CLASSIFICATION OF MPEG-2 TRANSPORT STREAM PACKET LOSS VISIBILITY

Jihyun Shin and Pamela C. Cosman

Dept. of Electrical and Computer Engineering, University of California, San Diego

ABSTRACT

We classify the visibility of TS (Transport Stream) packet losses for SDTV and HDTV MPEG-2 compressed video streams. TS packet losses can cause various temporal and spatial losses. The visual effect of a TS packet loss depends on many factors, in particular whether the loss causes a whole frame loss or partial frame loss. We develop models for predicting loss visibility for both SDTV and HDTV resolutions for frame loss and partial frame loss cases. We compare the dominant predictive factors and the results for the two resolutions. We achieve more than 85% classification accuracy.

Index Terms— TS packet, support vector machine, classification, packet loss visibility

1. INTRODUCTION

Most prior work on the visual effects of packet loss studied average perceptual scores of videos subject to average packet loss rates. In contrast, our prior work focussed on predicting the visibility of individual packet losses [1, 2, 3], and provided loss visibility models for MPEG-2 and H.264 compressed videos. These models assumed that one horizontal slice of macroblocks was put in one packet. However MPEG-2 video streams are often packetized into Transport Stream (TS) packets of 188 bytes. The goal of this work is to predict the loss visibility for these common fixed-size packets. This would be useful for visibility-based unequal error protection, or for packet dropping strategies during congestion.

The visual effect of a TS packet loss varies considerably depending on many factors. We ignore the extreme case of sequence loss from losing a portion of the sequence header. We focus on frame loss (FL) when the packet contains a frame header, and partial frame loss (PFL), when it does not. TS packets when lost can cause various temporal errors. If a reference frame is lost, the error can propagate. We conducted a subjective experiment with human observers to determine the visibility of individual TS packets with different characteristics. Experiments were conducted on two different frame sizes: SDTV (720 × 480) and HDTV (1920 × 1088). Depending on the fraction of viewers who saw each loss, the packet is classified as leading to a visible or invisible loss. We developed computable classifiers to predict this class.

In this paper, Section 2 describes the design and ground truth results of our subjective test. Section 3 describes the factors used for classification as well as the support vector machine classifier. Section 4 presents results and conclusions.

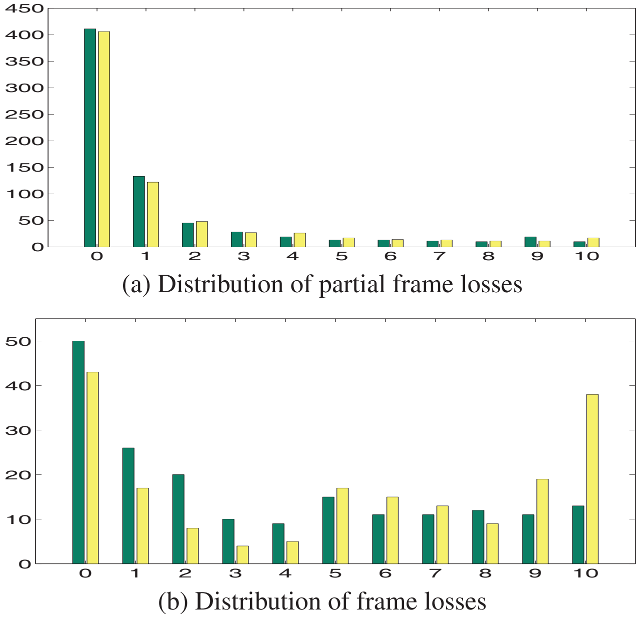
2. SUBJECTIVE EXPERIMENT

For the subjective tests, we conducted a single stimulus test, where only the lossy video being evaluated is shown (no original video). The observer was situated approximately six picture heights from the screen for the SD videos and three picture heights for the HD videos, conforming to ITU and DSL Forum Recommendations [4, 5]. A Group of Pictures (GOP) consists of 15 frames and the frame rate was 30 frames per second. Within the GOP, we refer to the P frames and B frames as follows:

$$\cdots IB_1B_2P_1B_1B_2P_2B_1B_2P_3B_1B_2P_4B_1B_2I\cdots$$

DVD-quality MPEG-2 video was used without audio. The nine raw video sources, with widely varying motions and textures, are in HD format; SD versions are obtained by downscaling with bicubic interpolation. The videos were concatenated to form a 20-minute sequence. Each subject watches one set: a lossy 20-minute SD video and the corresponding HD version. The experiment typically takes one hour, which includes an introductory session and a break. When viewers see a glitch, they press the space bar. Three different lossy video sets were generated by dropping TS packets randomly with pre-determined numbers for the loss types from the original video set. The videos are each observed by ten people. Table 1 shows the pre-determined numbers for the loss types. To create the lossy videos, one TS packet was dropped in every 4 second interval. The packet was dropped during the first 3 seconds of the interval and the last 1 second was used as a guard interval. This allowed the observer time to react to each individual loss, and also allowed the decoder to terminate any possible error propagation from the loss, thereby isolating each individual TS packet loss. For each resolution, data is gathered for a total of 900 TS packet losses over 60 minutes of MPEG-2 video. Of the 900 losses, 188 affected an entire frame and 712 losses lead to a partial loss.

This work was supported by Futurewei Technologies, Inc. and by the Center for Wireless Communications at UCSD.



(a) Distribution of partial frame losses

(b) Distribution of frame losses

Fig. 1. Distributions for SDTV (left bars) and HDTV (right).

The lossy MPEG-2 compressed video streams were decoded by ffmpeg [6]. This decoder, on confronted with a frame header loss, will produce errors in the previous frame as well, because bits of the currently damaged frame are overwritten on the previous frame (frame header loss prevents the decoder from realizing that data from a new frame is being decoded). Because of the encoding frame order, header loss of an I-frame, P-frame, or B$_1$-frame damages a reference frame, and errors propagate until a new I-frame. For the PFL case, the ffmpeg decoder will conceal the loss using a weighted motion-compensated error concealment strategy.

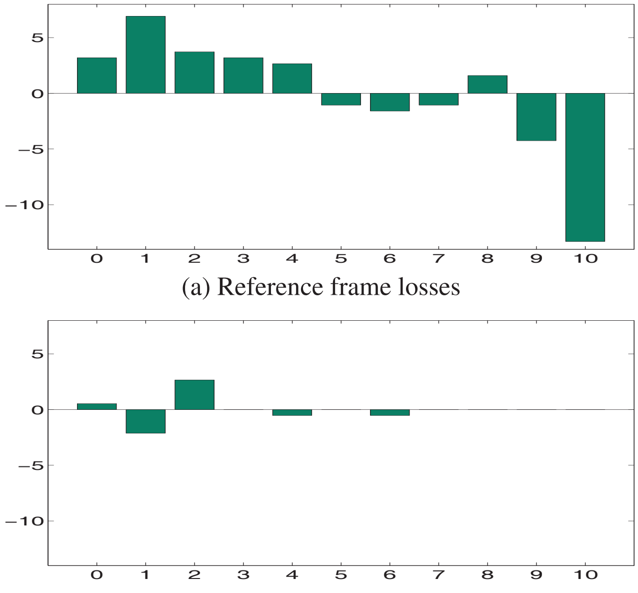
A TS packet loss was classified as invisible if two or fewer out of the ten viewers saw the loss. If seen by ≥ 3, the loss was categorized as visible. (This definition is arbitrary; it would be of interest to examine results using a more stringent definition.) For SDTV, 76.1% of the data are invisible, and for HDTV, 71.6%. Fig. 1 shows the number of observers who reacted to packet losses for PFL (Fig. 1(a)) and FL (Fig. 1(b)) cases. The distributions of visible PFL are similar for the two different resolutions. Over 80% of PFL packets are invisible for both resolutions. However, the FL cases for HDTV are more visible than they are for SDTV. Fig. 1(b) shows that the distribution of observed frame losses for HDTV is shifted to the right compared to the SDTV data. The fraction of invisible data is 51.6% for SDTV and 36.2% for HDTV.

As a preliminary exploration of the differences in observer responses, the data for whole frame losses was divided into reference frame losses and non-reference frame losses. Fig. 2(a) shows the subtraction of the distributions for SDTV and HDTV reference frame losses. Fig. 2(b) is the corresponding figure for non-reference frame losses. Fig. 2(b) appears to indicate that data of both resolutions are similar for non-reference frame losses. However, the distribution of reference frame losses shows a clear resolution-dependent trend: reference frame losses are more visible in HDTV.

Frame Loss Type	I	P$_1$	P$_2$	P$_3$	P$_4$	B$_1$	B$_2$
Number of losses	6	8	14	32	48	40	40

Partial Frame Loss Type	I	P$_1$	P$_2$	P$_3$	P$_4$	B
Number of losses	104	108	120	124	124	132

Table 1. Numbers of frame losses (top) and partial frame losses (bottom)



(a) Reference frame losses

(b) Non-reference frame losses

Fig. 2. SDTV distribution minus HDTV distribution.

3. SUPPORT VECTOR MACHINE CLASSIFIER

Tables 2 and 3 list the factors for the PFL and FL cases which will be extracted at the encoder from an individual TS packet to predict the visibility. Macroblocks which are part of the lost packet but which are perfectly concealed by the ffmpeg error concealment approach are not counted as being lost. Therefore the factors IMSE, MBnum, aveX and aveY which consider lost macroblocks do not include the perfectly concealed macroblocks in the computation.

The only factor in common between FL and PFL classifiers is TMDR, which is the number of frames affected by the TS packet loss. For B_2 frames, TMDR = 1. For other frames, TMDR depends on the distance to the next I frame. The maximum value TMDR can have is 17 because our GOP has 15 frames, so an I packet loss can corrupt all frames in the GOP as well as the last two B frames from the previous GOP.

Factor	Meaning
TMDR	Temporal Duration
IMSE	Mean square error per lost pixel
MBnum	Number of lost MBs
aveX	Average horizontal location of lost MBs
aveY	Average vertical location of lost MBs

Table 2. Extracted factors for PFL classifiers

Factor	Meaning
TMDR	Temporal Duration
ResEng	Average residual energy of luminance component per pixel after motion compensation in the lost MBs
aveMVx	Average motion in horizontal direction
aveMVy	Average motion in vertical direction
magMV	Average of motion vector magnitudes
angMV	Average of motion vector angles
stdMVmag	Variance of motion vector magnitudes
stdMVang	Variance of motion vector angles

Table 3. Extracted factors for FL classifiers

MBnum, aveX, and aveY are meaningful for PFLs, but not useful for FLs, as they have constant values (for SDTV, MBnum=1350, aveX=360, aveY=240) for FL packets.

IMSE (Initial Mean Square Error) is the mean squared error per pixel between the decoded videos with and without a TS packet loss evaluated over the pixels in lost macroblocks. IMSE, a factor used for PFL, is ill-defined for FL since the lost frame is discarded. Instead of attempting to estimate a value for IMSE, the FL classifier instead uses ResEng, the average residual energy of the luminance component per pixel after motion compensation. Since there are no motion vectors in an I-frame, when the packet being evaluated occurs in an I-frame, the motion vectors of the previous frame (non I-frame) are used for the six factors related to motion vectors.

We divide the data into four data sets, based on resolution (SDTV/HDTV) and loss type (FL/PFL). For each data set, the factors are extracted. Instead of using these factors directly as features of the support vector machine (SVM) classifier [7], we calculated a Feature Matrix:

$FM[j,i] = \sum_{n=1}^{N}(C_n(P_{i,n} - P_{j,n}))^2$ for $1 \leq i, j \leq K$ where K is the number of data points, N is the number of factors, C_n is the absolute correlation coefficient between the loss visibility and the nth factor, and $P_{i,n}$ denotes the value of the nth factor for the ith data point. The values of C_n are listed in Table 4. For PFL, $N = 5$ and $K = 712$, and for FL, $N = 8$ and $K = 188$. After calculating a feature matrix, each column is normalized to zero mean and unit variance. We train a SVM classifier with the feature matrix under a radial basis function kernel using the LIBSVM software [8].

Factor of PFL	SDTV	HDTV
TMDR	0.2005	0.1830
IMSE	0.0468	0.1541
MBnum	0.4302	0.3985
aveX	0.1409	0.0207
aveY	0.0444	0.0345

Factor of FL	SDTV	HDTV
TMDR	0.4662	0.5183
ResEng	0.1766	0.1407
aveMVx	0.0688	0.0313
aveMVy	0.0566	0.0411
magMV	0.4009	0.3300
angMV	0.4367	0.3332
stdMVmag	0.0684	0.0524
stdMVang	0.1671	0.1085

Table 4. Absolute values of correlation coefficient

4. RESULTS

The performance of a classifier is measured by 8-fold cross validation. We calculated performance with *all* combinations of factors. The best set of $i + 1$ factors always included the best set of i factors. Figs. 3 and 4 show classification accuracy as we add more factors. The order of adding factors is given in Tables 5 and 6. The first line in Tables 5 and 6 shows the accuracy for the default classifier which uses no factors and classifies all PFL packets and SDTV FL as invisible, and which classifies all HDTV FL packets as visible.

Frame Loss Classifiers: Only three factors, TMDR, magMV, and stdMVmag, contribute significant performance improvement in both resolutions. If a SDTV FL classifier uses only TMDR, the performance is merely 2.7% lower than with the magMV factor used alone. However, the HDTV FL classifier performance using only magMV is 11.7% worse than with TMDR alone. Therefore, TMDR is a very important factor in both FL classifiers although the listed orders are slightly different. TMDR values differ for reference and non-reference frames, and also differ among reference frames. HDTV tests showed only 5% of non-reference frame (B_2) losses are visible and 79.7% of reference frame losses are visible. The performance improvement of the final FL classifiers is 36.6% for SDTV and 27.1% for HDTV.

Partial Frame Loss Classifiers: The performance of the classifiers is 86.5% and 83.4% for SDTV and HDTV resolutions. This represents improvements of only 3.8% and 2.6% by using five factors relative to the default classifier which classifies all TS PFL packets as invisible. Because the training sets of PFL are so unbalanced towards the invisible class, the SVM does not have sufficient opportunity to learn about visible data. Fig. 4 shows the performance improvement by adding factors. Four factors are useful for the SDTV case, and only two factors for HDTV. The number of lost macroblocks is the most significant factor for both resolutions.

Discussion and Conclusions:

- TS packet losses are more visible for HD than for SD. This is probably because in our viewing conditions (which follow the DSL Forum recommendations) HD pictures occupy a much larger field of view than SD. One would expect it to be harder to detect glitches in pictures with a smaller angular field of view; this intuition is borne out by the SD FL data.
- Prediction accuracy for the SD PFL classifier is slightly higher than for the HD PFL classifier. This may be because, for our encoding parameters, the 188-byte TS packet size covers a higher fraction of the frame size in the SD case than in the HD case, and prediction may be slightly easier with a higher fraction of the frame data available.
- Subjective experiments show that, for whole frame losses, there is a substantial difference in the visibility of the loss for HDTV and SDTV for reference frame losses, but not for non-reference frame losses. Reference frame losses in HD are mostly visible whereas reference frame losses in SD are not.
- Over 80% of partial frame losses are invisible in both resolutions. The ability of a PFL classifier to extract useful parameters from a TS packet and improve on the baseline visibility prediction is very limited.
- In contrast, for TS packets which lead to whole frame loss, extraction of only three simple factors from the TS packet (the time duration of the loss, and two factors that depend on motion vectors) can hugely improve the prediction of packet importance (loss visibility).

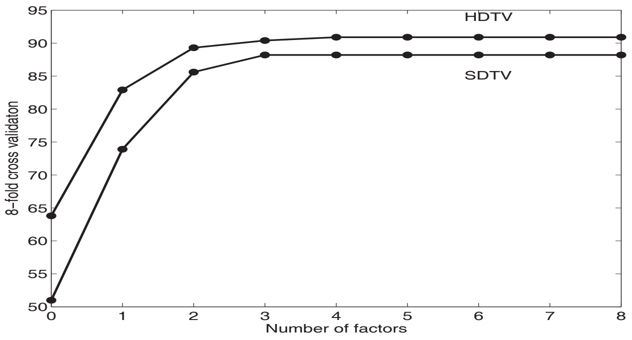


Fig. 3. FL classification accuracy vs. number of factors

5. REFERENCES

[1] S. Kanumuri, P. Cosman, and A. Reibman. A generalized linear model for MPEG-2 packet-loss visibility. *Intl. Packet Video Workshop*, Dec 2004.

[2] S. Kanumuri, P. Cosman, A. Reibman, and V. Vaishampayan. Modeling packet-loss visibility in MPEG-2 video. *IEEE Transactions on Multimedia*, 8:341–355, Apr 2006.

SD Factors	Accuracy	HD Factors	Accuracy
None (inv)	51.6%	None (vis)	63.8%
magMV	73.9%	TMDR	82.9%
TMDR	85.6%	stdMVmag	89.3%
stdMVmag	88.2%	magMV	90.4%
stdMVang	88.2%	stdMVang	90.9%
angMV	88.2%	angMV	90.9%
aveMVy	88.2%	aveMVy	90.9%
aveMVx	88.2%	aveMVx	90.9%
ResEng	88.2%	ResEng	90.9%

Table 5. Performance from adding factors in FL classifiers

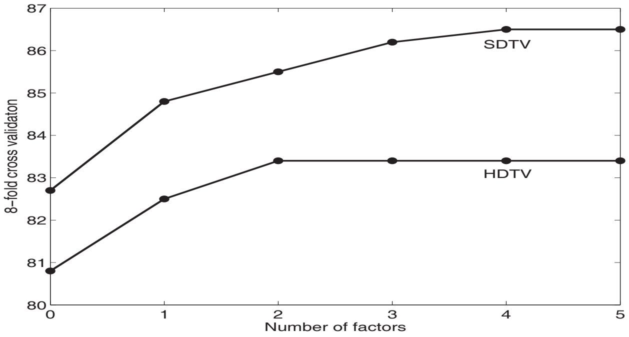


Fig. 4. PFL classification accuracy vs. number of factors

[3] T.-L. Lin, Y. Zhi, S. Kanumuri, P. Cosman, and A. Reibman. Perceptual quality based packet dropping for generalized video GOP structures. *ICASSP*, 2009.

[4] ITU-R BT.710-4 Subjective Assessment Methods for Image Quality in High-Definition Television. Jan 1998.

[5] DSL Forum Technical Report TR-126: Triple-play Services Quality of Experience (QoE) Requirements. Dec 2006.

[6] The official website of ffmpeg : http://ffmpeg.org/.

[7] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[8] C.-C. Chang and C.-J. Lin. The official website of LIBSVM : http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

SD Factors	Accuracy	HD Factors	Accuracy
None (inv)	82.7%	None (inv)	80.8%
MBnum	84.8%	MBnum	82.5%
TMDR	85.5%	TMDR	83.4%
aveX	86.2%	aveX	83.4%
aveY	86.5%	aveY	83.4%
IMSE	86.5%	IMSE	83.4%

Table 6. Performance from adding factors in PFL classifiers