NETWORK-BASED IP PACKET LOSS IMPORTANCE MODEL FOR H.264 SD VIDEOS

Yueh-Lun Chang, Ting-Lan Lin and Pamela C. Cosman

University of California, San Diego
Dept. of Electrical & Computer Engineering
9500 Gilman Dr., La Jolla, CA 92093-0407

ABSTRACT

We conduct an objective experiment in which Video Quality Metric (VQM) scores are computed on compressed video GOPs following fixed-sized IP packet loss, and then construct a network-based model to predict these VQM scores. The model is created for H.264 SDTV videos using a no-reference method, meaning that we only use the information from the bitstream but have no access to the original video. The model can be computed at the packet level and requires no frame-level reconstruction.

Index Terms— SDTV, IP, fixed-sized packet loss, video quality, VQM

1. INTRODUCTION

When video is transmitted through a network, there can be packet losses due to congestion or collision. Different video packet losses will cause different impact on the video quality, and it is crucial for an intermediate router to know the visual importance of each packet to decide which ones to drop during congestion. Many of the previous research works have focused on the average quality of video subjected to an average packet loss rate. However, we would like to emphasize the influence on the video quality of an isolated or individual packet loss.

Our prior work [1] built a generalized packet loss visibility model using subjective tests for different encoding standards and GOP structures. The model was applied to packet prioritization for a video stream. Each packet was assigned a priority bit at the encoder so the router could perform smart dropping when the network was congested. In [2], we allocated more Forward Error Correction (FEC) bits to high visibility packets to give them more protection, so as to minimize end-to-end video quality degradation due to packet losses. The models in [1] and [2] are encoder-based models, which are assessed by full-reference methods and need parameters such as MSE, type of camera motion, and information on scene cuts. These demand access to the original video at the encoder and have high computational complexity. In contrast to an encoder-based model, for a network-based model, the original video information is unavailable in the network, and the computational capability is also limited.

In addition to making a network-based model, a second goal of the current paper is to build a model for fixed-sized packets. In the network, video is typically packetized in one of two ways: it can be segmented into a variable-sized packet which contains a constant area in the frame; the other way is using fixed-sized packets which may correspond to different pixel areas but whose sizes in bits are the same, such as MPEG-2 Transport Stream packets. In our previous work [3], we proposed a packet loss visibility model for H.264 SD and HD videos for variable-sized packets which contain one slice for each Network Abstraction Layer (NAL) unit. In this paper, we would like to construct a network-based model for fixed-sized IP packets to predict visual importance using an objective experiment.

The paper is organized as follows: In Section II, the design of the objective experiment is described. In Section III, we discuss the factors used to predict the quality scores of a loss, and the model based on these factors. Section IV presents results and discussions, while Section V summarizes our conclusions.

2. OBJECTIVE EXPERIMENT

We conduct an objective experiment to construct a visual importance model. Nine SD resolution H.264 videos are used for our experiment. The encoder is H.264 JM9.3, and the settings can be found in Table I. These settings adhere to ITU and DSL Forum Recommendations [5, 6]. It is high quality compression so there are few encoding artifacts between the coded videos and original ones. In these videos, each slice contains a horizontal row of Macroblocks (16×16 pixels) in a frame, and each NAL contains one slice. There are 300 frames in each video and the content includes various types of motion, texture characteristics and camera operations. The decoder is FFMPEG [7] due to its high efficiency and wide use in industry. For error concealment, the FFMPEG decoder begins by estimating, for each lost macroblock, whether it is more likely to have been intra coded or inter coded. Based on the estimate, the algorithm uses one of two different approaches to conceal each lost MB [8].

Resolution	720×480
Bitrate	2.1 Mbps
Profile	Main profile, Level 3
Frame rate	30 fps
GOP	IBBPBBPBBPBBPBB 15/3

Table 1. Summary of the objective experiment setup for SD videos. These settings adhere to Recommendations [5] and [6].

2.1. Video Quality Evaluation

There are many objective ways to evaluate the video quality such as MSE or PSNR. In our experiment, we use a tool called Video Quality Metric (VQM) [9]. VQM was developed by the Institute for Telecommunication Science to provide an objective measurement for perceived video quality. It is a computable quality metric to evaluate the processed version by comparison with the original lossless video. It assigns a score to an entire video, or to a segment of video such as a GOP (group of pictures). VQM scores range from zero to one, where a lower score means higher quality with less degradation. It has been shown to correlate well with human perception of the video quality and has been adopted by ANSI as an objective video quality standard. Here we would like to develop a model to predict VQM scores using simple features that can be extracted from individual packets.

2.2. Packetization

In order to packetize the H.264 SD videos into fixed-sized packets that can be transmitted through the network, the detailed steps are described in this subsection.

2.2.1. Transport Stream

The Transport Stream (TS) is defined in MPEG2-Part1 [10]. It is a digital container format that encapsulates different types of information such as video, audio or data. In [10], it describes how to mux several streams into a single one. The transport stream uses fixed-sized packets as its basic transfer unit. There are many advantages to using fixed-sized packets. It is convenient to detect the start and end of a frame and also easy to recover from packet loss or corruption.

A freeware tsMuxer [11] developed by the company SmartLabs is used to mux H.264 videos into regular TS packets. Each TS packet is fixed-sized with 188 bytes in length and only contains information from the same frame.

2.2.2. IP Packet

Although the transport stream specifies how to packetize multimedia information, the actual transmit unit over the network is an IP packet. Some major applications of video transmission over IP are: conversational applications such as video telephony and videoconferencing, the download of complete, pre-coded video streams, and IP-based streaming such as YouTube [12].

By the protocol specification, the size of an IP packet is variable and can be up to 64 kbytes, but this size is rarely used. The reason is that a large IP packet needs fragmenting in order to pass onto the Ethernet since the payload size of the maximum transfer unit (MTU) for an Ethernet packet is 1500 bytes. To avoid splitting and recombining IP packets larger than the MTU payload size, we took the size of each IP packet to be less than 1500 bytes. Specifically, in our experiment settings, one IP packet contains seven TS packets (188×7=1316 bytes). The packet size would exceed the limitation of 1500 bytes if more than 7 TS packets were included. Figure 1 shows the entire encoding and packetization process from original video to IP packets. Our goal is to construct a model to predict the VQM score associated with each IP packet, that is, the VQM score for the GOP that would result from the loss of that single IP packet.

2.3. Lossy Test Videos

In our experiment, we drop an IP packet from a GOP to create a lossy video and use VQM to evaluate its quality after packet loss. There are three possibilities: 1) a packet contains only one slice or a part of one slice, 2) a packet contains more than one slice, 3) a packet contains a frame header. These will cause the loss of 1) one slice, 2) several slices, and 3) an entire frame.

In order to calculate a VQM score, the number of frames in the original video and in the lossy video must be the same. If a frame header is dropped, the number of frames in the lossy video cannot be kept the same. Moreover, loss of a slice header in an I frame will cause a serious degradation to the video due to the way FFMPEG decodes (the decoder does not work properly when the first slice is lost). Based on these reasons, the following two types of packets were considered to be the most important:

1. An IP packet with any frame header
2. An IP packet with an I slice header

Among all the IP packets from our test videos, less than 5% of them contain a frame header or I slice header, and these packets with the highest priority will not be dropped in our experiment. The reason for not including these in the experiment is that our goal for predicting packet-level VQM scores is to allow a router to choose which ones to drop. For these packets of highest priority, it is already known that dropping them should be avoided if at all possible. The goal therefore is to guide the router in choosing which of the other $> 95\%$ of packets should be selected for dropping in case of congestion. After the dropping is performed for a GOP, the FFMPEG

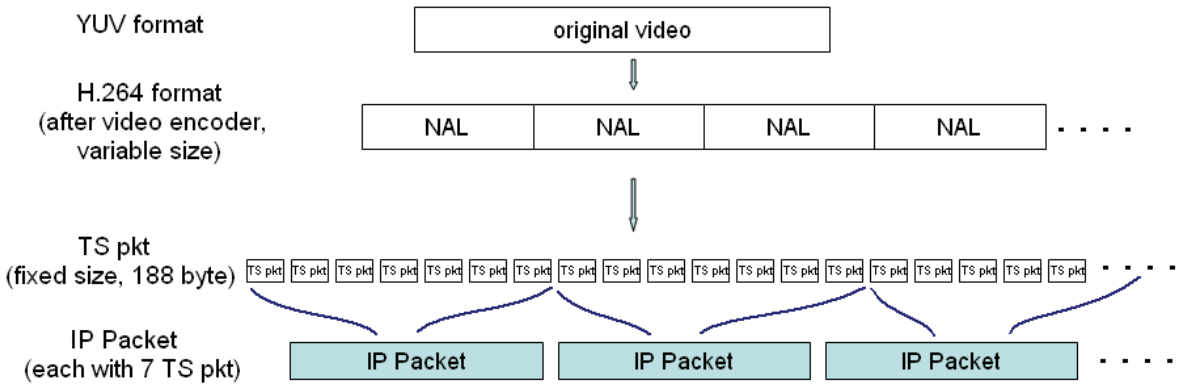

Fig. 1. The encoding and packetization procedure from original video to IP packets.

decoding and error concealment are run, and then the VQM score is calculated to obtain the objective video quality score for this GOP.

The last three frames of every GOP are excluded from being the location of the packet loss because the VQM algorithm ignores differences between the videos under comparison which occur at the end of the GOP. A total number of 931 IP packet losses are divided equally and randomly among all the I frames, P frames, and B frames.

3. FEATURES AND MODEL BUILDING

Our network-based model is built by using a no-reference method that only has access to the bitstream in the network while the original video is unavailable. The information we use does not require pixel data. This is desirable because the parameter extraction process can be made very efficient at a network node since it does not involve motion compensation (requiring reference frame), deblocking filter and frame reconstruction. In this section, the candidate features are described first and then the modeling approach will be explained.

3.1. Features

The features used to construct the model are introduced here. They can be classified into two categories: content independent and content dependent features. A buffer is used to aggregate some number of IP packets for feature extraction. Many IP packets contain no slice start code, so we have to gather information from their adjacent packets. However, there is no need for frame level reconstruction.

Content independent features only require the general information of the packet, for example, spatial and temporal location or frame type. The content independent features we considered are the following:

1. **TMDR** stands for time duration. It is the maximum number of frames that can be affected by the packet loss due to error propagation. TMDR=1 for non-reference frames. For reference frames, TMDR depends on the distance to the next I frame.

2. **DevFromCenter**=abs(Height-floor(N/2)) indicates how far the loss is from the center slice (in vertical direction) of the frame. Height indicates the spatial location of the packet, and N is the number of slices in a frame. In our experiment, the number of slices for SDTV videos is 30.

3. **IsIFrame**, **IsPFrame** and **IsBFrame** are boolean factors which are set when the packet is in an I, P or B frame.

4. **NAL_num** is the total number of slices in the packet, and **NAL_size** is the aggregate size in bits for every slice contained in the packet. Recall that a slice is one horizontal row of macroblocks. For example, consider an IP packet which contains one partial I slice whose total size spanning across several packets is 16000 bits. For this packet, *NAL_num* is 1 and *NAL_size* is 16000. For an IP packet which contains two partial P slices whose sizes are 8144 and 11488 bits, the *NAL_num* is 2 and *NAL_size* is 8144+11488=19632.

Content dependent features require the actual content of the lost packet, such as the motion in each direction. The motion-related features take calculations over all macroblocks in the lost packet to get their mean, maximum, or

variance of motion information. **MaxMotX**, **MeanMotX**, and **VarMotX** are the maximum, mean, and variance of the motion vectors in the x direction, while **MaxMotY**, **MeanMotY**, and **VarMotY** are the maximum, mean, and variance of the motion vectors in the y direction. **MaxMotA** and **MeanMotA** are the maximal and mean phase, where $MotA = \arctan(MotX/MotY)$. **MotM** is the magnitude of mean motion vectors in the x and y directions. It equals $\sqrt{MeanMotX^2 + MeanMotY^2}$. **MeanRSENGY** is the mean residual energy after motion compensation. This is calculated from the DCT coefficients, so no inverse DCT or pixel information is needed [8]. We used the term after logarithm, and 10^{-7} is added before taking the log to avoid a log of zero problem. **MaxInterparts** is the maximal number of inter macroblock partitions in the lost packet.

To construct the model, the above features are used as well as their interaction terms, which are the products of two features.

3.2. Modeling Approaches

We model the VQM score using logistic regression, a type of generalized linear model (GLM). GLM is an extension of classical linear models [13, 14]. It can be represented as

$$g(p) = \gamma + \sum_{j=1}^{K} x_j \beta_j \tag{1}$$

where $\beta_1, \beta_2, ...\beta_K$ are the coefficients of the K factors, and γ is the constant term. Coefficients β_j and γ are usually unknown and need to be estimated from the data. $g(.)$ is the link function, and p is the expected value of the predicted term, i.e., VQM score in our experiment. The canonical link function for logistic regression is set to be a logit function:

$$g(p) = log(\frac{p}{1-p}) \tag{2}$$

The simplest model is a null model which only has one parameter: the constant term γ. At the other extreme, the full model contains as many factors as there are data points. The goodness of fit for a GLM can be determined by its deviance. Deviance is a general term of variance. By definition, the deviance is zero for the full model, while the deviance is positive for all the other models. A smaller deviance means a better model fit. To obtain the model coefficients for the candidate factors, an iterative feature selection technique is implemented by Matlab.

To prevent overfitting, a 10-fold cross validation is applied. The data is randomly segmented into 10 groups, and we use nine out of the ten sets as the training set and the remaining as the test set. The procedure is repeated ten times, each time choosing a different set for testing.

4. RESULTS AND DISCUSSION

Figure 2 shows the histogram distribution of the actual VQM scores in our objective experiment. Higher VQM scores mean worse degradation of the video quality. In Figures 3 and 4, the plots of deviance and correlation of the actual and predicted VQM scores versus the factor numbers included are presented. While the null deviance is 71.8, the deviance of the model can be reduced to less than 45. The correlation gets higher when more factors are included. In Figure 4, however, there is a breakpoint as the factor number is around 10. The curve becomes nearly flat after this point, which means the improvement of the correlation is not much even if more factors are added to the model.

The nine most significant factors are chosen for our final model. The factors and their coefficients are listed in Table II in order of importance. The importance of a factor can be defined by the amount of deviance reduced for GLM. Notably, the factors in the model are somewhat complicated. Each factor is the interaction of two features rather than a single term. It is sometimes hard to directly interpret the meaning of factors by the sign of the coefficients since these factors are not independent of each other. (Refer to [15], which explains why sometimes the coefficient sign is not what we expect.)

We observed the following about the effect of factors on quality:

1. The frame type of the lost packet plays a crucial role in our model, and losses in P frames were most damaging. This may seem counter-intuitive. Typically one may consider that a packet loss from an I frame would cause more degradation to the video quality, while in our model, a packet loss from a P frame actually resulted in the worst quality. This is because we packetized the video using fixed-sized packets. For SDTV videos in our experiment, an I frame generally contains 200∼400 TS packets (approximately 30∼60 IP packets), whereas a P frame contains less than 100 TS packets (approximately 15 IP packets). The detailed statistics of TS and IP packet numbers for each video are shown in Table III. So one IP packet from an I frame covers on average 3.3% of the frame's area, whereas one IP packet from a P frame includes on average 13.3% of the frame's area. Sometimes the corrupted area could be as much as one-fourth or one-third of the whole frame, so the damage is worse. The actual VQM scores from different frame types are shown in Figure 5. The histogram of VQM scores in P frames is shifted to the right compared to the histogram for I frames, and the mean VQM score resulting from a packet loss in a P frame is 0.0886, higher (worse) than that in I frames, 0.0718.

2. Residual energy is quite important as well. Higher residual energy usually implies that the motion in the video is more complicated, or the texture is widely

Factor Number	Factors	Coefficients
Intercept (γ)	1	-4.1445
1	IsPFrame $\times$ log(MeanRSENGY $+10^{-7}$)	2.8457e-1
2	TMDR $\times$ NAL_size	-5.1068e-8
3	TMDR $\times$ NAL_num	3.9130e-2
4	NAL_num $\times$ IsPFrame	-2.1074e-1
5	NAL_size $\times$ MaxMotA	1.6223e-5
6	DevFromCenter $\times$ MotM	8.2868e-3
7	NAL_num $\times$ IsIFrame	1.4706
8	NAL_size $\times$ IsIFrame	-6.0151e-5
9	DevFromCenter $\times$ MaxMotA	-1.4744e-2

Table 2. Table of factors in the order of importance. The $\times$ symbol means interaction.

varied. A positive sign of the coefficient means that a packet loss with high residual energy will corrupt the video more and result in a higher VQM score.

3. Two out of the top three factors relate to TMDR. This indicates that error propagation duration is very important to determining packet loss impact on quality. Higher TMDR means that the corruption lasts longer and causes worse quality with higer VQM score. Therefore TMDR should be positively correlated with VQM score. For the two factor coefficients related to TMDR, however, one has a positive sign and the other has a negative sign. These terms can be factored to single features, and the effect on TMDR is the combination of them. For example, in our model, the part related to TMDR is:

$$-5.1068 \times 10^{-8}(TMDR \times NAL\_size) \\ +3.9130 \times 10^{-2}(TMDR \times NAL\_num)$$

This can be rewritten as:

$$TMDR(-5.1068 \times 10^{-8} NAL\_size \\ +3.9130 \times 10^{-2} NAL\_num)$$

So the coefficient of TMDR can be considered a variable β_{TMDR}, where

$$\beta_{TMDR} = -5.1068 \times 10^{-8} NAL\_size \\ +3.9130 \times 10^{-2} NAL\_num$$

Considering the range of NAL_size and NAL_num, β_{TMDR} is always a positive quantity, so that TMDR has an overall positive correlation with VQM score, as expected.

4. Not only the temporal but also the spatial information is important. Six factors are associated with NAL_size or NAL_num. These terms correlate with the corrupted area within one frame, and imply that the influence of spatial region which could be affected by the lost packet is quite prominent.

   Since NAL_num is the total number of slices in the packet, a larger value of NAL_num means a larger contaminated area and should generally mean higher VQM score. Since we do not drop packets with an I slice header, the NAL_num of a lost packet in an I frame is always 1, while it could be any number from $1 \sim 30$ for a lost packet in a P or B frame. As we mentioned before, packet loss in a P frame usually causes the worst degradation to the quality and the highest VQM score, while the damage is less bad from a packet loss in an I frame and it is the least in a B frame. Therefore, the effect of NAL_num is separated out by the boolean features IsI/P/BFrame.

   Although the factor NAL_num appears in the model in three different interaction terms (TMDR$\times$NAL_num, IsIFrame$\times$NAL_num, and IsPFrame$\times$NAL_num), the effect of NAL_num can be explained simply according to frame type.

   The coefficient of NAL_num for a loss in an I frame is a constant number $\beta_{NAL\_num\_I}$ since, for I frames, $TMDR = 15$ and $IsIFrame = 1$, so

$$\begin{aligned} \beta_{NAL\_num\_I} &= 3.913 \times 10^{-2} TMDR \\ &\quad +1.4706 IsIFrame \\ &= 2.0576 \end{aligned}$$

   The coefficient of NAL_num for a loss in a B frame is also a constant number $\beta_{NAL\_num\_B}$ since, for B frames, $TMDR = 1$, so

$$\begin{aligned} \beta_{NAL\_num\_B} &= 3.913 \times 10^{-2} TMDR \\ &= 0.0391 \end{aligned}$$

   Comparing the values of these two constants, we see $\beta_{NAL\_num\_I}$ is greater than $\beta_{NAL\_num\_B}$, so that the

average VQM score for a loss in an I frame will be higher.

The coefficient of NAL_num for a loss in a P frame is a variable $\beta_{NAL\_num\_P}$ depending on $TMDR$, where

$$\beta_{NAL\_num\_P} = 3.913 \times 10^{-2} TMDR \\ -2.1074 \times 10^{-1} IsPFrame$$

Recall that VQM does not count quality degradation in the last three frames of a sequence, so the TMDR value for a loss in a P frame could be 6, 9 or 12. This makes $\beta_{NAL\_num\_P}$ always a positive coefficient.

In summary, the coefficients of NAL_num are positively correlated with VQM scores, which means that a larger damaging area is always worse regardless the frame type.

5. The spatial location of the lost packet also plays a part. Analyzing the coefficient of DevFromCenter by the same method for TMDR, it generally carries a negative sign. Larger DevFromCenter means the damage is further away from the center of the video, so it's less visible with a lower VQM score.

6. MotM, the magnitude of motion, has a positive coefficient sign in the model since more movement means that a packet loss will cause a more serious degradation in quality and hence a higher VQM score.

7. Since IsIFrame, IsPFrame and IsBFrame are boolean factors and only take effect on a specific frame type, our model can be viewed in another way. Factors 1 and 4 are used in the model only for P frames. Factors 7 and 8 are used in the model only for I frames. These boolean factors (IsI/P/BFrame) construct submodels for each frame type.

5. CONCLUSION

We proposed a network-based visual importance model of fixed-sized IP packets for SD H.264 videos. The proposed model allows an intermediate node in the network to efficiently estimate the visual importance of a packet by information at the packet level. Our results from the objective experiment show that, for a fixed-sized IP packet, frame type is a quite significant factor in the model. Our most novel result is finding that a fixed-sized packet loss in a P frame is on the average worse than one in an I frame. Previous studies found that I-packet losses caused the worst degradation, but that result was for packets of fixed pixel area. For our packets which are of fixed size in bytes, a P-packet covers a much larger pixel area than an I-packet, and so causes more quality degradation when lost. The temporal and spatial location are also noteworthy for prediction.

Changing the fixed size of the packet or changing the resolution of the video would likely affect the model, and this would be of interest to study in the future.

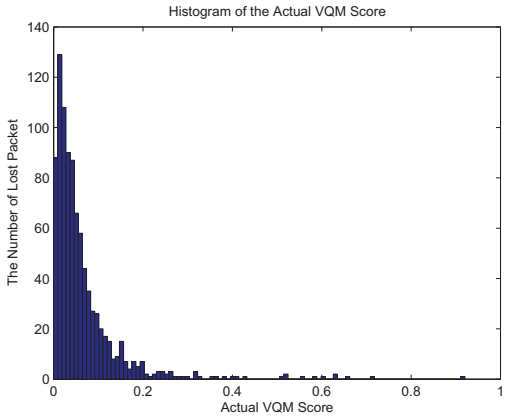

Fig. 2. Histogram of VQM scores from the objective experiment.

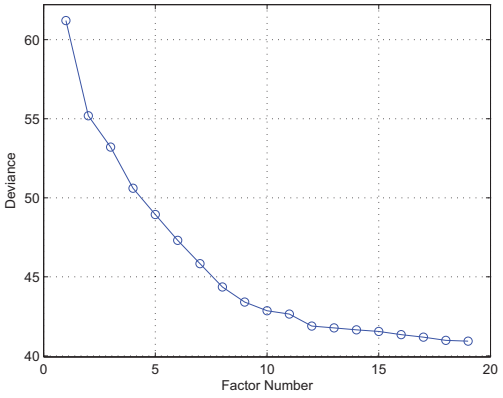

Fig. 3. Deviance reduction as additional factors are included in the model.

Video Name	Avg. number of TS / IP pkt in I Frame	Avg. number of TS / IP pkt in P Frame	Avg. number of TS / IP pkt in B Frame
Air show	319.8 / 45.6	134.2 / 19.2	70.9 / 10.1
Earth	276.8 / 39.5	89.6 / 12.8	41.7 / 6.0
Formula	233.9 / 33.4	90.7 /12.9	32.7 / 4.7
Golf	377.6 / 53.9	74.7 / 10.7	20.6 / 2.9
Hawaiian	383.7 / 54.8	78.0 / 11.1	29.8 / 4.3
Indianapolis	382.4 / 54.6	74.6 / 10.7	23.3 / 3.3
New York	369.5 / 52.8	71.9 / 10.3	21.0 / 3.0
Soccer	236.5 / 33.8	79.8 / 11.4	31.2 / 4.5
Stories	271.7 / 31.1	85.4 /12.2	26.3 / 3.8

Table 3. Table of the statistics for number of TS and IP packets in each video by frame type.

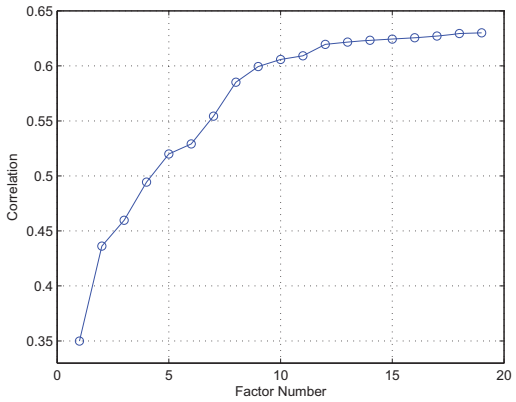

Fig. 4. Correlation between predicted and actual VQM score as additional factors are included in the model.

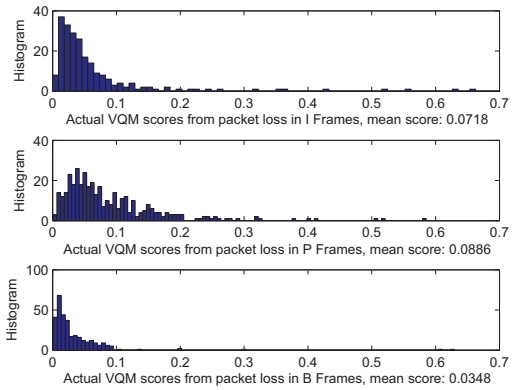

Fig. 5. Histogram of VQM scores from different frame types.

6. REFERENCES

[1] T-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. Cosman, A, Reibman. A Versatile Model for Packet Loss Visibility and its Application to Packet Prioritization. IEEE Transactions on Image Processing, Vol. 19, No. 3, pp. 722-735, March 2010.

[2] T.-L. Lin and P.C. Cosman. Efficient optimal RCPC code rate allocation with packet discarding for pre-encoded compressed video. IEEE Signal Processing Letters, Vol. 17, No. 5, pp. 505-508, May 2010.

[3] T.-L. Lin and P. Cosman. Network-based packet loss visibility model for SDTV and HDTV for H.264 videos. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010.

[4] S. Kanumuri, S.G. Subramanian, P.C. Cosman, A.R. Reibman, and V. Vaishampayan. Predicting H.264 Packet Loss Visibility using a Generalized Linear Model. in *ICIP*. IEEE, Oct. 2006, pp. 2245 – 2248.

[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] T.-L. Lin, J. Shin and P. Cosman. Packet dropping for widely varying bit reduction rates using a network-based packet loss visibility model. IEEE Data Compression Conference (DCC), pp.445-454, 2010.

[9] The website for VQM software: `http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm`.

[10] ISO/IEC standard 13818-1, also known as ITU-T Rec. H.222.0.

[11] The website for tsMuxeR software: http://www.smlabs.net/tsmuxer_en.html

[12] S. Wenger. H.264/AVC Over IP. *IEEE Transactions on Circuits and System for Video Technology*, July 2003.

[13] P. McCullagh and J. A. Nelder, Generalized Linear Models, 2nd ed. Chapman & Hall, 1989.

[14] D. W. Howmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. Wiley-Interscience, 2000.

[15] G. M. Mullet. Why Regression Coefficients Have the Wrong Sign. Journal of Quality Technology, vol. 8, No. 3, pp. 121-126, Jul. 1976.