NETWORK-BASED PACKET LOSS VISIBILITY MODEL FOR SDTV AND HDTV FOR H.264 VIDEOS

Ting-Lan Lin and Pamela C. Cosman

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

ABSTRACT

We conduct subjective experiments on visual quality following packet loss, and then construct models to predict these visual importance scores. The models are fully self-contained at the packet level, meaning that they use only information within one packet to predict the importance of that packet, requiring no frame-level reconstruction nor any information on the reference frame. Models are created for SDTV and HDTV resolutions, and the differences in the important factors between them are discussed.

Index Terms— SDTV, HDTV, packet loss, video quality, network monitoring

1. INTRODUCTION

Since different video packets have different impact on visual quality when dropped, it is important for an intermediate router to estimate the visual importance of each packet to know which ones to drop during congestion. Our prior work [1] built a generalized packet loss visibility model for different GOP structures. We assigned each packet a priority bit at the encoder so that the router could perform smart dropping during congestion. The model in [1] is an encoder-based model; it requires factors such as Initial MSE, type of camera motion, information on the reference frame and on scene cuts. This is applicable at the encoder where the reference frame is available, and where the computational capability is high. In the current work, we focus on a network-based model where the complexity must be limited, and in any case, reference frames are not necessarily available because packets may be out of order or because there are multiple streams and the network node cannot afford to decode and reconstruct them.

A second goal is to explore the difference between SDTV and HDTV packet visibility models. Subjective results in [2] showed that displays should guarantee a large screen with high contrast to achieve the higher expectation for watching HDTV than for watching SDTV. The work in [3] concluded that people prefer SDTV with high quality over HDTV with low quality. These works comparing SDTV and HDTV are not concerned with packet loss visibility. One related paper is [4], which studied region of interest (ROI) determination for SDTV and HDTV. The study showed that the ROI of a video is identical for both SDTV and HDTV. Also, losses occurring in the top and the bottom regions of the picture were not generally in the ROI.

This paper is organized as follows: In Section 2, the subjective tests are described. In Section 3, we discuss self-contained factors that relate to packet loss visibility, and the models based on these factors. Section 4 presents results and discussion.

2. SUBJECTIVE EXPERIMENTS

The video encoder is H.264 JM9.3. Encoder settings (Table 1) adhere to ITU and DSL Forum Recommendations [5, 6]. Each Network Abstraction Layer (NAL) packet contains a horizontal row of Macroblocks (16×16 pixels) in a frame. There are 30 packets per SD frame, and 68 per HD frame. The raw video sources are in HD format, and the SD versions are obtained by downscaling the HD videos by bicubic interpolation. Nine videos with widely varying motion and texture characteristics are concatenated into a 20-minute sequence.

The decoder is FFMPEG [7] due to its high efficiency and wide use in industry. For error concealment, the FFMPEG decoder begins by estimating whether each lost macroblock is more likely to have been intra or inter coded. For example, in P and B frames, if more than half of received macroblocks are intra coded, the algorithm will guess that all lost macroblocks in the frame were coded intra. For the macroblocks which are guessed to be intra coded, FFMPEG conceals using a weighted average of uncorrupted neighboring blocks. For the macroblocks which are guessed to be inter coded, the algorithm estimates the forward and backward motion vectors by using the colocated future and past motion vectors.

Each subject watches a lossy HD video and the corresponding SD version, 40 minutes in total. The experiment takes one hour, which includes an introductory session and a break. When viewers see a glitch, they press the space bar. To allow observers enough time to respond to each individual loss, only one packet loss occurs for every 4-sec interval. The loss occurs in the first 3 seconds, and the fourth second allows any error propagation to terminate. During the 40 minutes of video, there are 600 packet loss data points obtained from a

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

	SD	HD
Resolution	720×480	1920×1080
Bitrate	2.1 Mbps	10 Mbps
H.264 Profile	Main profile Level 3	Main profile Level 4
Viewing Distance	6H	3H
Frame rate	30 fps	
GOP	IBBPBBPBBPBBPBB 15/3	

Table 1. Summary of the subjective experiment setup for SD and HD videos. H is the height of the video.

subject. These losses are divided equally among I frames, P frames and B frames. There are three different loss realizations; each of the three 40-minute lossy video pairs is watched by 10 people. The ground truth packet loss visibility for a specific packet can be obtained as the number of people who see the loss artifact divided by 10. With three loss realizations, each evaluated by 10 people, we have ground truth visibility for 600×3=1800 packets (900 for SD, 900 for HD resolution).

3. FEATURES AND MODEL BUILDING

In this section, we first introduce candidate factors associated with a packet. Next, we build models using these parameters to predict, for each packet, the packet loss visibility results of our subjective experiment.

Content dependent factors depend on the actual video content at the location of the loss. The ones we use all involve taking a mean, maximum, or variance computed over all macroblocks in the packet. **MeanRSENGY** is the mean residual energy after motion compensation. **MaxRSENGY** denotes the maximal residual energy after motion compensation. Following the way these factors were used in [1, 8], we used the above two terms after logarithm because they were shown to be more correlated with packet loss visibility (we add 10^{-7} before taking the log to avoid a log of zero problem). **MeanMotX** and **MeanMotY** are the mean motion vectors in the x and y directions. **MaxMotX** and **MaxMotY** are the maximal motion vectors. **VarMotX** and **VarMotY** are the variances of the motion vectors. **MotM** is $\sqrt{MeanMotX^2 + MeanMotY^2}$. To compute the factors related to phase of motion vectors, we only consider macroblocks with non-zero motion, for which the phase is well defined. **MeanMotA** is the mean phase. **MaxMotA** is the maximal phase. **MaxInterparts** is the maximal number of inter macroblock partitions in the packet.

Content independent factors depend on, for example, the spatial location or frame type of the loss, but do not depend on the actual video content at the location of the loss. **TMDR** is the maximum number of frames to which the error from this packet loss can propagate. TMDR=1 for non-reference frames. For reference frames, TMDR depends on the distance to the next I frame. **Height** is the spatial location where the

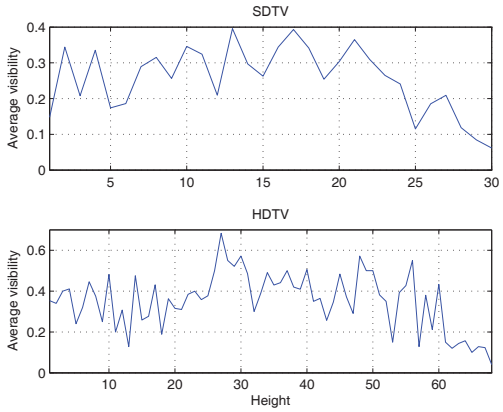

Fig. 1. Average packet loss visibility versus **Height**

loss occurs; the top slice in a frame has Height=1, and the bottom slice in a frame has Height=N, where N is the number of packets in a frame (30 for SDTV and 68 for HDTV). Most of the factors mentioned above have a monotonically increasing (or decreasing) relationship with the average packet loss visibility. However, this is not the case for Height. The plots of average packet loss visibility versus Height are in Fig.1. Although the data are noisy, we see the trend that average packet loss visibility is highest near the middle of the screen, and decreases as we move to the top or bottom. This is difficult to capture by a linear relation, therefore, we create **DevFromCenter** = abs(Height-floor(N/2)) to indicate how far away the loss occurs from the vertical center of the frame.

In addition to these content independent and content dependent factors, we also consider the interactions between factors in one category and factors in the other, as well as between factors within the content independent category.

The motion information mentioned above is estimated by the network node where reference frames are not available. In some cases, the "true" values for those quantities require the reference frames. For example, the "direct" mode of coding a macroblock assumes that an object is moving with constant speed, so the motion vector for the current MB is copied from the previous co-located MB. Within a packet, we do not have any information on the previous co-located macroblock. We instead copy the motion vector from a spatial neighbor. This way, the model is fully self-contained at the packet level, and can be implemented at a network node.

In the experiment and data analysis, we assume each viewer's response is an independent observation of the average viewer (for whom we are developing the model). Therefore, each viewer response can be considered iid with probability p for seeing a particular packet loss. Hence, we choose a generalized linear model (GLM) with the logit function as link function, since it can predict a probability parameter in a binomial distribution. We want to know the probability that a packet loss artifact will be observed when the packet is lost. A GLM with a logit function for the binomial distribution has the form

$$log(\frac{p}{1-p}) = \gamma + \sum_{j=1}^{P} x_j \beta_j \quad (1)$$

where $\beta_1, \beta_2, ..., \beta_P$ are the coefficients of the P factors considered for prediction, and γ is the constant term. Often the parameters of the GLM are estimated such that the resulting model has the least deviance (the deviance is a generalization of the residual sum of squares). This treats data points equally, no matter how far they are from the regression line. However, outliers may distort the results. To give unequal treatment to data points to suppress outliers, we minimize the M-estimator [9]; data points farther from the regression line have smaller weights, and contribute less to the final modeling result. We chose the "Fair" function as the M-estimator function, shown in Figure 2. The M-estimator is computed as the sum of the weighted residual squares, where the weight of each data point is computed by the residuals in the previous iteration. The M-estimator function in Figure 2 is chosen to avoid the weights of the curve going close to zero at the two ends, because we do not want to have a final model that has least M-estimator just because most of the data points are at the two ends. The model developing procedure uses 4-fold cross validation to prevent the model overfitting the data, so an average M-estimator is produced for a set of factors. The factor which most reduces the average M-estimator goes next into the model. This procedure repeats until there is no improvement in the average M-estimator by including an additional factor. We develop GLM models for both SD and HD resolution videos. The best factors chosen for them and their corresponding coefficients are listed in Tables 2 and 3. Figures 3 show the decrease of the M-estimator as additional factors are incorporated in the SD and HD models.

4. RESULTS AND DISCUSSION

From Figure 3, we see the best M-estimator value is 0.1096 for the SDTV model and 0.1201 for the HDTV model. If we compare against an encoder-based model which uses initial MSE, requiring the reference frame and frame reconstruction, as a factor, the encoder-based models perform better as expected; the minimum achievable M-estimators are 0.1067 for SDTV and 0.1172 for HDTV, as shown in Figure 3. However, the performance difference is slight; the network-based model performs almost as well as the encoder-based model, but the former is suitable for a router as it uses no information from reference packets or pixel domain processing.

We can not properly interpret the model by the sign of the coefficients in Tables 2 and 3 if the factors correlate with each other [10], however the order in which factors are added to the model provides an indication of their importance. The most important factors in both SDTV and HDTV relate to TMDR, indicating that error propagation duration dominates the packet loss visibility regardless of resolution. However spatial location of the loss affects the visibility differently between models. In Fig.1, the maximum average loss visibility

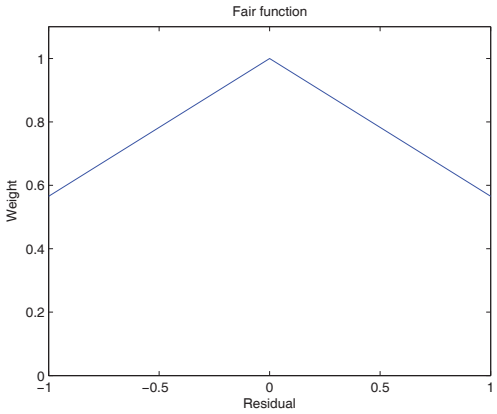

Fig. 2. The Fair function versus the residual.

is 0.3957 at Height=13 for SDTV, and 0.6833 at Height=27 for HDTV; they are both near the middle slice. The minimum average loss visibility is 0.0615 at Height=30 for SDTV, and 0.0400 at Height=68 for HDTV; they are both at the bottom slice. Packet losses in the center are more visible than those at the bottom. What is more, given that the average packet loss visibility for all losses in SDTV is lower than that in HDTV (0.2565 and 0.3506), it is surprising that the average loss visibility of the bottom packet in HDTV is lower than that in SDTV, and the ratios of maximum loss visibility to minimum loss visibility are 6.4341 and 17.0825 for SDTV and HDTV respectively. In the viewing conditions of Recommendations [5, 6], HD requires a larger viewing angle. The viewing angles are (vertical, horizontal)=$(9.52°, 14.25°)$ for SDTV, and $(18.92°, 33°)$ for HDTV. Therefore, a viewer who watches HDTV may not fully realize what happens in the edge area of a frame. Prior research [4] found that losses occurring in the top and the bottom regions of the picture were not generally in the region-of-interest. We would add to this that, for HDTV, losses occurring at the top and bottom are less likely to be noticed not only because they are not in the ROI but also because of the larger viewing angle.

Conclusion: We propose self-contained packet loss visibility models for SDTV and HDTV. These network-based models perform only slightly less well than the much more complicated non-self-contained models that could be implemented only at the encoder. The proposed models allow a network node to efficiently evaluate the visual importance of packets just by information contained in each packet. No reference information or frame reconstruction is required for the predicting factors. This model can be useful to evaluate packets in the network in case of congestion. The study found that packet loss is more visible in HDTV than in SDTV. And due to a wider viewing angle for HDTV, the spatial location of the packet loss in HDTV matters more than in SDTV. For both SDTV and HDTV models, the temporal duration of the error propagation is a very important factor for a packet to be visible.

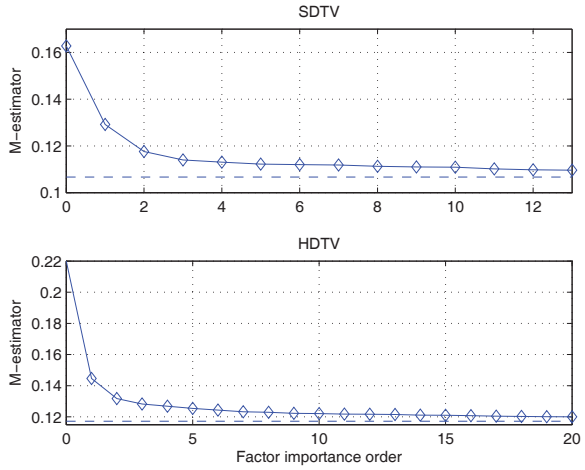

Fig. 3. M-estimator value decreases as important factors are included in the SDTV and HDTV model. Numbers on x-axis denote the index in factor order shown in Table 2 and 3. The dashed horizontal line denotes the minimum M-estimator value of the SDTV and HDTV encoder-based models.

5. REFERENCES

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

[2] M. Ardito, M. Gunetti, and M. Visca. Influence of display parameters on perceived HDTV quality. *IEEE Transactions on Consumer Electronics*, 42, Feb 1996.

[3] S. Péchard, M. Carnec, P. Le Callet, and D. Barba. From SD to HD television: effects of H.264 distortions versus display size on quality of experience. *ICIP*, 2006.

[4] F. Boulos, W. Chen, B. Parrein, and P. Le Callet. A New H.264/AVC Error Resilience Model Based on Regions of Interest. *Packet Video*, June 2009.

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

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

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

[8] A. R. Reibman and D. Poole. Predicting packet-loss visibility using scene characteristics. *Packet Video*, 2007.

[9] W. Rey. *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer, 1983.

[10] G. Mullet. Why regression coefficients have the wrong sign. *Journal of Quality Technology*, 8(3), 1976.

Order	Factors	Coefficients
α	1	-2.6407
1	TMDR×MaxMotA	-4.7591e-3
2	**DevFromCenter**×MaxMotA	2.2996e-2
3	**Height**×MeanMotA	-8.8462e-4
4	TMDR×log(MeanRSENGY +10^{-7})	3.5954e-3
5	TMDR×MeanMotY	-1.6431e-2
6	**DevFromCenter**×TMDR	-1.0164e-2
7	**DevFromCenter**×MeanMotY	5.3172e-3
8	TMDR	2.3680e-1
9	TMDR×MaxInterparts	-5.6283e-3
10	TMDR×MotM	4.9349e-3
11	**Height**×**DevFromCenter**	-3.1830e-3
12	**Height**×MaxInterparts	2.1661e-3
13	TMDR×VarMotY	5.1232e-4

Table 2. Table of factors in the order of importance for SD GLM model. The × symbol means interaction. Bolded factors relate to spatial location.

Order	Factors	Coefficients
α	1	-3.0413
1	TMDR×log(MaxRSENGY +10^{-7})	9.1743e-3
2	**Height**×**DevFromCenter**	-2.1129e-3
3	**Height**×TMDR	3.4239e-4
4	TMDR×MaxMotA	6.0561e-2
5	**Height**×MotM	9.9631e-4
6	**Height**	3.2186e-2
7	**DevFromCenter**×MeanMotY	1.3397e-3
8	**Height**×VarMotX	-2.0544e-5
9	TMDR×VarMotX	3.8690e-4
10	TMDR×MeanMotX	3.3589e-3
11	**DevFromCenter**×TMDR	-4.7789e-3
12	log(MaxRSENGY +10^{-7})	-6.5376e-2
13	**DevFromCenter**	7.6811e-2
14	**Height**×MaxInterparts	7.9892e-4
15	**DevFromCenter**×MaxInterparts	-9.3612e-4
16	**DevFromCenter**×MaxMotY	-6.7759e-4
17	**DevFromCenter**× log(MeanRSENGY +10^{-7})	3.9123e-3
18	TMDR×MeanMotY	2.1333e-3
19	VarMotY	2.3235e-4
20	TMDR×log(MeanRSENGY +10^{-7})	3.1425e-3

Table 3. Table of factors in the order of importance for HD GLM model. The × symbol means interaction. Bolded factors relate to spatial location.