PERCEPTUAL QUALITY BASED PACKET DROPPING FOR GENERALIZED VIDEO GOP STRUCTURES

Ting-Lan Lin[◇] *, Yuan Zhi*[◇] *, Sandeep Kanumuri* [⋆]
Pamela C. Cosman[◇] *, Amy R. Reibman*[†]

[◇] Dept. of ECE, UCSD [†] AT&T Labs Research
[⋆] DoCoMo Communications Laboratories USA, Inc.

ABSTRACT

Our work builds a general visibility model of video packets which is applicable to various types of GOP (Group of Pictures). The data used for analysis and building the model come from three subjective experiment sets with different encoding and decoding parameters on H.264 and MPEG-2 videos. We consider factors not only within a packet but also across its vicinity to account for possible temporal and spatial masking effects. This model can be useful for an intermediate router in a congested network to drop less visible packets to maintain overall video quality. Experiments are done to compare our perceptual-quality-based packet dropping approach with existing Drop-Tail and Hint-Track-inspired cumulative-MSE-based dropping methods. The result shows that our dropping method produces videos of higher perceptual quality for different network conditions and GOP structures.

Index Terms— Video coding, Perceptual video quality model, Packet visibility, Packet prioritization, Packet discarding policy.

1. INTRODUCTION

For video transmission in a network, video quality at the receiver can be highly affected by packet losses. Prior research to understand the relationship between packet losses and visual quality degradation includes [1, 2, 3, 4], where average packet loss rate (PLR) was used to model average video quality. In [3], a random neural network model was used to assess quality given different bandwidth, framerate, packet loss rate, and I-block refresh rate. Subjective video quality was modeled in [5] using packet loss locations, cumulative MSE (sum of MSE over all frames affected by the packet loss) and error propagation. The prediction of objective distortion by MSE is discussed in [6], and [7] uses three different metrics to estimate the MSE caused by a packet loss.

Rather than studying how packet losses affect overall perceptual quality, or relate to MSE, our past work concentrated on characterizing the visibility of a packet loss. We conducted a subjective experiment on packet losses in MPEG-2 bitstreams to develop a classifier for visible/invisible packets, and a generalized linear model (GLM) [8] to predict packet loss visibility [9]. The visibility model for H.264 video packets is developed and discussed in [10]. And in [11, 12], we found that the factors in a packet's spatial or temporal neighborhood, as well as factors related to proximity to a scene cut and camera motion, are important to predict the visibility of packet losses. GLM visibility models across different codecs, coding parameters (such as GOP types) and error concealment strategies were also built in [11, 12].

In this paper, we present a new packet-loss visibility model based on a more general strategy for factor inclusion than in [11, 12]. The goal of this paper is to demonstrate the effectiveness of a packet prioritization application based on our visibility model. Using packet priorities, an intermediate router can intelligently drop low-priority packets. For existing approaches on packet classification, [13] sets a packet as high/low priority based on the cumulative MSE due to the packet loss, network status and end-to-end QoS constraint. Also, the Rate-Distortion Hint-Track method was proposed in [14, 15], where a congested intermediate router drops packets from different streams to minimize the sum of the cumulative MSE, constrained on the sum of the outgoing rates to be less than the bandwidth of the outgoing link. Note that the cumulative MSE is computationally expensive to measure since it includes the MSE due to error propagation. Therefore, instead of the cumulative MSE, the *initial* MSE, which only considers errors in the lost packet, is used for our factor consideration, and it shows good correlation with visibility. The most significant difference between our approach and the above-mentioned methods is that we do not use MSE (or PSNR) as a quality metric to develop our method; our model is built from subjective experiments. We compare our visibility-based packet dropping strategy with the cumulative-MSE-based method and the widely-used Drop-Tail policy with simulations using NS-2 [16]. The comparisons are made using the well-known perceptual quality metric VQM (Video Quality Metric) [17].

This paper is organized as follows: Section 2 introduces the experiment settings for three different subjective tests. Section 3 describes the analyzed factors for visibility in terms of signal attributes classified by the level of accessibility. In Section 4, we illustrate our GLM model building strategy integrating multiple data sets with different sizes and incorporating significant factors. Section 5 presents the experimental comparison and conclusions.

2. DATASETS FOR THREE SUBJECTIVE TESTS

To develop a robust visibility model for packet losses, we collect data from three subjective experiments [9, 10, 18] with different encoding and decoding parameters in MPEG-2 and H.264. The parameters are summarized in Table 1. The encoding rates in the three tests were set such that there are no obvious encoding artifacts. This allows us to concentrate on impairments induced by packet loss. H.264 videos have one slice (a row of macroblocks) per Network Adaptation Layer Unit (NALU), and each packet loss is equivalent to the loss of one slice. For MPEG-2 videos, we packetize the video by taking fixed

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

	Test 1 [18]		Test 2 [9]	Test 3 [10]
Spatial resolution	720x480		720x480	352x240
Frame rate (fps)	30	24	30	30
Duration (minutes)	7.3	8.9	72	36
Standard	MPEG-2		MPEG-2	H.264
GOP structure	I-B-B-P-		I-B-B-P-	I-B-P-
I-frame insertion	adaptive		adaptive	fixed
GOP length	≤ 13	≤ 15	≤ 13	20
concealment	Default		ZMEC	MCEC

Table 1. Summary of subjective tests' parameters and their datasets

size segments from the bitstream. Therefore, it is possible that a packet loss leads to two consecutive slice losses. A whole frame loss is also possible if the lost packet is a frame header. The decoder concealment strategies are Default, ZMEC (Zero-Motion Error Concealment), MCEC (Motion-Compensated Error Concealment) for the three tests. The videos used across the three experiments had different motion and spatial texture, type of camera motion and object motion. Refer to [12] for details of these three experiments.

All three experiments aimed to obtain ground truth for each packet loss. Each packet loss was evaluated by 12 viewers, and whenever they see a visible artifact or a glitch, they respond by pressing the space bar. We calculated the probability of visibility of a loss as the number of viewers who saw the loss divided by 12.

3. DESCRIPTION OF ANALYZED FACTORS

In this section we discuss factors that affect the visibility of packet losses. We define the original video frame at time t as $f(t)$, the compressed video frame as $\hat{f}(t)$, and the decoded video frame as $\tilde{f}(t)$ (with possible packet losses). The error is $e(t) = \hat{f}(t) - \tilde{f}(t)$. The factors are classified by accessibility of the original video as a reference when measured. To reduce the complexity of the model, here, unlike in [12], we do not consider full-reference factors. Also, we do not restrict ourselves to use only average measurements or maximum measurements as [12] does. Due to space constraints, in the following, we only introduce factors that are significant to and incorporated in our final model.

3.1. Reduced-Reference (RR) measurements

The RR measurements for a packet can be obtained when a video encoder or video server reliably provides per-MB information based on $e(t)$, $\hat{f}(t)$ and $\tilde{f}(t)$, assuming knowledge of the decoder concealment strategy. We found that **IMSE**, **ISSIM** (the average MSE and SSIM [19] among all MBs in an initial packet loss) and **MaxIMSE** (the maximum per-MB MSE over all MBs in the initial packet loss) are significant to the packet loss visibility. To measure the motion information (x,y) per MB that is independent of any codec, we do a forward motion estimation using 16x16 motion blocks from the *uncompressed* signal $f(t)$. (MOTX,MOTY) is the average motion vector, and **ResidEng** is the average residual energy after motion compensation, over MBs in a packet. We define high motion **HighMOT** to be TRUE if $MOTM = \sqrt{MOTX^2 + MOTY^2} > \sqrt{2}$.

Reference-Scene-related factors are shown to be important by exploratory data analysis (EDA) in [12]. A method for detection of quick scene cuts from $\hat{f}(t)$ was presented in [20]. We label each packet loss by the distance in time between the first frame affected by the packet loss and the nearest scene cut, either before or after. This is DistFromSceneCut, and is positive if the packet loss happens after the closest scene cut in display order, and negative otherwise. DistToRef per MB describes the distance between the current frame (with the packet loss) and the reference frame used for concealment. This variable is positive if the frame at which the packet loss occurs uses a previous (in display order) frame as reference, and negative otherwise. We define **FarConceal** to be TRUE if MaxDistToRef (maximum of |DistToRef| in a slice) ≥ 3. In this inequality, MaxDistToRef has units of frames. We also define a Boolean variable, **OtherSceneConceal**, which is TRUE if |DistFromSceneCut| $<$ |MaxDistToRef|, where the compared variables must be of the same sign (same direction). In this inequality, the compared variables have units of seconds. If the compared variables have different signs, OtherSceneConceal is FALSE. OtherSceneConceal describes whether the packet loss will be concealed by an out-of-scene reference frame which will increase the visibility of packet loss. To account for the depressed visibility immediately *before* a scene cut, we define **BeforeSceneCut** to be TRUE if $-0.4sec < \mathrm{DistFromSceneCut} < 0sec$. We classify scenes based on four camera-motion types: still, panning, zooming, or complex camera motions. Our previous work [12] shows significantly fewer viewers saw packet loss in still scenes than in panning or zooming scenes. Therefore, we define **NotStill** to be TRUE if motion type is not still.

3.2. No-Reference (NR) measurements

NR factors can be measured from the lossy pixels only (NR-P), lossy bitstream only (NR-B), or both bitstream and pixels (NR-BP). Factors found by these methods describe *exactly* the spatial extent, pattern, location, and temporal duration of the loss. Variants of these factors that are significant to the visibility are defined as follows: **SXTNT2** is TRUE when two consecutive slices are lost, and **SXTNTFrame** means all slices in the frame are lost. **Error1Frame** is TRUE if the packet loss lasts only one frame. As in our past work, we adopt the NR-B measurements to directly obtain the above factors. The signal $\hat{f}(t)$ at the location of the impairment can be estimated with either NR-P or NR-B using information from neighboring unimpaired frames.

4. GENERALIZED LINEAR MODEL

This section describes our proposed packet visibility model. We choose a GLM to predict the probability of a packet loss being observed [9]. Since our three datasets have different sample sizes, for fairness, we use an equal number of samples from each of the datasets as training data, and the remaining samples in each dataset as test data. We then apply the method discussed in [9] to estimate the model parameters from the *training set*, and evaluate the performance error of the fitted model using the *test set* as follows:

$$q = \frac{1}{3}\sum_{k=1}^{3}\left[\frac{1}{N_k}\sum_{i \text{ in TestSet } k}(p_i - \tilde{p}_i)^2)\right], \qquad (1)$$

where $\tilde{p}_i$ is the predicted fraction of viewers who saw the i^{th} packet loss, and N_k is the number of samples in the test set of Dataset k. We choose four-fold cross-validation: we repeat the fitting process four times with 4 non-overlapping training sets, producing 4 fitted models and $q_j, j = 1, 2, 3, 4$. We then repeat this procedure for four different random seeds for different partitions of the data to validate the model. We define the average performance error of these sixteen

Factors	Coeff. for Final Model
Intercept	4.18061
$\log(1 - \text{ISSIM} + 10^{-7})$	0.22871
SXTNT2	-0.41208
SXTNTFrame	-1.47672
Error1Frame	-0.33009
$\log(\text{MaxIMSE} + 10^{-7})$	0.27578
$\log(\text{ResidEng} + 10^{-7})$	-0.61219
HighMOT	0.18290
NotStill	0.73364
BeforeSceneCut	-1.14434
OtherSceneConceal	2.08966
$\log(\text{IMSE} + 10^{-7})$	0.30492
$\log(\text{IMSE} + 10^{-7})$: FarConceal	0.25720

Table 2. Factors of the final model. Note that the colon (:) means "interact with"

models as Q. For further factor refinement, we use Q to decide if a specific factor is significant and should be included in the model: for each considered factor added to the model, we calculate a Q by the 4-seeds-4-folds GLM modeling process. We *include* a factor only if the model with that factor included has smaller Q than the model without that factor. To obtain the factor coefficients, we use the fitting from the seed that achieved the lowest performance error. The factors and coefficients of our final model are summarized in Table 2. Since the model is developed based on data from different GOP types, and the factors are not GOP-type-specific, our generalized packet visibility model can be used in various GOP types.

5. EXPERIMENT RESULTS

Our generalized-GOP visibility model can serve as a tool to prioritize packets. In the server, we compute the visibility of each packet based on its relevant information as shown in Table 2. Each packet is then labeled by 1 bit as low priority when its visibility is less than 0.25, and high priority otherwise. This is a perceptual-quality based packet prioritization (PQ) method. To compare with the notion of using cumulative MSE in Hint Track [15], a packet prioritization method, cMSE (Cumulative MSE prioritization method), is designed. At the encoder, the cumulative MSE is computed by summing MSE over all the frames that are affected by a packet loss. Then we assign a 1-bit priority by comparing the cumulative MSE of each packet to a threshold, which is derived such that we have approximately the same number of high-priority packets for a given video stream for both cMSE and PQ prioritization. For both policies, the intermediate router can drop the packets of low priority in case of network congestion, using the priority bit. One widely-implemented packet dropping approach is Drop-Tail (DT), which drops packets at the end of the buffer queue in the router when the network is congested. We will compare our approach with these two methods in terms of received video quality, measured by the VQM (Video Quality Metric) developed by ITS [17].

We simulate the experiment using NS-2 [16] in a network topology (Fig. 1) where two videos are transmitted simultaneously as sources S1 and S2 (variable-bit-rate encoded in $r1$ and $r2$ bps on average) to destination D. Packets belonging to both videos compete for space in the queuing buffer (of size BF bits) at intermediate node I. The bottleneck link's bit-rate is constant at R bps. When instantaneous rates of S1 and S2 sum to more than R, packets accumulate in the buffer. If this condition persists, the buffer will eventually overflow and packets are dropped in accordance with a policy. At destination D, the quality of received videos is evaluated using VQM, which ranges from 0 (excellent quality) to 1 (poor quality).

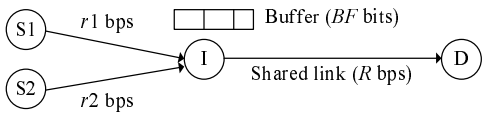

Fig. 1. Topology of experimental network

Six videos (two sets of videos with still, low and high motion) of 10s duration are coded at bit-rates (R/2 bps) using H.264/AVC JM Version 9.1. We form 9 pairs from six videos so we have a balanced representation where each type of video is competing twice with the traffic with all the three types. Each simulation of a pair of videos produces two lossy videos, one for each source video. Therefore, there will be 18 decoded videos in total for each policy. We compare our PQ policy with DT and cMSE in a network condition (R, BF). Videos from the same competing pair among different dropping policies are compared: a policy wins when the VQM score of its resulting video is lower than the other, and a tie occurs when two policies have identical scores.

To show the effectiveness of our approach across different GOP structures, we conducted this procedure with *IPPP*, *IBBP* and *Pyramid* (Fig. 2). Table 3 shows the comparison results for different buffer sizes when bottleneck rates are fixed. When comparing to DT, we can observe large winning margins with IPPP in all cases (the winning ratio, defined as $\#wins/\#losses$, is 5 when aggregated for all buffer sizes). We also do well with IBBP in each buffer size (winning ratio=2). With Pyramid, we win for two of the buffer sizes and tie for the third one. When comparing with cMSE, the proposed method gives a consistently large advantage with Pyramid (6.14) and IBBP (5). However, with IPPP, we lose slightly.

Table 4 demonstrates the performances of different policies for different bottleneck rates while the buffer size is fixed. An important observation is that we perform relatively better in a higher encoding rate (R=1200 kbps) than in lower ones. This is because the data building the model were collected from videos with no obvious coding artifacts. With that said, the performance of our model is quite robust to lower encoding rates. For overall performance with IBBP, the proposed PQ strategy performs very well at all encoding rates, and the winning ratios when aggregated for all encoding rates are about 1.57 over DT, and 3.9 over cMSE. With Pyramid structure, we have good ratio over cMSE (2.17), while the ratio is smaller (1.07) when comparing to DT. And with IPPP, we outperform the DT by 4.40, but we lose slightly against cMSE.

From both tables, we observe that our scheme works well in most of the cases (five out of six GOP-Competitor pairs). In particular, for all the network conditions, our performance is almost always better than the cMSE approach, and always better than DT, a widely implemented dropping method in existing intermediate routers. In addition, PQ on average has a lower (better) VQM score (0.214) than cMSE and DT do (0.247 and 0.244) over different comparisons in Tables 3 and 4.

Conclusion: This paper proposes a generalized packet-loss visibility model for various GOP structures, developed by factors of different information levels and analyzed on multiple data sets. A PQ method to intelligently drop packets in the intermediate router is

Pyramid (R=1200)				**IBBP** (R=1200)				**IPPP** (R=1200)			
vs. DT	Wins	Losses	Ties	vs. DT	Wins	Losses	Ties	vs. DT	Wins	Losses	Ties
BF=200	10	8	0	BF=80	12	6	0	BF=80	18	0	0
BF=400	9	9	0	BF=100	12	6	0	BF=100	13	5	0
BF=600	9	5	4	BF=120	12	6	0	BF=120	14	4	0
vs. cMSE	Wins	Losses	Ties	vs. cMSE	Wins	Losses	Ties	vs. cMSE	Wins	Losses	Ties
BF=200	15	3	0	BF=80	16	2	0	BF=80	9	9	0
BF=400	16	2	0	BF=100	14	4	0	BF=100	7	11	0
BF=600	12	2	4	BF=120	15	3	0	BF=120	6	12	0

Table 3. Proposed PQ compared to DT and cMSE: Fixed bottleneck rate (R kbps) and varied buffer size (BF kbits).

Pyramid (BF=300)				**IBBP** (BF=80)				**IPPP** (BF=80)			
vs. DT	Wins	Losses	Ties	vs. DT	Wins	Losses	Ties	vs. DT	Wins	Losses	Ties
R=800	9	9	0	R=800	11	7	0	R=800	13	5	0
R=1000	9	9	0	R=1000	10	8	0	R=1000	13	5	0
R=1200	10	8	4	R=1200	12	6	0	R=1200	18	0	0
vs. cMSE	Wins	Losses	Ties	vs. cMSE	Wins	Losses	Ties	vs. cMSE	Wins	Losses	Ties
R=800	13	5	0	R=800	14	4	0	R=800	7	11	0
R=1000	13	5	0	R=1000	13	5	0	R=1000	7	11	0
R=1200	11	7	4	R=1200	16	2	0	R=1200	9	9	0

Table 4. Proposed PQ compared to DT and cMSE: Fixed buffer size (BF kbits) and varied bottleneck rate (R kbps).

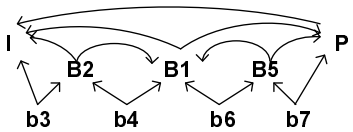

Fig. 2. Pyramid GOP structure; A B-frame in upper case can be used for reference while the ones in lower case can not. The numbers indicate the coding order within the group.

designed based on our visibility model, and the experiment results show that PQ is better than the policy using cumulative MSE as used in the Hint-Track method in most of the cases, and outperforms the widely-implemented Drop-Tail in all cases.

6. REFERENCES

[1] G. W. Cermak. Videoconferencing Service Quality as a function of bandwidth, latency, and packet loss. *T1A1.3/2003-026, Verizon Laboratories*, May 2003.

[2] B. Chen and J. Francis. Multimedia Performance Evaluation. *AT&T Technical Memorandum*, February 2003.

[3] 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.*, 12(12):1071–1083, Dec 2002.

[4] C. J. Hughes et al. Modeling and subjective assessment of cell discard in ATM video. *IEEE Trans. Image Processing*, 2(2):212–222, April 1993.

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

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

[7] A. R. Reibman et al. Quality monitoring of video over a packet network. *IEEE Trans. Multimedia*, 6(2):327–334, Apr 2004.

[8] P. McCullagh and J. A. Nelder. Generalized Linear Models 2^{nd} Edition. *Chapman & Hall*.

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

[10] S. Kanumuri et al. Packet-loss visibility in H.264 videos using a reduced reference method. *IEEE ICIP*, Oct 2006.

[11] A. R. Reibman and D. Poole. Characterizing packet loss impairments in compressed video. *IEEE ICIP*, Sept 2007.

[12] A. R. Reibman et al. Predicting packet-loss visibility using scene characteristics. *Packet Video*, pages 308–317, Sept 2007.

[13] D. Quaglia et al. Adaptive packet classification for constant perceptual quality of service delivery of video streams over time-varying networks. *ICME*, 3:369–72, July 2003.

[14] J. Chakareski and P. Frossard. Rate-Distortion Optimized Bandwidth Adaptation for Distributed Media Delivery. *ICME*, 2005.

[15] J. Chakareski and P. Frossard. Rate-distortion optimized distributed packet scheduling of multiple video streams over shared communication resources. *Multimedia, IEEE Transactions on*, 8:207 – 218, April 2006.

[16] NS Project. http://www.isi.edu/nsnam/ns/.

[17] VQM. http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm.

[18] Y. Sermadevi and A. R. Reibman. Unpublished subjective test results. Sept 2002.

[19] Z. Wang et al. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Im. Proc*, 13, Apr 2004.

[20] A. Hanjalic. Content-based analysis of digital video. *Kluwer Academic Publishers, Boston*, 2004.