Visibility of individual packet losses in MPEG-2 video

Amy R. Reibman
AT&T Labs – Research
amy@research.att.com

Sandeep Kanumuri
Univ. Calif. at San Diego
skanumur@code.ucsd.edu

Vinay Vaishampayan
AT&T Labs – Research
vinay@research.att.com

Pamela C. Cosman
Univ. Calif. at San Diego
pcosman@code.ucsd.edu

***Abstract*— The ability of a human to visually detect whether a packet has been lost during the transport of compressed video depends heavily on the location of the packet loss and the content of the video. In this paper, we explore when humans can visually detect the error caused by individual packet losses. Using the results of a subjective test based on 1080 packet losses in 72 minutes of video, we design a classifier that uses objective factors extracted from the video to predict the visibility of each error. Our classifier achieves over 93% accuracy.**

I. Introduction

Since the first papers on video transport over networks appeared, a long-standing problem has been "What packet loss rate (PLR)[1] can viewers accept?". Target thresholds on acceptable PLR have ranged from 10^{-9} or lower [1], [2] to 10^{-6} [3], with even higher PLR assumed in recent work. Clearly, the choice of the proper threshold is confounded by a number of issues.

1) Viewer expectations, involvement, and task. For example, viewers may tolerate errors if they can't get that video content any other way but not if they usually receive the same service error-free.
2) Environmental viewing conditions. Background lighting, monitor characteristics, and viewing distance all affect the viewing experience.
3) Each loss creates an error with a different visual impact. Encoding parameters, decoder concealment, video content (moving or still), packet size, burstiness of losses, and the actual location of the loss; these all play fundamental roles in whether a particular loss will be visible or invisible to most viewers.

A number of subjective studies are available regarding the perceived impact of packet losses [4], [5], [6], [7]. The effect of losses depends heavily on the scene content and amount of motion [4]. However, to our knowledge, all attempts to consider the perceptual impact of packet losses have examined the combined impact of multiple packet losses. Because not all packet losses create the same visual impact, different realizations of video content and packet loss may lead to vastly different visual quality. Thus, in these studies, many different realizations of both packet loss and video content are necessary to reduce the variability of the observer responses.

In this paper, we take a different approach, by considering the visual impact of each individual packet loss. Our eventual goal is to create a network-based video quality monitor that is real-time, per-stream, and accurate enough to answer the question: do the *specific* packet losses affecting this *specific* video being transported degrade its visual quality?

To accomplish this, we take the following steps. First, we conduct a subjective test in which viewers who are shown MPEG-2 videos with injected packet losses are asked to indicate when they see an artifact in the displayed video. Data is gathered for a total of 1080 individual packet losses over 72 minutes of MPEG-2 video. We purposely leave open the question of when a viewer will find a given frequency of visible packet losses to be objectionable or annoying. "Ground truth" regarding packet loss visibility is defined by the results of these subjective tests.

The data gathered from the subjective test could be correlated with the output of any number of objective quality metrics (including [8], [9], [10]), to understand how these can be used for characterizing packet loss errors. However, because we are interested in monitoring the video quality within the network [11], we would like a metric that operates on the compressed bitstream. Most available video quality metrics require either video information prior to encoding, or a completely decoded bitstream (or both).

Therefore, our second step is to develop a tree-based classifier that labels each possible packet loss as either visible or invisible. The classifier uses objective factors extracted from the video, including factors that are independent of the video content (temporal duration, initial spatial extent, and vertical position) and factors that depend on the underlying video content (motion and initial error). We achieve better than 93% classification accuracy.

This paper is organized as follows. Section II gives an overview of MPEG-2 packet losses and their impact. We also describe objective factors that are relevant in predicting packet loss visibility. Section III describes our subjective test. In Section IV, we describe our tree-based classifier that predicts the visibility of each packet loss. Section V concludes.

[1]We consider here only those packets which are still lost after all error control (including FEC, retransmission, etc). Thus, we consider the PLR seen specifically by the compressed video decoder.

II. Packet loss in MPEG-2 video

MPEG-2 is typically packetized in one of two ways. First, video can be segmented and packetized into small fixed-size packets (like ATM cells or MPEG-2 Transport Stream packets), in which case a single packet loss might force the decoder to discard either a slice or an entire frame. Second, a variable-sized packet can contain one or more entire slices. In both cases, a packet loss corresponds to the loss of one or more slices.

The initial error caused by a packet loss propagates in space and time as a result of the video decoding algorithm. The exact error due to packet loss can be completely described by (a) the initial error for each macroblock in the lost packet, and (b) the macroblock type and (c) motion information for subsequently received macroblocks [11]. The latter two control the temporal duration and spatial spread of the error.

We expect the visibility of a loss to depend on a complex interaction of its location, the video encoding parameters, and the underlying characteristics of the video signal itself. For example, the texture and motion of the underlying signal may potentially mask the error. To isolate the impact of the various parameters, one approach could be to inject different error amplitudes against an identical signal background, as was done in [12] for blocky, blurry, and noisy artifacts. However, for packet losses, the error itself is highly dependent on the underlying signal. Therefore, we must take a different approach.

We have independent control over the location, initial spatial extent, and temporal duration of each loss we inject. The other factors depend on the signal. Thus, we can choose whether to lose a single slice, multiple slices, or an entire frame, and we can choose the loss to be in a B-frame (which would last a single frame) or in a reference frame (which would typically last more than one frame). In choosing the location of the loss, we should distribute the locations vertically within the frame, and we should also choose representative samplings from both still and active regions of the sequence.

III. Subjective test

We use a single-stimulus test, in which the viewers' task was to indicate when they saw an artifact, where an artifact was defined simply as a glitch or abnormality. We wanted viewers to be immersed in the viewing process and not scrutinizing the video for any possible impairment. Thus we chose DVD-quality MPEG-2 video[2] from travel documentaries. Audio was not presented, and the video decoder used zero-motion concealment.

We chose twelve 6-minute DVD-quality video sequences, for a combined length of 72 minutes. We grouped the sequences into 4 sets, each consisting of three of the 6-minute sequences. This limited a viewing session to 18 minutes so as not to tire or bore the viewers. During each 18-minute viewing session, a viewer evaluated a set of video with a short break between each sequence. Some viewers participated in more than one viewing session, although never on the same day. Each set of video (and hence each packet loss) was evaluated by 12 viewers.

Viewers were told that the videos they were watching would have impairments caused by packet losses, and that when they saw something unexpected in the video they should respond by pressing the space bar. They were asked to keep their finger on the space bar so they would not be distracted by that task. The lighting condition was typical of an office environment and the viewer was positioned approximately six picture heights from the screen.

A total of 1080 packet losses were randomly injected in these videos such that every non-overlapping four-second interval contained one packet loss in the first three seconds. The one-second guard interval ensured a viewer had sufficient time to respond to each individual error. Inside the three-second interval available for each loss, we distributed the losses such that overall, 30% affected an entire frame, 10% affected two adjacent slices, and 60% affected a single slice. Further, we chose to have 30% of the losses to be in B-frames (and hence have a temporal duration of one frame), and the remaining 70% evenly distributed across the available P- and I-frames in the 3-second interval. Finally, the video we selected was highly varied, with many different motion types and amounts of spatial texture. Therefore, we believe our packet losses occur across a representative set of diverse signal background types.

We label each of the 1080 packet losses with the responses from the 12 viewers: seen or not seen. Figure 1 shows the histogram of the number of viewers who responded to each packet loss. From these responses, we define the ground truth regarding the visibility of an error. We define an error to be *visible* if 75% or more viewers responded to it. Similarly, an error is *invisible* if 25% or fewer viewers responded to it. The remaining errors are *indeterminate*. Of the 1080 total errors shown to viewers, 732 were invisible, 195 were visible, and 153 were indeterminate. We do not concentrate here on the 14% of errors that were indeterminate, but instead focus on understanding the 927 visible and invisible errors.

IV. Objective factors and classifier

In this section, we consider objective factors that can be extracted from a complete video bitstream. We first examine the effect of individual factors on the visibility of packet losses as defined by our human viewers, and then present several objective classifiers based on these factors.

A. Factors affecting visibility

We consider a total of nine objective factors. We consider first three content-independent factors: temporal duration (TMDR), initial spatial extent (SPXNT), and the vertical position (or height) of the error (HGT). The values of these three factors were chosen for each packet loss

[2] 720 pixels, 480 lines, and 60 fields per second.

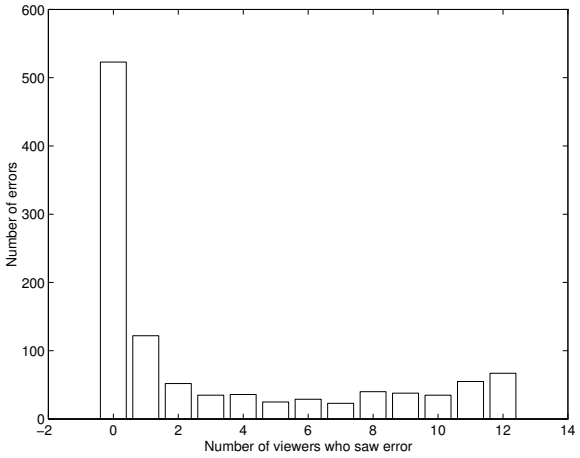

Fig. 1. Number of viewers who saw each error.

when we designed the subjective test, as described above. They do not depend on the underlying video content, and they can be easily computed or extracted from a partially-received bitstream using only the information in the video headers.

Overall, the correlation between TMDR, SPXNT, and HGT and the number of viewers who saw each error is low: 0.051, 0.29, and -0.13, respectively. However, some trends can be observed. Of those errors with TMDR=1 (i.e., B-frame errors), only one error is visible and the remaining 119 are invisible. Of the full-frame errors, 39% are visible, while only 13% of the single- and double-slice errors are visible. A higher percentage of losses in the bottom third (84.5%) are invisible than those in the top (73.3%) or middle third (69.8%).

Next, we consider content-specific factors that depend on the video content at the location of the loss: motion and the MSE of the initial error (IMSE). For a particular packet loss, these content-specific factors cannot be exactly obtained from a bitstream in which the packet is already lost; however, they are available from the complete bitstream. We average these content-specific factors across all macroblocks initially lost.

Intuitively, motion plays a key role in the visibility of losses. We use an MPEG-2 encoder to compute motion vectors for each pair of adjacent frames and summarize the motion within a slice by the average motion in both the x- and y-directions, denoted by MOTX and MOTY respectively. In addition to average motion, we also consider several other motion-related variables to predict the visibility of packet losses: the variance in MOTX and MOTY denoted by VARMX and VARMY, respectively, and the energy in the residual after motion compensation, RSENGY, for each slice. These additional parameters help in determining whether the calculated motion vectors represent the underlying scene motion well or not. For example, if RSENGY is high, the motion vectors probably do not represent the actual scene motion well. Similarly, if the motion variance is high, the motion is inconsistent.

The initial Mean Square Error of a packet loss, IMSE, has a significant impact on the visibility of a packet loss. For the frame affected by a packet loss, we compute the IMSE by evaluating the per-frame MSE between the decoded images using the complete bitstream and the bitstream with loss. Assuming zero-motion concealment, IMSE is easily computed using a decoder that receives the entire bitstream (without losses).

The correlation coefficients between MOTX, MOTY, and IMSE and the number of viewers who saw an error are 0.40, 0.26, and 0.44, respectively. Not unexpectedly, visible errors are much more likely to have large motion and large IMSE than invisible errors. Most errors with small motion are invisible to most viewers: only 11 out of 330 packet losses with both MOTX and MOTY less than 0.5 (half a pixel) are visible.

B. Objective Classifiers

We compare four objective classifiers that classify each packet loss to be visible or invisible to an "average" human observer. Each classifier is a decision tree; the classifier traverses a tree where the path at each node is based on a binary decision using one of the nine factors discussed above. During the training of the tree, a node is split to minimize the probability of misclassification.

Two of the classifiers start with a sub-tree, denoted **root-tree**, that is based on our earlier observations regarding the impact of short temporal duration, small motion, and spatial extent. **Root-tree** consists of the following decisions. First, all packet losses with temporal duration of one frame (TMDR$\leq$1) are classified as *invisible*. This introduces only one misclassification. Second, all packet losses with small motion, defined by (MOTX$\leq$0.5 & MOTY$\leq$0.5), are classified as *invisible*. This results in 11 misclassifications. At this stage, we have classified 450 of the 927 errors, with only 12 misclassifications. Next, we split the tree based on the initial spatial extent (SPXNT$<$15) without making any decision. This split is based on the earlier observation that sub-frame (single- and double-slice) and full-frame losses have different probabilities of being visible and allows these two cases to be treated differently.

At this stage, we apply CART [13], a well known statistics tool for tree-structured data analysis, to classify the data in each of the two nodes. CART splits the frame-loss node using (IMSE$\leq$55.947), where losses with smaller IMSE are classified as *invisible* and the remaining losses are classified as *visible*. For the slice-loss case, CART produces the tree shown in Figure 2. Seven of the eight available parameters are used in decisions, but TMDR is not used. This is the first classifier we consider.

We note that in Figure 2, the initial decision is based on IMSE. Because CART uses a greedy algorithm to find the best split, we conclude that after partitioning the data set using **root-tree**, the single most important factor at that stage for predicting visibility is IMSE. Thus, the second classifier we consider uses only **root-tree** and IMSE. It

Classifier	Misclass.		Accuracy	
	RS	CV	RS	CV
root-tree + 9-factor CART	36	62	96.12%	93.31%
root-tree + IMSE	78	82	91.59%	91.15%
9-factor CART	50	71	94.61%	92.34%
1-sec MSE	122	127	86.84%	86.30%

TABLE I

MISCLASSIFICATIONS AND ACCURACY FOR EACH CLASSIFIER, DURING RESUBSTITUTION (TRAINING) AND CROSS-VALIDATION.

classifies full-frame errors as described above and classifies sub-frame errors as *invisible* if (IMSE$\leq$18.834) and as *visible* otherwise.

Our third classifier is designed by applying CART to all nine factors above. The decision tree for this classifier is not shown due to space constraints; however, it contains 10 terminal nodes, and the initial split is based on IMSE.

The fourth classifier is designed using linear regression applied to the one-second MSE for each packet loss. Despite its known shortcomings at accurately characterizing video quality, MSE has traditionally been used to evaluate the impact of packet loss, since it summarizes the overall impact of the packet loss. We measure MSE between the decoded video without loss and the decoded video with each packet loss, over any one-second interval that contains the error. With this classifier, errors with one-second MSE smaller than 3.621 are classified as *invisible*.

C. Classifier Performance

We use the four classifiers to classify all non-indeterminate packet losses into visible and invisible losses. The classifiers are not run on indeterminate losses for lack of a ground truth for comparison. The performance of the four classifiers is shown in Table I. Entries under "RS" correspond to the performance during the resubstitution phase, which classifies the training data. Entries under "CV" correspond to the cross-validation phase, where the data is partitioned into 10 equal partitions and each of the 10 partitions is classified using a tree trained using the other nine partitions.

The CART-based classifier that begins with **root-tree** performs best, achieving over 93% accuracy on the cross-validation test set. The classifier obtained using CART alone does not perform as well because CART is a greedy algorithm; each split is only locally optimal, and may not be the best split for the overall tree. The split on spatial extent is such an example, where CART does not see that a split on spatial extent is advantageous at the global level. Using **root-tree** plus IMSE achieves over 91% accuracy on the cross-validation test set, while using only the one-second MSE performs the worst.

V. CONCLUSIONS

When evaluating the quality of video in a network, PLR alone is insufficient because the impact of losses depends heavily on video content. Our study here is a first step toward developing accurate perception-based video quality monitors within the network. Further work is needed to incorporate our current classifier into a video quality monitor that uses only information available from bitstreams that already have losses [11]. Further work is also needed to generalize our classifier to other environments: non-isolated errors, or different compression algorithms.

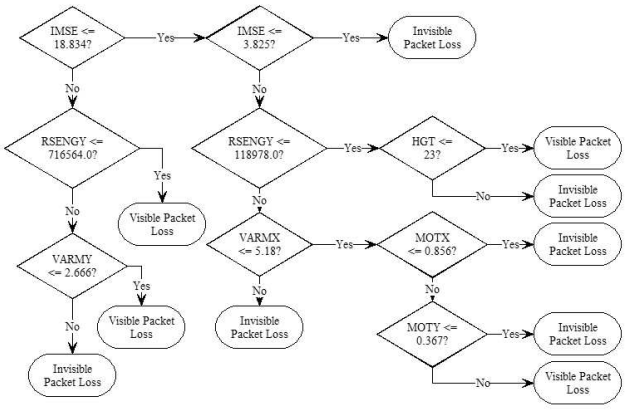

Fig. 2. Classifier for SPXNT<30, TMDR>1, not-small motion.

We expect the Visible Packet Loss Rate (VPLR), i.e. the rate of losses causing visible errors, will be more meaningful than PLR, because a threshold on acceptable VPLR is likely to be invariant with scene content. Our classifier could be used to assign to each scene a *probability of visibility* for a random loss. Coupled with an understanding of how people tolerate frequent visible errors, this could allow better network design for video transport.

REFERENCES

[1] W. Verbiest and L. Pinnoo, "A variable bit rate video codec for asynchronous transfer mode networks", *IEEE Journal on Selected Areas in Comm.*, vol. 7, no. 5, pp. 761–770, June 1989.

[2] W. Verbiest et al. "The impact of the ATM concept on video coding", *IEEE JSAC*, vol. 6, no. 9, pp. 1623–1632, Dec. 1988.

[3] R. Aravind et al. "Packet loss resilience of MPEG-2 scalable video coding algorithms", *IEEE Trans. on Circuits and Syst. for Video Tech.*, vol. 6, no. 5, pp. 426-435, Oct. 1996.

[4] C. J. Hughes et al. "Modeling and subjective assessment of cell discard in ATM video," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 212-222, April 1993.

[5] S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks", *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 12, no. 12, pp. 1071-1083, Dec. 2002.

[6] B. Chen and J. Francis, "Multimedia Performance Evaluation", AT&T Technical Memorandum, February 28, 2003.

[7] Verizon Laboratories (Gregory W. Cermak), "Videoconferencing Service Quality as a function of bandwidth, latency, and packet loss", T1A1.3/2003-026, May 6, 2003.

[8] S. Hemami and M. Masry, "Perceived quality metrics for low bit rate compressed video", *Int. Conf. on Image Proc. (ICIP)*, pp. 721–724, Sept. 2002.

[9] H. R. Wu et al. "Vision-model-based impairment metric to evaluate blocking artifacts in digital video", *Proceedings of the IEEE*, vol. 90, no. 1, pp. 154 –169, Jan. 2002.

[10] S. Wolf and M. Pinson, "In-service performance metrics for MPEG-2 video systems", IAB, Montreux, Switzerland, Nov 12-13, 1998.

[11] A. R. Reibman et al. "Quality monitoring of video over a packet network", *IEEE Trans. Multimedia*, to appear, 2004.

[12] M.S. Moore et al. "Defect visibility and content importance implications for the design of an objective video fidelity metric " *ICIP 2002*, pp. III.45-48, June 2002.

[13] L. Breiman et al. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.