PERCEPTUAL IMPACT OF BURSTY VERSUS ISOLATED PACKET LOSSES IN H.264 COMPRESSED VIDEO

Ting-Lan Lin and Pamela C. Cosman

University of California, San Diego – ECE

Amy R. Reibman

AT&T Labs – Research

ABSTRACT

When video packets are lost in congested networks, one loss pattern creates a different visual impact than another. We conduct a subjective experiment with H.264 videos and conclude that isolated losses are better than bursty losses in terms of perceptual video quality. A network-implementable video quality model is developed for a router to drop packets so as to achieve good visual quality.

Index Terms— Video codecs, Packet loss, Subjective video quality metrics, FMO (Flexible Macroblock Ordering)

1. INTRODUCTION

In a network, packets can be corrupted in transmission or can be dropped by intermediate routers due to congestion. Considerable research has been done to explore how packet losses impact video quality. In [1], the authors evaluated quality by computational metrics from a no-reference model as well as a subjective test. They found that simple quality metrics (such as blockiness, blurriness and jerkiness) do not predict quality impairments in a packet-loss environment as well as when there are no packet losses. The joint effect of encoding artifacts by MPEG-2 and ATM cell loss ratio was considered [2] where quality was measured with the MPQM perceptual quality metric. In [3], a subjective experiment showed that higher motion videos are more sensitive to cell losses. Also, low visibility of losses could be maintained when network loading increases by increasing the number of videos multiplexed.

Although objective metrics sometimes do not reflect perceptual quality well, PSNR (Peak Signal to Noise Ratio) and MSE (Mean Square Error) are commonly used to measure video quality. The relation between PSNR and perceptual quality scores is considered in [4]. This work on low-resolution, low-bitrate videos found that packet losses are visible when the PSNR drop is greater than a threshold, and the distance between dropped packets is crucial to perceptual quality. The prediction of objective distortion by MSE is discussed in [5], which concludes that bursty losses produce larger distortion than an equal number of isolated losses.

The visibility of lost slices after motion-compensated error concealment was investigated in our prior work in [6, 7]. We showed that most (84%) lost slices are invisible. With this model, we used one bit to denote high/low priorities for slices during encoding so slices of low priority could be dropped first by the router during congestion. The result showed that visibility-based dropping is better in terms of visual quality than the traditional DropTail packet dropping policy which drops consecutive packets off the tail of the queue [8].

In this paper, we consider a different question: if a router must drop L slices and does not have access to visibility-based priority information, which slices should be dropped to minimize the perceptual impact? As a related question, we are concerned with how to group multiple slices into a single packet at the encoder so that, if the packet is lost, the quality degradation will not be as severe. We conducted a subjective experiment to investigate the visual impact of different spatial and temporal patterns of packet loss. Our work differs from [4] in two main respects. They characterize subjective quality based on PSNR, and we linearly regress our subjective data on several different factors. Also, instead of only low-rate coding, we use rates from 200 kbps to 600 kbps. (At high rates, this allows us to characterize the packet-loss effect with minimal encoding artifacts.) In Section 2, four different packet dropping strategies and the evaluation experiment are described. In Section 3, the performances of different dropping methods are compared by non-parametric analysis. Section 4 constructs a prediction model for video quality based on network-accessible factors for a given number of lost packets.

2. EXPERIMENT SETUP

Encoding: We used H.264/AVC JM Version 12.1 to encode SIF resolution videos (15 horizontal slices per frame, 22 macroblocks of 16×16 pixels per slice). The frame rate is 30 fps, and encoding rates are 200, 400 and 600 Kbps. There are 20 frames, IBPBP... in a GOP. Rate control and loop filtering are enabled; the initial quantization parameter is set to 40.

Loss Types : We consider four types or patterns of losses. In Spatial Adjacent (SA), consecutive slices are lost within one frame; this corresponds to the DropTail policy which drops consecutive packets. In Same Frame (SF), slices are lost within one frame but they are not all adjacent (i.e., the

This work was supported in part by the National Science Foundation, the Center for Wireless Communications at UCSD, and the UC Discovery Grant program.

case of SA is excluded). In Same Location (SL), lost slices are in spatially identical locations in different frames. In Different Frame and Location (DFL), lost slices are in different spatial locations and in different frames in the video. For SL and DFL, each frame includes at most one lost slice. Each lossy video has only one loss type involving either L =4, 6 or 8 lost packets. To prevent the case of all lost slices being invisible to the subject, we use the slice-visibility model from [6] to randomly choose one visible slice in the video. We then develop the four loss types based on that location, where the chosen visible slice can be any slice from the 1st to the Lth slice in the loss event. For real-time video, delay exceeding 500ms is unacceptable, so at 30 fps, at most 15 frames can be buffered. So in patterns SL and DFL, we constrain a loss event to span less than 15 frames.

Subjective Evaluation: We define a *comparison set* to be the rating of SA, SF, SL, DFL and No-Loss (NL) versions (in randomized order) of a video at one encoding rate and loss length. The original uncompressed source video can be viewed as a known reference. Given that this original has a score of 100, the viewer is asked for a relative rating for the lossy videos. The rating procedure for a comparison set was realized by SAMVIQ (Subjective Assessment Methodology for Video Quality) [9]. The observer rated the overall quality of a video by sliding a bar on a scale from 0 to 100, marked bad (0-20), poor (20-40), fair (40-60), good (60-80) and excellent (80-100). They can fine tune the score using keyboard up/down arrows, with the numerical score shown to the viewer. They can watch the videos more than once, and can refine the rating.

In each comparison set, if the NL version did not score highest, the viewer was asked to redo that comparison. Most viewers passed this consistency check without redoing the comparison. For the lowest rate videos (200 Kbps), coding artifacts and packet losses are not easily distinguishable, leading some viewers to fail the consistency check initially. But most passed on the second try, and all passed by the fourth try. We generated 360 comparison sets, using 4 original videos with 3 encoding rates and 3 lengths of loss, each of which had 10 different random locations for loss insertion. Each comparison set was rated by 5 subjects. Each subject rated 18 comparison sets (2 original videos, 3 rates, 3 lengths of loss, and one loss location), providing 90 quality labels (5 loss patterns per comparison set), taking about 1 hour to complete. In all, 100 subjects generated 9000 quality labels. To characterize the original videos, we use temporal motion information (tmi) and spatial information (si) as defined in [10]. First we calculate differences of the luminance of successive frames. The tmi is the maximum spatial standard deviation of each difference frame. The si is the maximum spatial deviation of Sobel-filtered luminance frames. The videos used were *Bullfight* (tmi,si=35,73), *Wagon* (24,123), *Sport* (32,63) and *Dancing* (14,85).

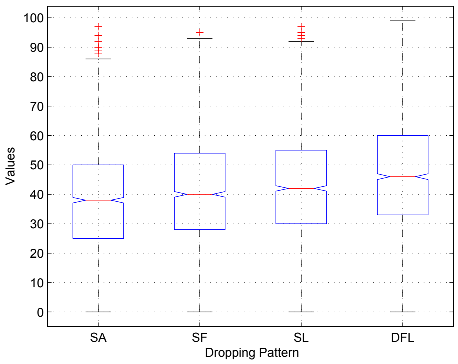
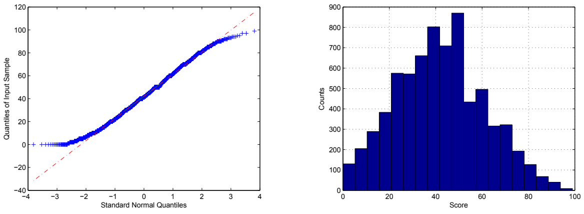


Fig. 1. Boxplot of data from different dropping patterns. Each box indicates 1st quartile, median and 3rd quartile.



(a) (b)

Fig. 2. (a) QQ plot of Score data versus Standard Normal. (b) Histogram of scores obtained from 100 subjects.

3. ANALYSIS OF DROPPING PATTERNS

The purpose of our experiment is to see how each dropping pattern affects perceptual video quality. We start using a boxplot comparison (Figure 1), which shows the differences among four distributions. Specifically, DFL (46.15) has the highest mean perceptual score, followed by SL (42.47), SF (41.29) and SA (38.49) in descending order. So bursty losses degrade the video quality the most, and separating losses both spatially and temporally improves quality more than either individually. In fact, the improvement in mean quality for DFL relative to SA is larger than the sum of the improvements for SL and SF. Note that this boxplot comparison is *unpaired*, meaning that data are not paired by loss length, encoding rates, and source videos in comparing patterns. Paired comparison will be discussed next.

First, we examine the normality of the data with a QQ plot (Quantile-Quantile plot) [11] in Fig. 2(a). At both ends of the curve, there are deviations from normal quantiles. There are hard limits at both ends of the scale; viewers cannot give scores below 0 or above 100. This phenomenon can also be seen in Fig.2(b). By the JB (Jarque-Bera) test [11] for normality, the p-value is essentially zero and thus we reject the assumption that the data are normal. Therefore, we resort to non-parametric paired analysis. The Wilcoxon Signed Rank Test (paired comparison) [12] compares the medians of any pair of dropping patterns in a one-sided test where the H_1 alternative is that the median of one dropping pattern is greater than that of the other. Let x_i and y_i be the values in dropping patterns A and B in the ith comparison set. Define $w = \sum_{i=1}^{n} r_i z_i$ where r_i is the rank of $|x_i - y_i|$ among all $|x_j - y_j|$, and $z_i = 1$ if $x_i - y_i > 0$ and $z_i = 0$ otherwise. Here $n = 1800$ (360×5), the number of total comparison sets. The statistic for the test,

$$Z = \frac{w - [n(n+1)]/4}{\sqrt{[n(n+1)(2n+1)]/24}}, \qquad (1)$$

distributes approximately as Normal(0,1) when $n > 12$. The largest p-value is 0.032 ($< 5\%$), which occurred for the SL-SF comparison, and the other pairwise comparisons had p-values all smaller than 10^{-5}: they all significantly rejected H_0 at the 95% level. Therefore, DFL has the highest median score in a paired comparison with any of the others, followed by SL, SF and SA in descending order, which is consistent with the boxplot of Figure 1. We conclude that dropping packets in isolation, especially in temporal isolation, provides better video quality compared to temporally consecutive or spatially adjacent dropping (e.g., DropTail).

4. VIDEO QUALITY RATIO PREDICTION

After considering the effects of dropping patterns for a range of encoding artifact levels, we now focus on the data at 600kbps. These contain minimal encoding artifacts and let us explore the effects of packet losses alone. To investigate the impact of packet loss given a number L of packets dropped, we define the ratio $R_Q = QualityScore/L$. Given L, we want R_Q to be large. An interesting interpretation of R_Q can be seen from Figure 3. This shows quality from burst loss (SA) and isolated loss (DFL) regressed on packet loss rate. The slopes can be interpreted as the average quality loss per additional dropped slice. As DFL has a less steep slope than SA, dropping an additional packet in isolation hurts quality less compared to an additional packet loss in a spatial burst.

$Factors$	Set1	Set2	Set3
$Constant$	5.890	18.128	18.40
TR	1.218	0.395	0.358
SR	1.177	0.40	0.336
$NumI$	—	-1.74	-1.582
$NumP$	—	-1.80	-1.714
$NumB$	—	-1.54	-1.548
$NumHighPri$	—	—	-0.205
$Correlation$	0.420	0.8551	0.8571

Table 1. Regression coefficients for R_Q and correlation

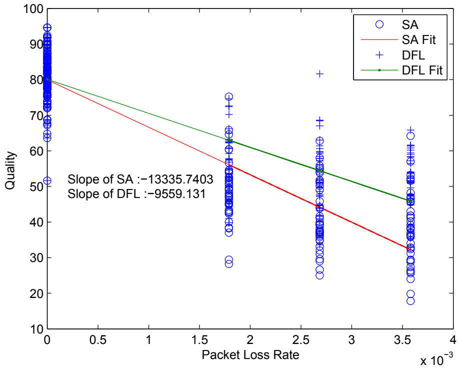


Fig. 3. Scatter plot of quality loss versus packet loss rate.

For a given number of losses, isolation appears to help quality, so this motivates us to quantify the isolation as a predictor of quality score. For a given loss, we define its temporal range (TR) and spatial range (SR) to be the difference between max and min values of the dropped frame number and slice number, respectively, scaled by $1/L$. Given L, if the dropped packets are more spread out temporally or spatially, TR or SR will be higher. Another factor that can be available in the router for predicting quality after packet loss is the type of packet dropped. $NumI$ is the number of dropped packets that are in an I frame, with similar meanings for $NumB$ and $NumP$. Furthermore, we assume that each packet carries one bit of visibility information from the visibility model [6] calculated at the encoder. We say a packet is of high priority when its visibility > 0.25, and define $NumHighPri$ as the number of packets of high priority dropped.

To analyze how each factor affects the responses R_Q, we use linear regression analysis, where the regression coefficients are found by the least-squares approach. Since each comparison was rated by 5 people, we average them as $QualityScore$ for each comparison set. We start from the factor sets that are most easily obtained and then increase the number of factors. Table 1 shows the regression coefficients for different sets of factors for responses R_Q, and their Pearson correlation coefficients between the true and predicted responses, characterizing the prediction accuracy [13]. In Set1, we consider only SR and TR. The table indicates the correlation coefficient is 0.420. Set2 includes the number and type of packets dropped, which increases the correlation coefficient dramatically to 0.8551. Set3 further includes the number of packets of high priority dropped, but this yields limited improvement. One reason may be that the visibility model includes the packet type as an important factor already. This is an encouraging result from a practical point of view: if the encoder includes the single bit of visibility information with each encoded slice, the bitstream would no longer be standard-complaint; however, our model shows that in predicting the quality score, the visibility information is of negligible additional utility after considering other factors. This suggests that the router's choice of which packets to drop from its congested buffer can adequately be made using easily obtainable standard-compliant information $(TR, SR, NumI, NumB, NumP)$. To validate the use of linear regression, the normality of the residuals is tested by the JB test : the p-values are 0.051, 0.057 and 0.062 $(> 5\%)$ for the three models Set1, Set2 and Set3, and thus we cannot reject normality. To conclude, when forced by congestion to drop packets, the router can minimize the damage to video quality by increasing isolation (SR, TR) among the lost packets. But to predict the video quality accurately, the type of packets should be considered.

Packetization for H.264 Flexible Macroblock Ordering (FMO): FMO in H.264 allows packets to contain arbitrary non-consecutive groups of macroblocks from one frame, with the grouping varying from frame to frame. Putting adjacent slices in a packet will likely have the best compression rate due to spatial redundancy; however, separated slices in a packet can minimize damage to video quality when that packet is lost, according to our previous conclusion. The goal here is to develop a model that predicts the visual effect of packet losses within one frame, which can eventually lead to a design algorithm for making FMO decisions.

Since standard-compliant FMO allows packetization only within one frame, we consider here a linear regression model for R_Q that only uses within-frame variables and dropping possibilities. Specifically, we do not include TR, $NumI$, $NumP$, and $NumB$ in our prediction model, and we consider only SA and SF in which lost slices are scattered across one frame. FMO decisions occur at the encoder, so content-specific factors describing each slice can also be included in the model. The significant factors considered, along with their regression coefficients, are listed in Table 2. The resulting correlation coefficient between the actual and predicted scores is 0.8722. The use of linear regression is also validated by the JB test with p-value 0.0714. Therefore, using this table, a packetization strategy could be developed; one could try different combinations of slice groups to achieve the best R_Q.

Conclusion: We showed by nonparametric analysis that bursty losses, corresponding to the traditional DropTail policy, perform worst statistically compared to dropping strategies involving temporal or spatial isolation for packet losses. Furthermore, a video quality model composed of network-accessible factors was developed to provide a dropping criterion for a router. When we concentrate on losses within one frame, factors corresponding to slices are also found correlated to the quality ratio. The relation can be used to attain better video quality for FMO when encoding.

Factor Description	Regress. Coef.
Constant	18.549
Spatial Range (SR)	0.3345
Number of slices in a packet	-1.1086
Average initial MSE (IMSE)	-7.52×10^{-4}
Maximum of IMSE	1.88×10^{-4}
Average motion in y-direction	-2.33×10^{-2}
Maximum magnitude of overall motion	1.13×10^{-2}
Average RSENGY (residual energy after motion compensation)	-1.8×10^{-3}
Maximum RSENGY	-9.27×10^{-5}
Fourth root of MAXIMSE	-0.2108

Table 2. Model factors and their regression coefficients

5. REFERENCES

[1] S. Winkler and R. Campos, "Video quality evaluation for internet streaming applications," *SPIE, Human Vision and Electronic Imaging VIII*, vol. 5007, pp. 104–115, Jan. 2003.

[2] O. Verscheure, P. Frossard, and M. Hamdi, "Joint impact of MPEG-2 encoding rate and ATM cell losses on video quality," *GLOBECOM*, vol. 1, pp. 71–76, 1998.

[3] C. J. Hughes et al., "Modeling and subjective assessment of cell discard in ATM video," *IEEE Trans. Image Process*, vol. 2, pp. 212–222, April 1993.

[4] T. Liu, Y. Wang, J.M. Boyce, Z. Wu, and H. Yang, "Subjective Quality Evaluation of Decoded Video in the Presence of Packet Losses," in *ICASSP*. IEEE, April 2007, pp. 1125–1128.

[5] Y. J. Liang et al., "Analysis of Packet Loss for Compressed Video : Does burst-length matter?," in *ICASSP*. IEEE, 2003, vol. 5, pp. 684–687.

[6] S. Kanumuri, S. G. Subramanian, P. C. Cosman, A. R. Reibman, and V. Vaishampayan, "Predicting H.264 Packet Loss Visibility using a Generalized Linear Model," in *ICIP*. IEEE, 2006, pp. 2245–2248.

[7] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video," *IEEE Trans. Multimedia*, vol. 8, pp. 341–355, April 2006.

[8] S. Kanumuri, *Packet Loss Visibility and Packet Prioritization in Digital Videos*, Ph.D. thesis, UCSD, 2006.

[9] "ITU-R Recommendation BT.500V11: Methodology for the Subjective Assessment of the Quality of Television Pictures," *ITU-R 211/11*, 2004.

[10] "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," *ITU*, 1999.

[11] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer, 2nd edition, 2002.

[12] R. Larsen and M. Marx, *An Introduction to Mathematical Statistics and Its Applications*, Pearson Edu,4th ed.

[13] "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," VQEG, 2000.