Characterizing packet-loss impairments in compressed video

Amy R. Reibman and David Poole
AT&T Labs – Research
{amy,poole}@research.att.com

Abstract— **We examine metrics to predict the visibility of packet losses in MPEG-2 and H.264 compressed video. We use subjective data that has a wide range of parameters, including different error concealment strategies and different compression standards. We evaluate SSIM, MSE, and a Slice-Boundary Mismatch (SBM) metric for their effectiveness at characterizing packet-loss impairments.**

I. Introduction

The growing popularity of transmitting compressed video over the Internet increases the need for quality assessment methods that can accurately characterize how the network is affecting the video quality seen by the end-user. Accurate quality assessment is essential when specifying requirements for, designing, and testing systems that transport video over networks.

Significant progress has been made on methods that assess video quality for applications other than video transmission. Image quality metrics like SSIM [1] and those evaluated in [2] are becoming more accurate at predicting quality of individual frames of video. Full-reference (FR) video quality metrics like the Continuous Video Quality Evaluation (CVQE) metric [3] incorporate the temporal aspects of human perception. Reduced-reference (RR) metrics like [4] extract low-bandwidth information at the sender to be sent reliably to the receiver to estimate resulting video quality. No-Reference (NR) methods such as the blurring metric in [5] only use information available in the bitstream or at the decoder.

Recent work considers the visual quality produced by network impairments. A random-neural-network model is developed in [6] to assess quality given different bandwidth, frame-rate, packet loss rate, and I-block refresh rate. However, quality is evaluated only in an average sense without considering the impact of source content. Average performance across an entire video sequence is also the focus in [7], which uses MSE to assess quality for different compression standards and different concealment techniques. MSE of a packet loss impairment (PLI) is estimated using a NR metric in [8]. Fluidity impairments caused by freeze frames resulting from packet loss are assessed in [9]. NR metrics that directly measure the length and strength of PLI are in [10], [11].

Our recent work focuses on predicting the visibility of PLI for both MPEG-2 [12] and H.264 [13] separately. Our goal in the current work is a quality metric that is targeted specifically for packet loss impairments yet comprehensive enough to be effective for a variety of compression standards, encoding parameters, and decoding strategies. A secondary goal of this paper is to systematically examine the ability of SSIM [1] to predict the visibility of PLI, both with and without assistance from other RR and NR factors. Despite its popularity, we believe this is the first time SSIM has been systematically evaluated on PLI.

Section II describes the subjective test datasets we use, emphasizing the diversity of the parameters underlying the data. Section III describes the attributes of PLI with an eye toward (a) identifying new attributes to measure and (b) emphasizing which features of a PLI are dependent on the compression standard and which are independent of it. Measurements that characterize PLI are described, including a new method to compute SSIM for the initial impairment in an RR setting and an improved NR Slice-Boundary Mismatch (SBM) metric. Section IV compares SSIM and MSE, and presents a set of general models that predict the visibility of PLI for two compression standards and three concealment techniques without prior specification of standard or concealment technique.

II. Subjective datasets

Table I summarizes the three datasets obtained with subjective testing that we use in this study. All three subjective tests are based on the same methodology: a single-stimulus test in which the viewers' task was to indicate by pressing the space-bar when they saw an artifact. Twelve viewers were used to label each packet loss in all three tests. For each of the three tests, packet losses were injected into the video such that every non-overlapping four-second interval contains a single packet loss.

Datasets 1 and 2 use video compressed by MPEG-2 at spatial resolution 720 by 480 with an adaptive GOP structure in which an I-frame is inserted at each scene change. Dataset 3 uses H.264 video at spatial resolution 352 by 240 with a fixed GOP structure. The video content used in each test is highly varied in its motion and spatial texture. The signal attributes are all statistically identical across the three tests. The video content in Dataset 3 is identical to half the video content in Dataset 2, while the content in Dataset 1 is unique.

One important difference among the tests is the decoder concealment strategy. Dataset 3 uses Motion-compensated

	Data 1 [14]		Data 2 [12]	Data 3 [13]
Spatial resolution	720x480		720x480	352x240
Frame rate (fps)	30	24	30	30
Duration (minutes)	7.3	8.9	72	36
Compression standard	MPEG-2		MPEG-2	H.264
GOP structure	I-B-B-P-		I-B-B-P-	I-B-P-
I-frame insertion	scene adaptive		scene adaptive	fixed
GOP length	≤ 13	≤ 15	≤ 13	20
concealment	default[1]		ZMEC	MCEC
Losses	108	107	1080	2160
Losses in B-frames	14%		14%	50%
Full-frame losses	20%		30%	0%
Mean num. viewers who saw each loss	4.56	5.13	3.11	1.32
Null Pred. error	0.14599		0.12236	0.041571
Initial mean sq. pix. error	5.245		3.919	1.708

TABLE I

SUMMARY OF SUBJECTIVE TEST DATASETS

Error Concealment (MCEC) as detailed in [13]. Dataset 2 uses zero-motion error concealment (ZMEC), while Dataset 1 uses a naive error concealment that is typical of software decoders[1]. One common feature of all the decoder error-handling strategies is that the video decoder only processes slices that are completely received. Table I shows both that significantly more viewers saw each loss in Dataset 1 and Dataset 2 than in Dataset 3, and that the initial MSE (IMSE) for those pixels whose packet was lost is very different due to the different concealment strategies.

III. ATTRIBUTES OF PACKET-LOSS IMPAIRMENTS

A. *Description of attributes*

In general, the impairment caused by a packet loss depends on the encoding parameters, how the decoder handles errors, the packetization strategy, and the video content. Let the original uncompressed video frame at time t be $f(t)$, the compressed video frames be $\hat{f}(t)$, and the decoded video frames be $\tilde{f}(t)$. The error is $e(t) = \hat{f}(t) - \tilde{f}(t)$. The PLI then can be characterized by attributes of (a) the error $e(t)$, (b) the decoded signal including error, $\tilde{f}(t)$, and (c) the encoded signal (without impairment) at the location of the impairment, $\hat{f}(t)$.

The error caused by the impairment, $e(t)$, can be characterized by its support and its amplitude. The support is characterized by size, spatial pattern, duration, and location. The error $e(t)$ may have somewhat different characteristics depending on the compression standard. For example, H.264 allows Flexible Macroblock Ordering (FMO), which may alter the spatial pattern of the error. Using long-term prediction in H.264 can improve error attenuation [15], but the initial error at the time of the loss depends more heavily on the underlying video content and the decoder concealment than on the compression standard itself. Further, the size, location and duration of the error are not influenced by the choice of compression *standard* (although they may depend on the encoding *parameters*).

The decoded signal, $\tilde{f}(t)$, at a PLI has several attributes that are likely to affect packet-loss visibility. A lost frame is likely to introduce temporal edges and a lost slice to introduce both temporal and horizontal edges into the decoded signal. Vertical edges may also be introduced with FMO, or when the impairment propagates into subsequent frames. Moving vertical edges that are continuous in the encoded signal may also become disjointed in the decoded signal due to the impairment. All of these edge artifacts are likely to increase the visibility of the impairment.

Attributes of the encoded signal without PLI, $\hat{f}(t)$ at the location of the PLI can also affect visibility. Texture, luminance, and motion masking may each reduce visibility of the PLI. Motion tracking may enhance visibility of PLI in smoothly moving regions, yet local signal variance and motion variability may hide the PLI. Clearly, signal attributes do not depend on the compression standard.

B. *Measuring attributes*

To obtain an accurate quality metric for PLI, we must measure the effects described above. In this section, we describe measurements of these attributes that can be extracted with a FR, a RR, or NR video quality metric and discuss limitations of these measurements.

1) Full-Reference measurements: MSE, a FR metric, measures the error, $e(t)$, directly. It characterizes the error amplitude in part, but cannot quantify the spatio-temporal frequency characteristics of the error. MSE only indirectly measures attributes like error size and duration, but cannot capture any information about error location or pattern. Finally, MSE is clearly incapable of characterizing anything about the decoded pixels $\tilde{f}(t)$ or the underlying signal $\hat{f}(t)$.

SSIM, also a FR-metric, characterizes the error, $e(t)$, to the same degree and accuracy as MSE; however, it also incorporates some information about the signal at the location of the impairment. It captures statistical information about $\tilde{f}(t)$, but does not directly measure the decoded impairment attributes listed above.

For both MSE and SSIM, the averaging (or pooling) interval is an important consideration. In this paper, we consider MSE-1 and SSIM-1, which are averages across all pixels in a one-second interval that contains the PLI. We also consider other pooling below.

2) Reduced-Reference measurements: For RR measurements, we envision a video encoder or video server reliably providing per-macroblock (MB) information about $e(t), \hat{f}(t)$ for video quality assessment and having the video quality metric compute similar information about $\tilde{f}(t)$. The measurements of $e(t)$ assume knowledge of the decoder concealment strategy and may have reduced accuracy when using MCEC since its estimate of the missing motion depends on the received data.

[1] To improve speed, many software decoders merely swap pointers between forward and backward reference frames. In this "default concealment", missing macroblocks are never overwritten. So a reference frame is "concealed" using data from *two* reference frames ago, while B-frames are "concealed" using data from the most recent *B-frame*.

Max-IMSE, defined as the maximum per-MB MSE over all MBs in the initial impairment, was shown to be a useful measure in [13]. Here, we also consider Min-ISSIM, the minimum per-MB SSIM over all MBs in the initial impairment. It can be shown that a per-MB *initial* SSIM can be computed in an RR framework using

$$\mathrm{SSIM} = \frac{(2\hat{\mu}\tilde{\mu} + C_1)}{(\hat{\mu}^2 + \tilde{\mu}^2 + C_1)} \left[1 + \frac{(\hat{\mu} - \tilde{\mu})^2 - \sigma_e^2}{(\hat{\sigma}^2 + \tilde{\sigma}^2 + C_2)} \right]$$

where $\hat{\mu}$,$\tilde{\mu}$, $\hat{\sigma}^2$, $\tilde{\sigma}^2$ are the local means and variances of $\hat{f}(t)$ and $\tilde{f}(t)$ respectively, σ_e^2 is the MSE, and C_1 and C_2 are constants described in [1]. We use a MB-sized uniform window to compute local means and variances instead of the 11x11 Gaussian window proposed in [1] to reduce storage and transmission requirements for an RR metric. We also consider IMSE and ISSIM to quantify the initial error over the frame with the initial PLI.

We also consider the following RR signal descriptors: MotMean, MotVar, and ResidEng (the residual energy after motion compensation) as considered in [12], as well as SigMean and SigVar. Because our goal is a standards-independent PLI metric, we measure these directly from the underlying signal, independent of the compression algorithm, using 16x16 motion blocks. We store these using one value each for a complete row of MBs.

3) No-Reference measurements: NR measurements may be based on pixels (NR-P), the lossy bitstream (NR-B), or both (NR-BP). Using the lossy bitstream, we can exactly measure error size, pattern, location, and duration. However, NR-P methods can only exactly measure information about $\tilde{f}(t)$. Both can estimate attributes of the signal at the location of the impairment using information from neighboring unimpaired frames.

As in our past work, we consider here the NR-B measurements of spatial extent, temporal duration, and location. We also present a NR-BP Slice-Boundary Mismatch (SBM) metric based on the NR-P metric presented in [11].

The NR-P metric in [11] applies a PLI detection and estimation stage to the decoded pixels to measure combined impairment length and strength. Its detection stage is based on the assumption that it is unlikely for the signal $\hat{f}(t)$ to have edges at MB boundaries. Unfortunately, in the over 200,000 frames in our combined subjective tests, the detection process in [11] detects PLI in one-third of them, even though less than 6% have PLI. Further, the detection process misses 60% of the frames with PLI. It is unable to detect full-frame impairments and often does not detect the PLI that propagate into subsequent frames because the impairment is no longer aligned with MB boundaries.

Therefore, we replace the NR-P detection process in [11] with a NR-B detector that uses information from the received bitstream to exactly pinpoint impairment location. Our NR-P estimation process differs from that in [11] in minor details only. We define the SBM metric based on the impact of the impairment on slice boundaries, and apply the metric only on the boundaries between slices that contain non-zero errors.

$$\mathrm{SBM} = \frac{1}{r-1} \sum_{i=1}^{r-1} \mathrm{SBM}(i) * I(i)$$

where there are r rows of MBs, $\mathrm{SBM}(i)$ is the Slice-Boundary Mismatch along the i-th MB boundary, and $I(i)$ indicates we detected an impairment in one of the slices on either side of the boundary. Let $S_u(i)$ and $S_d(i)$ be the sum of absolute row difference up above and down below the MB row-boundary i, respectively, and let $S_m(i)$ be their mean. Then,

$$\mathrm{SBM}(i) = \max\{(S_b(i) - S_m(i))/S_m(i), 0\}$$

when both $S_b(i) > T$ and $S_m(i) > T$, where $S_b(i)$ is the summed absolute row difference across MB row-boundary i, and T is a noise threshold set to 10 here. In our studies below, we consider Max-SBM, the maximum value across all impaired frames.

IV. Results

We use logistic regression [16] to predict packet loss visibility, as in [12], [13]. Logistic regression is a special case of a regression using a Generalized Linear Model (GLM) where the link function is the logit function. Our goal is to estimate p_i, the fraction of viewers who saw error i. We fit our models using a fraction of the data (training set) and evaluate it using the remaining samples (test set). Performance is the prediction error averaged using four-fold cross-validation (CV) and four initial random seeds. Each training set has an equal number of samples from each Dataset, and each Dataset provides equal weight in the final prediction error. To improve performance, we fit using $\log(1 - SSIM)$ and $\log(MSE)$ rather than each variable directly.

All model performance in this section is illustrated in both Table II and Figure 1. The first shows the CV prediction error averaged across all 3 Datasets for all the models considered, while the second shows the CV prediction error in each Dataset for a subset of the models considered. We now describe our models.

We begin by examining the relative importance of the various error and signal+error descriptors: SSIM, MSE, and SBM. Table II(a) shows the CV prediction error for each individual factor. MSE outperforms SSIM for all pooling strategies, with the Min/Max pooling strategy performing best. Prediction error for SBM alone is quite poor although it is better than no model (the null error).

Next, we examine the impact of adding NR error factors to the single-factor fits examined above. We compare two sets of models: those using only one of SSIM, MSE, and SBM, and those using one of those measures along with two NR error descriptors: initial spatial extent and temporal duration. Duration is measured in seconds to account for the disparate GOP structures and the presence of both video and film in the datasets. To account for the disparate spatial resolutions, spatial extent is measured relative to

Primary factor	Primary factor alone	Primary factor + duration and spatial extent
SSIM-1	0.062449	0.051103
ISSIM	0.065483	0.054315
MinISSIM	0.058437	0.056216
MSE-1	0.057580	0.050207
IMSE	0.060313	0.052822
MaxIMSE	0.057014	0.055604
MaxSBM	0.084992	0.077192

(a)

Combined model 1: 1-sec. pooling	0.03924
Combined model 2: Initial per-frame error	0.04155
Combined model 3: MB-based error	0.04652
Null-error	0.10331

(b)

TABLE II

AVERAGE CV PREDICTION ERROR ACROSS ALL DATASETS FOR VARIOUS MODELS. (A) INDIVIDUAL FACTORS WITH OR WITHOUT DURATION AND SPATIAL EXTENT. (B) MODELS INCORPORATING ERROR, SIGNAL AND SIGNAL+ERROR FACTORS.

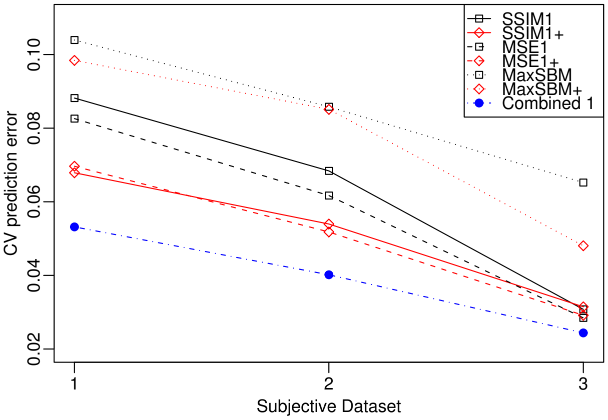

Fig. 1. Prediction error within each Dataset using one-second pooling. SSIM: solid lines, MSE: dashed lines, SBM: dotted lines.

the full frame size. The latter set of models are designated by adding a "+" to the name in the figure.

Each model improvement is statistically significant when the NR error factors are added, although the largest improvement is seen for SSIM-1 and ISSIM. Figure 1 shows that SSIM1+ performs nearly identically to MSE1+, although slightly better on Dataset 1 and slightly worse on Datasets 2 and 3. Adding the NR error factors helps Max-SBM+ for Datasets 1 and 3, but it still performs poorly.

Next, we consider logistic regression models that use all error characteristics (Duration, InitialSpatialExtent, MSE, SSIM, and SBM) as well as signal characteristics (SigMean, SigVar, MotMean, MotVar, ResidEng). We tentatively add each new factor individually and decide to include it in our model only if the average CV prediction error decreases. We also explore some nonlinear mappings of each factor to see if improved fitting performance is possible. We present a set of three models according to the type of pooling used for the error $e(t)$: one-second pooling across all frames affected by the PLI, per-frame averaging across the initial frame affected by the PLI, and maximum-over-macroblock pooling across the initial frame affected by the PLI. Table III summarizes the final factors in each model. Results in both Figure 1 and Table II(b) demonstrate that MSE, SSIM, and SBM all combine to obtain substantially better performance across all Datasets.

Model 1: 1-sec pooling	Model 2: Initial error	Model 3: Max-MB error
$\log(\text{MSE1})$	$\log(\text{IMSE})$	$\log(\text{MaxIMSE})$
$\log(1 - \text{SSIM1})$	$\log(1 - \text{ISSIM})$	$\log(\text{MinISSIM})$
$\log(\text{MaxSBM})$	—	—
Duration	(Duration$<$ 0.05)	(Duration$<$ 0.05)
InitialSpatialExtent	InitialSpatialExtent	—
$\lvert\text{SigMean} - 128\rvert$	$\lvert\text{SigMean} - 128\rvert$	$\lvert\text{SigMean} - 128\rvert$
$\log(\text{SigVar})$	$\log(\text{SigVar})$	$\log(\text{SigVar})$
$\text{MotMean} > 1/\sqrt{2}$	—	$\text{MotMean} > 1/\sqrt{2}$
MotVar	—	—
$\log(\text{ResidEng})$	$\log(\text{ResidEng})$	$\log(\text{ResidEng})$

TABLE III

FACTORS IN EACH COMBINED MODEL

V. CONCLUSIONS

Methods to characterize the PLI in the decoded signal are still in their infancy. On its own, Max-SBM is unable to accurately predict visibility of PLI, but incorporating it into a comprehensive PLI detection model improves prediction accuracy. Combining SSIM and MSE into the same model also statistically improves performance.

REFERENCES

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

[2] H. R. Sheikh et al., "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms", *IEEE Trans. Image Proc.*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.

[3] M. A. Masry, S. S. Hemami, "A Metric for Continuous Quality Evaluation of Compressed Video With Severe Distortions," *Signal Proc.: Image Comm.*, vol. 19, no. 2, Feb. 2004.

[4] S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system", *First Int'l Workshop on Video Proc. and Quality Metrics*, Jan 2005.

[5] P. Marziliano et al, "Perceptual blur and ringing metrics: Application to JPEG2000", *Sig. Proc.: Image Comm.*, pp. 163–172, Feb. 2004.

[6] 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.

[7] S. Tao et al., "RealTime Monitoring of Video Quality in IP Networks", *Proceedings NOSSDAV'05*, pp. 129–134, June 2005.

[8] A. R. Reibman et al. "Quality monitoring of video over a packet network", *IEEE Trans. Multimedia*, vol. 6, pp. 327-334, April 2004.

[9] R. R. Pastrana-Vidal and J.-C. Gicquel, "Automatic quality assessment of video fluidity impairments using a no-reference metric", *Int'l Workshop on Video Proc. and Quality Metrics*, Jan 2006.

[10] R. V. Babu et al., "No-Reference metrics for video streaming applications", *International Workshop on Packet Video*, Dec 2004.

[11] H. Rui et al., "Evaluation of packet loss impairment on streaming video", *J. of Zhejiang University SCIENCE*, vol. 7, April 2006.

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

[13] S. Kanumuri et al., "Packet-loss visibility in H.264 videos using a reduced reference method", *IEEE Int. Conf. Image Proc.*, Oct. 2006.

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

[15] M. Budagavi and J. D. Gibson, "Multiframe video coding for improved performance over wireless channels", *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 252–265, February 2001.

[16] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley-Interscience.