overall scores by all viewers, the viewer scores are normalized. Specifically, we set each viewer's lowest score to "0" and his/her highest score to "100", and linearly scale the rest of his/her scores.

3.2. Subjective quality vs. error length and PSNR drop

To examine the impact of error length and loss severity on the perceptual quality, we created several sequences with a single loss (losing two consecutive frames) in the middle of a sequence but differing in error length and loss severity. The error length is defined as the number of frames starting from the lost frame to the end of next IDR frame in the bitstream. This is because the error propagation usually does not stop until the I-frame in the next GOP. The only exception is when a scene change happens before the end of the GOP. Sequences with different error lengths are created by varying the loss position within a GOP. To measure the severity of a loss, we first determine the PSNR drop for each affected frame, which is the difference between the PSNR of the frame decoded in the absence of the loss, and the PSNR of this frame decoded with packet loss. We then find the biggest PSNR drop among all affected frames, which is simply referred as PSNR drop. There are totally 8 sequences with similar error length (12 frames) but different PSNR drops, and 6 sequences with similar PSNR drops (about 10 dB) but different error lengths.

Figure 1 shows the relations between the mean overall score among all the viewers and the error length, and that between the mean overall score and the PSNR drop. It is clear that the perceptual quality degrades as the error length or PSNR drop increases, but the relation is non-linear. The beginning flat portions in both Figures 1(a) and 1(b) suggest that viewers do not "see" the loss when the error length or PSNR drop is smaller than a certain threshold, $EL_{\min}$ or $PD_{\min}$, and the end flat portion in Fig. 1(a) suggests that viewers think the sequences are equally "bad" once the PSNR drop exceeds a certain threshold $PD_{\max}$.

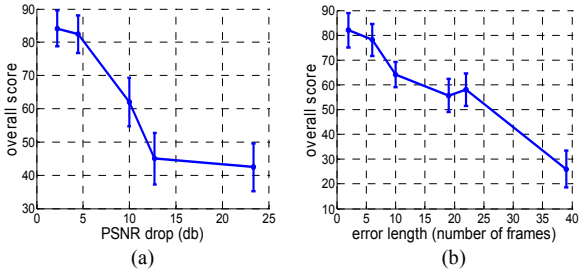
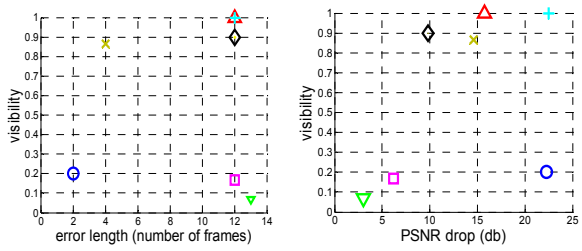


Fig.1 Relations between overall score and (a) PSNR drop (b) error length. The vertical bar around each point indicates the 95% confidence interval.

3.3. Error visibility

Figure 1 shows that when the error length or PSNR drop associated with a loss is smaller than a certain threshold, a viewer tends not to notice the loss. This packet loss "invisibility" phenomenon in MPEG-2 video was discussed in [4], however, to further investigate what affects error visibility in our case, we look into the continuous score curves from the viewers. We observed that, for some particular losses, most viewers did not lower their ratings after the loss, and those losses have either short error length or low PSNR drop. We measure the visibility of an error by the percentage of total number of times viewers "saw" the error, as indicated in the continuous score curves. Fig. 2 shows the relations between the visibility and the error length and PSNR drop, respectively. From Fig. 1 and Fig. 2, we observe that, for our test set-up, the minimal error length for an error to be noticeable ($EL_{\min}$) is about 3~4 frames, and the minimal PSNR drop ($PD_{\min}$) is around 5~7 dB. Note that when two consecutive frames are lost due to a transmission loss, with frame-copy error concealment, the first two frames in each error duration are both copied from the last frame before the loss, which do not incur much noticeable artifacts. Visible artifacts usually start to be seen right after the two concealed frames. This explains why $EL_{\min}$ >=3 in our experiment.



(a) (b)

Fig.2. Relations between error visibility and (a) error length (b) PSNR drop

3.4. Forgiveness effect

Existing research revealed that degradation in video materials can be "forgiven" or "forgotten" to some extent if the degradations are followed by good quality video [5][6]. In other words, for similar perceptual distortion caused by frame loss, the farther that loss locates away from the end of a sequence, the higher will be the subjective rating.

To verify the existence of such "forgiveness" effect and investigate the specific relationship between the perceptual quality and the distance (in time) from the occurrence of the loss to the end of the sequence, we generated 3 sequences, each a 40 second cut from the 60 second clip. Each sequence contains the same GOP with the same single loss (hence same PSNR drop and error length) in the middle. But the three cuts are shifted so that the GOP with loss is positioned in the beginning (15 sec.), middle (25 sec.), and end (35 sec.) of the three sequences, respectively.

Figure 3 shows the relation between the mean normalized overall score and loss position (distance to the end of sequence). From the figure, we find that the overall score for the sequence with loss happening at the end is significantly lower than the overall scores of the other two cases, which received similar ratings. Via paired-T significance test, we confirmed that the difference between the ratings for the sequence with end loss and the two sequences with beginning and middle losses are statistically significant, whereas the difference between the ratings for the two sequences with beginning and middle losses are insignificant. This result substantiates the existence of the "forgiveness" effect, and shows that the increase of the overall score (or the "forgiveness factor") is non-linearly related with the increase of the distance. In our test, the "forgiveness factor" stabilizes after 15 seconds, which means that viewers' memories do not differentiate very well for losses happened 15 seconds before. This result correlates well with those of previous work on the memory effect of human visual systems [7].

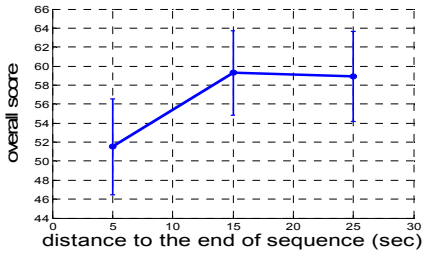


Fig. 3. Relation between the mean normalized overall score and loss position.

3.5. Proposed Objective Measure for Single Loss

Results in previous sections show that both error length and PSNR drop affect the subjective ratings. It is natural to ask if there is some composite objective measure that can reflect the effect of both. Towards this goal, we propose to use "PSNR drop sum" or PDS, which is the sum over the PSNR drops of all erroneous frames, formally defined as

$$PDS = \sum_{n=1}^{EL} PD_n \tag{1}$$

where PD_n is the PSNR drop for frame n, with n=1 denoting the first lost frame. To take into account of the clipping phenomenon shown in Figure 1(a), we further modify the PSNR drop (PD) of each frame by defining:

$$\alpha(PD) = \begin{cases} 0, & PD < PD_{\min} \\ PD - PD_{\min}, & PD_{\min} \le PD \le PD_{\max} \\ PD_{\max} - PD_{\min}, & PD > PD_{\max} \end{cases} \tag{2}$$

Considering the error length threshold, we only sum the PSNR drops for frames after a minimal error-length. The modified PSNR drop sum (MPDS) is expressed as

$$MPDS = \sum_{n=EL_{\min}}^{EL} \alpha(PD_n) \tag{3}$$

For the following results, we used: PD_{min}=5, PD_{max}=14, EL_{min}= 4.

Lastly, we weight the contribution from a single loss based on its distance to the end of the sequence, to take into account of the "forgiveness effect". Based on the non-linear trend observed in Fig. 3, we use an exponential decay weighting factor $w(d) = e^{-\gamma * d}$, where d is the distance (by number of frames) from last erroneous frame to the end of sequence, and γ is a constant that we determine through fitting objective measures to the model (=0.0014 for our test data). This leads to the Weighted MPDS (WMPDS) measure

$$WMPDS = w(d) \sum_{n=EL_{\min}}^{EL} \alpha(PD_n) \qquad (4)$$

Figures 4(a), 4(b) and 4(c) show the relations between the mean overall score and PDS, MPDS and WMPDS, respectively. We can see that MPDS and WMPDS are more linearly related to the perceptual quality. To quantify how well a model fits the measured perceptual ratings, we computed the linear correlation coefficient between the measured ratings and their corresponding PDS, MPDS and WMPDS values for all the testing sequences. The three measures have correlation coefficients of -0.8708, -0.9149, and -0.9275, respectively.

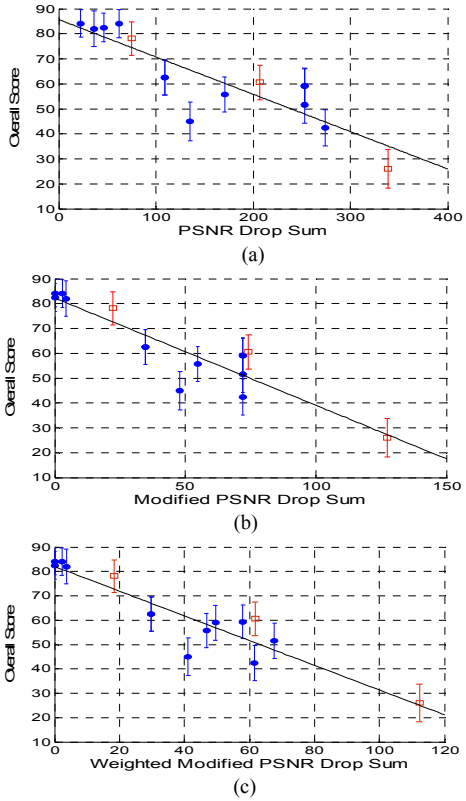


Fig. 4. Relation between overall score and (a) PDS, (b) MPDS, (c) WMPDS. "Solid" points correspond to sequences coded using 2 sec. GOP and QP=31 , and "Circle" points are coded using 4 sec. GOP and QP=30.

Note that the ratings of the sequences coded using QP=30 are generally a little higher than those coded using QP=31. This may be due to the fact that sequences coded using a smaller QP have slightly higher quality in error-free durations (35.72 dB in average PSNR vs. 35.22 dB). The proposed MPDS considers only the effect of erroneous frames on the perceptual quality. We will study how to take into account the error-free quality (or QP) of an encoded sequence in future work.

4. CONCLUSIONS

In this paper, we examined the impact of error length, PSNR drop, and loss location on the perceptual quality of a decoded video subjected to a single transmission loss (which causes the loss of two consecutive frames). We found that an error is visible only if the error length or PSNR drop exceeds a certain threshold, and that the perceptual quality is approximately linearly related to the error length and PSNR drop, subject to some clipping in the beginning and end portion. We proposed an objective measure, MPDS. To take into account the forgiveness effect, we further propose to weight the MPDS of each loss by a distance-dependent weighting factor. Both MPDS and WMPDS measures were shown to correlate quite well with the objective scores for a large set of test sequences.

When a sequence contains multiple losses, the perceptual rating depends on both the PSNR drop sums of individual losses, as well as the loss pattern (whether they are evenly spread or clustered). We hypothesize that the distances among multiple errors may be an important factor affecting overall perceptual quality. As part of the subjective test described here containing 28 test sequences, we have already obtained perceptual quality ratings for sequences with multiple losses which differ in loss number as well as loss pattern. We are in the process of analyzing the data and deriving an objective measure that takes into account loss position and loss pattern.

REFERENCES

[1] S. Winkler, et al.,."Video quality evaluation for mobile streaming applications," *SPIE Conf. Visual Communication and Image Processing,* Lugano, Switzerland, 2003.

[2] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[3] S. Bandyopadhyay, et al., "Frame loss error concealment for H.264/AVC," ISO/IEC MPEG and ITU-T VCEG, JVT-P072, 2005.

[4] S. Kanumuri, et al., **"**Modeling packet-loss visibility in MPEG-2 video,**"** *IEEE Trans. Multimedia*, April, 2006.

[5] V. Seferidis, et al., "Forgiveness effect in subjective assessment of packet video," *Electronics Letters*, 1992.

[6] D.S Hands, "Temporal characterization of forgiveness effect," *Electronics Letters*, Vol.37, 2002.

[7] M. Pinson, and S. Wolf, "Comparing subjective video quality testing methodologies" *SPIE Conf. Visual Communications and Image Proc.,* Lugano, Switzerland, July, 2003.