Predicting H.264 Packet Loss Visibility using a Generalized Linear Model

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

Sitaraman G. Subramanian
Univ. Calif. at San Diego
sganapat@code.ucsd.edu

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

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

Abstract— **We consider modeling the visibility of individual and multiple packet losses in H.264 videos. We propose a model for predicting the visibility of multiple packet losses and demonstrate its performance on dual losses (two nearby packet losses). We extract the factors affecting visibility using a reduced-reference method. We predict the probability that a loss is visible using a generalized linear model. We achieve MSE values (between actual and predicted probabilities) of 0.0253 and 0.0398 for individual and dual losses respectively. We also examine the effect of various factors on visibility.**

Index Terms: Video codecs, Quality Control

I. Introduction

Compressed video streams transmitted over heterogeneous networks experience visual quality impairments due to various factors such as delay, jitter, packet loss, drift (due to scalability) and compression artifacts. Being able to quantify the perceptual quality degradation due to these factors is important for a network quality monitor. Objective quality measures such as PSNR do not correlate well with subjective results. The current work concentrates on modeling the quality degradation in H.264 video due to packet losses by predicting the packet loss visibility.

Predicting the visibility of a packet loss is useful for several reasons. Packets which are perceptually important can be given more error protection (unequal error protection). If one assigns perceptual importance levels to each of the packets at the encoder, the packets with lower importance level can be discarded, if the buffer in a network node overflows. Thirdly, packet loss visibility can be very useful for accurate, real-time network quality monitoring.

In our previous work [1], we looked at the problem of predicting the visibility of individual packet losses in MPEG-2 bitstreams. However, video transmission over internet or wireless links is typically characterized by bursty losses. Stuhlmuller et al. [2] analyzed the distortion (MSE) of isolated packet losses. They also used a linear additive model to quantify the distortion of multiple packet losses. In [3], the authors compare bursty losses with isolated losses of equal combined length. They conclude that: (a) the loss pattern has significant impact on distortion, (b) bursty loss produces larger distortion than an equal number of isolated losses. Chakareski et al. [4] proposed a scheme to predict the distortion incurred by bursty losses.

In this paper, we extend our previous work [1] in two ways. First, we model the visibility of isolated packet losses in *H.264* instead of MPEG-2 bitstreams. We use motion-compensated error concealment (MCEC) to conceal the losses, instead of the zero error concealment (ZEC). Second, we model the visibility of *multiple* packet losses for H.264. We define a multiple loss as a set of N individual losses occurring in close temporal proximity. We introduce a model framework to predict the visibility of a multiple packet loss and we examine the performance of this approach when $N = 2$ (dual loss). We compare the importance of factors between the isolated and dual loss cases.

Section II describes the design of the subjective experiment. Section III discusses the factors we use to predict the visibility of a loss. Section IV describes our modeling approach, while Section V provides results.

II. Subjective Tests

We conducted subjective tests to obtain ground truth on the visibility of packet losses. The viewers' task is to indicate when they see an artifact or abnormality. Our tests are single stimulus tests, so viewers are only shown videos with packet losses, not original videos. We assume a packet loss entails the loss of a single slice, namely a horizontal row of macroblocks (MBs)[1]. The initial error induced by a packet loss depends on the decoder's error concealment strategy. We use motion-compensated error concealment (MCEC). The initial error incurred with this method is less than that of the zero-error concealment method used in our previous work [1].

The video sequences in the subjective test are muted travel documentaries at SIF resolution (352 × 240) and 30 fps. They are encoded and decoded using the extended profile of H.264 JM Version 9.1 Codec. Our encoding structure is I B P B P B...P B with a GOP size of 20 frames. For P frames, we use two reference frames for motion compensation: one long-term and one short-term. The long-term reference frame is always the I frame of the current GOP. We follow the usual convention for the

[1] We do not consider the Flexible Macroblock Ordering (FMO) available in H.264.

1-4244-0481-9/06/$20.00 ©2006 IEEE

short-term reference frame to be the previous P frame. B frames use the future P frame and either reference frame for bidirectional prediction. We use fine quantization (28) without rate control, so that the only artifacts in the lossy videos are due to packet loss. The decoder conceals the lost slices using MCEC where the concealment motion vector is estimated as the median of the motion vectors of surrounding blocks.

We are interested in the visibility of individual and multiple packet losses. In practice, packet losses tend to occur together. Different losses within a multiple packet loss interact with each other; the overall effect is not the sum of individual effects. This interaction can be either physical or perceptual. Physical interaction is caused by inter-frame prediction. Perceptual interaction is caused by the close proximity (spatial or temporal) of the losses. Here, we study the overall effect of two packet losses (dual loss) occurring together. A dual loss is characterized by the spatial separation D between the two losses in MB units (i.e., $D = 1$ implies the two packet losses affect adjacent slices) and by the temporal separation T between the two losses in number of frames. In our case, D varies from 0 to 14 (15 slices in each frame). T varies from 0 to 5 frames (maximum separation of 1/6 of a second).

We choose six videos of 6 minutes each, divided into 4-second intervals called slots, producing 90 slots per video and 540 slots in total. A loss (individual/dual) is introduced in the first three seconds of each slot. The last second is reserved to create a guard interval, which prevents interaction between losses across slots and provides the viewer time to respond to the current loss before the next loss occurs.

We design 4 individual losses for each slot and use all 6 combinations of individual loss pairs to get 6 dual losses. Within each time slot, either exactly one of these 4 individual losses will appear, or exactly two of them will appear (dual loss). The individual losses are designed such that the set of dual losses are approximately distributed uniformly over D and T. Since we have 10 different losses (4 individual and 6 dual) that can be introduced in a slot, 10 different lossy versions are created from each source video. Each lossy video has both individual and dual losses, but in different slots. In total, we introduced 2160 individual losses and 3240 dual losses.

During the subjective test, a viewer is shown only one set of 6 lossy videos. Each set of lossy videos was evaluated by 12 viewers, for a total of 120 viewers. To help viewers understand their task, we show them a 1-minute pilot training video before the actual test. Viewers are told that they will watch videos which are affected by packet losses. Whenever they see a visible artifact or a glitch, they should press the space bar. After the subjective test, we generate a table of the viewers' boolean responses, corresponding to whether they saw a loss or not. This defines the ground truth for the probability of visibility for each loss.

III. Factors affecting Visibility

In this paper, we use a Reduced-Reference (RR) method for predicting the visibility of packet losses. A RR method has access to the decoder's reconstructed video (with losses) and factors extracted from the encoded video. We classify the factors that determine packet-loss visibility into Content-Independent and Content-Dependent Factors.

Content-Independent Factors depend only on the location of the loss, not on its actual content. We consider two content-independent factors. HGT [1] is defined as the height of the lost slice in a given frame, where slices are numbered from top to bottom. $FRAMETYPE$ is the type of frame (B/P/I) affected by the packet loss and is treated as a categorical factor. $FRAMETYPE\_ML$ is the counterpart of $FRAMETYPE$ for the dual (multiple) loss case. Between the two frames affected, $FRAMETYPE\_ML$ represents the type that belongs to a higher category. Category I is higher than category P, which is higher than category B.

Content-dependent Factors, on the other hand, depend on the actual video content at the location of the loss, such as Motion, Contrast, etc. We use the following:

1) **Initial Mean Squared Error (IMSE)** is the MSE between the error-free reconstructed MB and the lossy concealed MB. Factors $AVGIMSE$ and $MAXIMSE$ are the average and maximum IMSE of all MBs in a given slice.

2) **Residual Energy** is the energy (sum of squares of all DCT coefficients) of the residual after motion compensation. If a slice is lost, then even if MCEC does a perfect job of estimating the lost motion vectors, the resultant slice still differs from the original. The residual energy, calculated on a MB basis, is one way to assess the magnitude of this difference. $AVGRSENGY$ and $MAXRSENGY$ are the average and the maximum of the residual energy values of all the MBs in a given slice.

3) **Motion-Related Factors:** For computing motion-related factors, we first linearly scale the motion vectors in each partition of a MB so they represent the motion between two display frames. Then we assign each MB a single motion vector which is a weighted average of the motion vectors in all the MB partitions.

$MEANMAG$ and $MAXMAG$ are the mean and maximum magnitudes of all motion vectors of the MBs in a given slice. For computing phase-related factors, only MBs with non-zero motion are used. Further, we require at least half of the MBs to have non-zero motion. If not, we consider that phase information is undefined and set PH_UDEF (a boolean factor). If PH_UDEF is not set, $MEANPHASE$ and $MAXPHASE$ are the mean and maximum of all the defined phases of the MBs in the slice. In the dual loss case, we have two boolean factors PH_UDEF1 and PH_UDEF2. PH_UDEF1 is set if at least one of the losses has an undefined phase and PH_UDEF2 is set if both losses have undefined phase.

Since we cannot assign a value when the phase is undefined, we use $MEANPHASE\_V$ and $MAXPHASE\_V$ that are variants of $MEANPHASE$ and $MAXPHASE$. These variants take on the original values incremented by 1 when the phase is defined, and the value 0 when it is undefined.

$INTRASLICE$ (a boolean factor) is set when the lost slice is coded as an intra slice. In the case of dual losses, we have two boolean factors $INTRASLICE1$ and $INTRASLICE2$. $INTRASLICE1$ is set if at least one of the lost slices is coded as an intra slice and $INTRASLICE2$ is set if both the lost slices are coded as intra slices. Unlike MPEG-2, H.264 has motion vectors with variable block sizes. The number of macroblock partitions in H.264 can range from one for the coarsest partition to sixteen for the finest partition. $AVGINTERPARTS$ and $MAXINTERPARTS$ are the average and maximum number of Inter macroblock partitions in a slice.

We also explored various spatial clutter and contrast-based factors which statistically did not turn out to be useful.

For individual packet losses, we use the K factors described above to model visibility. For multiple losses, our goal is to develop a generic model for visibility irrespective of the number of individual packet losses (N) in the multiple loss. However, for each of the K factors in the individual loss case, we now have N values, one for each packet loss. Hence, we have a total of NK available factors for multiple losses. However, a generic model should select factors irrespective of N. Therefore, we derive $2K + 5$ factors representing a multiple loss as follows. We use the maximum and minimum of the N values from each of the K factors to form $2K$ new factors. They are named by attaching "MAX_" or "MIN_" as prefix to the factor name. We also use the maximum and minimum of D_i and T_i, the spatial and temporal separation between each pair of packet losses. The final factor is the number of packet losses in the multiple loss, N. In this paper, we demonstrate the effectiveness of this framework in predicting visibility of dual losses ($N = 2$). With only one pair, we use D and T directly, and set $N = 2$.

IV. Modeling Approaches

We model the probability of visibility using logistic regression, a type of generalized linear model (GLM) [6] whose link function is set to be the logit function. The simplest model (Null model) has only one parameter: the constant γ. At the other extreme, it is possible to have a model (Full model) with as many factors as there are observations. The goodness of fit for a GLM can be characterized by its deviance, defined in [6]. By definition, the deviance for the Full model is zero and the deviance for all other models is positive. A smaller deviance means a better model fit. Deviance is also useful in determining the significance of different factors.

All the factors described in section III may not be statistically useful. To identify the factors that are important and to build a good model, we follow the 6-stage approach described in [7]. The first stage involves a univariable analysis of each factor to identify factors that show no association with visibility and they are not considered further for multivariable analysis. The second stage involves building a multivariable model using a stepwise approach for adding factors, one at a time. We start with the Null model and at every step, we add the factor that causes the maximum decrease in deviance per degree of freedom. This gives us a list of models with increasing numbers of factors. In the third stage, we select the first model in the list whose cross-validated MSE (between actual and predicted probabilities) increases when the chosen factor is added. The significance of each factor in the selected model is verified and insignificant factors are dropped from the model. This model is called the preliminary effects model. In the fourth stage, we check for the correct parametric representation for each factor in the model. For example, a factor F might be better represented by F^2 instead of F. The model at this stage is called the main effects model. In the fifth stage, we look for any interaction factors that make intuitive sense and improve the prediction capability. Interaction factors are created as the product of pairs of main effect factors. This gives us the preliminary final model. In the sixth stage, we verify the importance of each factor in the preliminary final model and drop insignificant factors. This marks the completion of our model building process. The model at this stage is called the final model.

V. Results

We obtained our final models using the model building process described in section IV. $MAXIMSE$ is found to be better represented by $MAXIMSE\_S$ (scaling with a power of 1/4). We found the interaction between $MAXRSENGY$ and $INTRASLICE$ to be useful ($INTER\_IL$ - individual loss, $INTER\_ML$ - multiple loss).

Our final model has 8 factors for the individual loss case. Its residual deviance is 5237.7 whereas the null deviance is 8597.5 and the MSE obtained during cross-validation is 0.0253. Similarly, in the dual loss case, our final model has 9 factors. Its residual deviance is 10402.2 whereas the null deviance is 17802.3 and the MSE obtained during cross-validation is 0.0398.

The factors in the final models and their coefficients are listed in tables I and II for individual and dual loss cases. The values of the coefficients do not necessarily convey the importance of corresponding factors because these factors have different variances and ranges. However, the sign of the coefficients is important and informs whether a packet loss is more visible with a high or low value for a factor. Most of the factors in the final models for individual losses and dual

Factor	Coefficient
Constant γ	-2.750e+00
$MAXIMSE\_S$	5.141e-01
PH_UDEF	-1.419e+00
$MAXPHASE\_V$	-5.566e-01
$MAXRSENGY$	-4.481e-04
$FRAMETYPE - P$	8.333e-01
$FRAMETYPE - I$	9.778e-01
$AVGINTERPARTS$	-3.298e-01
$VARMOTX$	-1.861e-03
$INTER\_IL$	7.346e-04

TABLE I

FACTORS AND THEIR COEFFICIENTS IN THE FINAL MODEL (INDIVIDUAL LOSSES)

Factor	Coefficient
Constant γ	-1.769e+00
$MAX\_MAXIMSE\_S$	5.139e-01
PH_UDEF2	-1.813e+00
$MAX\_AVGINTERPARTS$	-3.273e-01
$MAX\_MAXPHASE\_V$	-6.127e-01
$MAX\_MAXRSENGY$	-5.315e-04
PH_UDEF1	-1.767e-01
$FRAMETYPE\_ML - P$	8.862e-01
$FRAMETYPE\_ML - I$	1.249e+00
MAX_HGT	-5.114e-02
$INTER\_ML$	4.700e-04

TABLE II

FACTORS AND THEIR COEFFICIENTS IN THE FINAL MODEL (DUAL LOSSES)

losses have one-to-one correspondence and corresponding coefficients have the same sign.

The importance of a factor in a model can be evaluated by the increase in the deviance that results when that factor is removed from that model. With this analysis, $MAXIMSE\_S$ is the most significant factor in predicting visibility, followed by $FRAMETYPE$ and $AVGINTERPARTS$. We observe the following about the effect of factors on visibility:

1) Factor $MAXIMSE\_S$ is directly proportional to visibility. This makes intuitive sense. If the initial MSE of a loss is high, one would expect the loss to be more visible. Visibility increases, as expected, in the order B, P and I for $FRAMETYPE$.

2) Factors $MAXRSENGY$, $AVGINTERPARTS$, $VARMOTX$ and MAX_HGT are inversely related to visibility. A high value for residual energy can occur when the signal has a lot of high frequency content and the motion is inconsistent (for example, a market crowded with people). In such a case, visibility is reduced due to masking effects. When $AVGINTERPARTS$ is large, the MB must be subdivided to achieve good motion compensation. Hence, the underlying motion is complex, generating spatial and temporal masking that makes the loss less visible. Similarly, a large $VARMOTX$ means that the motion is highly variable across the slice. The negative coefficient for MAX_HGT indicates that viewers' sensitivity to losses goes down as we move from the top to the bottom of the frame. When PH_UDEF is set (i.e, when the majority of the motion vectors in the slice are (0,0)), visibility decreases since concealment works well.

3) Horizontal motion causes losses to be more visible than vertical motion does. $MAXPHASE\_V$ is very significant and has a negative coefficient. One explanation for this arises because a slice is a horizontal structure, and a packet loss causes the loss of a slice. When there is horizontal motion, vertical edges longer than a MB cause discontinuous edges when concealed, and horizontal edges cause no new artifacts. On the other hand, when there is vertical motion, vertical edges do not cause new artifacts, and horizontal edges either appear twice or disappear completely depending on whether the concealing slice contains the horizontal edge or not. However, they do not cause discontinuous edges.

4) The effect of factors $MAXIMSE\_S$, $FRAMETYPE$, $MAXRSENGY$, $VARMOTX$, MAX_HGT is consistent with our earlier findings based on MPEG-2 videos [1]. However, in our earlier work which used zero-motion error concealment, the magnitude of the underlying motion was a highly significant factor for predicting visibility. Now that we use motion-compensated error concealment, motion is no longer a significant factor.

Conclusion: We considered the problem of modeling the visibility of individual and multiple packet losses in H.264 bitstreams, and explored the importance of new factors in predicting visibility. The factor $AVGINTERPARTS$ based on variable block size in H.264 turned out to be significant. Unlike our previous work [1] using zero error concealment, the amount of motion is no longer a significant factor in predicting packet loss visibility. Factors such as $MAXIMSE\_S$, $FRAMETYPE$, $MAXRSENGY$, $VARMOTX$ and MAX_HGT continue to be significant consistent with our earlier findings.

REFERENCES

[1] S. Kanumuri et. al., "Modeling Packet-Loss Visibility in MPEG-2 Video", *IEEE Trans. Multimedia*, vol. 8, pp. 341-355, April 2006.

[2] K. Stuhlmuller et. al., "Analysis of video transmission over lossy channels", *IEEE J. on Sel. Areas in Comm.*, vol. 18, no. 6, pp. 1012-32, June 2000.

[3] Y. J. Liang et. al., "Analysis of packet loss for compressed video: does burst-length matter?", *ICASSP*, vol. 5, pp. 684-687, April 2003.

[4] J. Chakareski et. al., "Distortion chains for predicting the video distortion for general packet loss patterns", *ICASSP*, vol. 5, pp. 1001-1004, May 2004.

[5] Y. F. Ma et. al., "A generic framework of user attention model and its application in video summarization", *Multimedia, IEEE Transactions on*, vol. 7, no. 5, pp. 907-919, Oct 2005.

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

[7] D. W. Hosmer and S. Lemeshow, "Applied Logistic Regression", 2^{nd} Edition, Wiley-Interscience.