Packet Loss Visibility of View+Depth Compressed Stereo 3D Video

Qing Song and Pamela Cosman
University of California, San Diego
Dept. of Electrical and Computer Engineering
9500 Gilman Dr, La Jolla, CA 92093-0407

***Abstract*—We conduct a human observer experiment on the visibility of fixed-sized packet losses in compressed 3D videos. The stereo videos are encoded using view+depth based on H.264/AVC. A prediction model of the visibility score is built with features extracted from the video. Results show good accuracy of the model.**

Keywords*—*view+depth 3D, human observer experiment, H.264/AVC, fixed-sized packet loss, error concealment.

I. Introduction

With the increasing popularity of 3D videos, the amount of data to be stored and transmitted greatly increases. This poses problems for fluctuating, band-limited wireless communications. One solution is view+depth coding, using one color view and its depth map instead of two color views for stereoscopic videos. At the decoder, the other view can be generated by depth image-based rendering (DIBR) [1]. The depth map includes the distance of objects in the scene from the camera but no information of texture, thus the amount of data can be reduced.

When videos are transmitted through a lossy channel, packets can be lost. Different packet losses can have different visual impacts. Hewage et al. found that the overall video quality is affected by both color view and depth packet losses, but prioritizing color view packets can vastly improve the overall video quality [2] [3]. Pinto et al. mentioned that losses in videos with low disparity is less annoying than in ones with high disparity [4]. However, they only evaluate the overall image quality and depth perception of the entire video sequence via PSNR or MOS (Mean Opinion Score). It remains unclear if an arbitrary color view packet has greater impact than any depth packet. For example, whether a P frame in the color view video is more important than an I frame in the depth map video was not addressed. It is also unclear if a packet located in a low disparity region should always cause less damage than a packet in a high disparity region. One may ask if other factors, such as the spatial location of the packet, can also affect the perceived quality. Simple objective metrics may not be satisfying to represent the complex 3D attributes. Therefore, we conduct a human observer experiment to measure the human perception of different types of packet loss. The video is encoded in view+depth format and packetized into fixed-sized packets. Based on the observer experiment, we build an encoder-based model to predict the visibility of packet losses. We employ logistic regression to build the prediction model. The predicted visibility can be used for forward error correction (FEC) and unequal error protection (UEP). Strong protection can be applied to the packets with high visibility by allocating more forward error correction to them.

Fig. 1. View+Depth Format

The paper is organized as follows: in Section II, we briefly overview the view+depth coding format. The observer experiment settings and results are explained in Section III. We describe the feature extraction, modeling approach and performance in Section IV. Section V concludes the paper.

II. View+Depth Coding Format

For stereoscopic 3D video, the view+depth format consists of the left color view and its depth map. The depth map is a 2D representation of the 3D surface. With the depth map, the 3D structure of the objects in the scene can be reconstructed. The depth map is a grayscale image, thus can be compressed in YUV 4:0:0 format (Fig. 1). It can be generated from the original left and right color view. In recent years, depth estimation has been extensively explored, and many of the algorithms are evaluated by the Middlebury Stereo Benchmark [5]. In our observer experiment, we employ the widely used Min-Cut algorithm [6] which is also used in MPEG Depth Estimation Reference Software [7]. The left color view and the depth map can be separately or jointly encoded.

At the decoder, the right view is synthesized from the decompressed left color view and depth map. If the two views are well rectified and parallel, the right view can be synthesized efficiently without a z-buffer [8]. The columns of the left image are warped from left to right image borders based on the 3D structure built from the depth information. The major problem is disocclusion. Some areas occluded in the left view can be visible in the right view. This results in holes in the synthesized right view. One solution is to fill the holes by spatial interpolation using neighboring pixels. Another one is preprocessing the depth map with a Gaussian filter so that much smaller disocclusion areas would appear in the right view [1]. Based on [9], the Gaussian filtering method yields the best visual quality. It only incurs a small geometric distortion

978-1-4799-2172-0/13/$31.00 ©2013 IEEE

but no visible flickers along the object edges. Therefore, in our observer experiment we also apply a 27×27 Gaussian filter with $\sigma = 9$ to preprocess the depth map before 3D warping. If disocclusion still remains after the filtering, the holes would be diminished to a very small area, and then we use the spatial interpolation method proposed in [10], which has similar performance to inpainting [11] but works more efficiently.

III. Human Observer Experiment

A. Setup of the Experiment

We conducted a human observer experiment in which the viewers were shown 3D videos with impairments caused by packet losses. The viewers were asked to press the space bar once they saw a glitch. To allow the viewers to have enough responding time, we insert at most one loss in every 4 seconds. The loss occurred in the first 3.2 seconds in each 4-sec interval, and the last 0.8 seconds would allow any error propagation to stop. The viewer was considered as having seen the loss if he/she responded within 1 second after the loss.

We encoded the left color view (denoted as *color* or *color video* below) and the depth map (denoted as *depth* or *depth video* below) separately, as we want to compare the impacts of losses in color and in depth. The encoder is H.264 JM 18.1 [12]. The color video is in YUV 4:2:0 format and the depth video is in 4:0:0 format. Quantization Parameter (QP) values of 26, 31, 36 and 41 for both color and depth are suggested in [13]. It was found that increasing the bitrate of depth can improve the quality of the synthesized right color view significantly [8]. Thus, we fix QP to 26 for both color and depth video. The test video is 21'20" long, including episodes from 3D films *Avengers* and *Harry Potter and the Deathly Hallows Part I*. The color video is HD (1920×1080), and has 30 frames per second. The depth video is downsampled by 2 in the horizontal and vertical direction, so the size of each depth frame is a quarter of a color frame, which is also suggested in [13]. The deblocking was turned off when depth was encoded. We use the hierarchical GOP structure and insert intra frames every 24 frames (0.8 sec per I frame, Fig. 2). There are 6 types of frames in the GOP: one type of I frame, two types of P frames and 3 types of B frames, classified by their time duration. The time duration, or the length of error propagation, is defined as the maximal number of frames affected by the error in one frame. For example, any loss in a P1 frame would affect itself, the next P2 frame and 21 B frames. The time duration of each type of frame is given in Table I. The video bitstream is divided into fixed-sized packets of 1316 bytes (equal to seven MPEG packets of 188 bytes in length), as recommended in [14]. Each packet includes at most one frame and would not include any information of the next frame. A packet would not split a macroblock either. So some packets could be less than 1316 bytes.

The decoder is JM 16.2. To conceal losses in color and depth I frames, we use spatial interpolation by taking the sum of weighted neighboring available pixels. To conceal losses in color P or B frames, we use motion-compensated error concealment (MCEC). The motion vectors of neighboring available (correctly decoded or concealed) macroblocks are extracted. The motion vector that minimizes the boundary

TABLE I. Maximal Number of Frames Affected

Frame Type	Time Duration
I	31
P1	23
P2	15
B1	7
B2	3
b	1

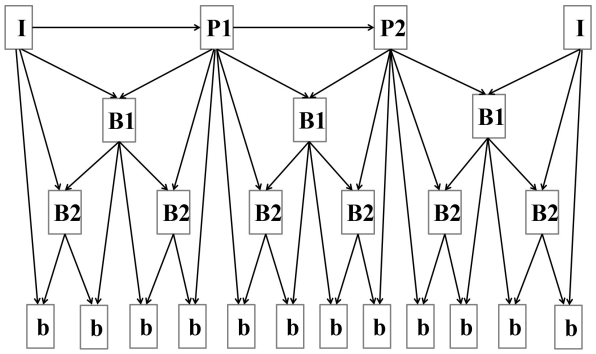


Fig. 2. Hierarchical GOP Structure

matching error [15] is taken to conceal the corrupted macroblock. If the neighboring macroblock is sub-partitioned, only the motion vectors of the blocks adjacent to the macroblock to-be-concealed are considered as candidates. For example, corrupted macroblocks are shown in dark gray in Fig. 3. To conceal MB #1, motion vectors of the blocks in light gray above and below MB #1 are considered as candidates when those macroblocks are sub-partitioned. Only the motion vectors of correctly decoded macroblocks are considered if they are available. The motion vectors of concealed macroblocks are considered only when none of the neighboring macroblocks is correctly decoded (the macroblock to be concealed is surrounded by other corrupted macroblocks). In Fig. 3, only the motion vectors of the blocks below MB #2 are candidates for MCEC. Though the macroblocks above MB #2 are concealed first, their motion vectors are considered unreliable so we do not use them. If all the neighboring macroblocks are intra-coded or the whole frame is lost, then no motion vector is available. In that case, we set the motion vector to zero, i.e., copying the co-located macroblock from the reference frame.

For losses in depth P or B frames, we conceal each macroblock by setting its motion vector as half of the average of the motion vectors extracted from the co-located macroblocks in the corresponding color frame, due to the high correlation between color and depth video. In our experiment, the co-located macroblocks in the color frame are always available (though their motion vectors may not exist), because there is at most one packet loss in every 4-sec interval, so if the loss is in a depth frame, then the corresponding color frame is intact. Since the depth maps are downsampled by 2 in each direction, the motion in the color frame can be twice the motion in the depth, and one macroblock in a depth frame corresponds to 4 macroblocks in the color frame. In Fig. 4, macroblocks in light gray shade in the color frame are extracted to conceal MB #3 in the depth frame. The efficiency of this method is shown in [16]. If all of the co-located color macroblocks are intra-coded, we simply set the motion vector of the corrupted

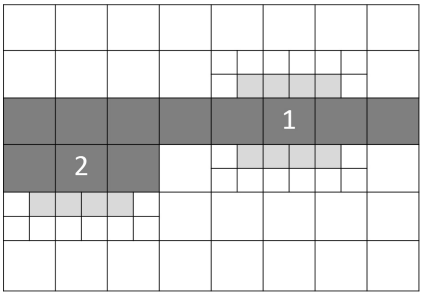
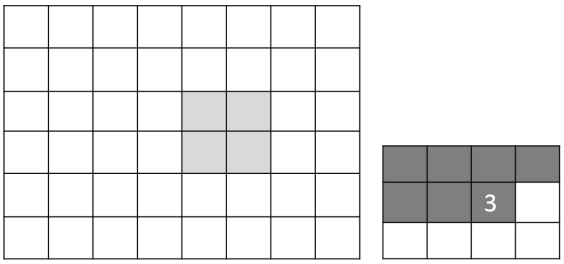


Fig. 3. Error concealment for color frame



(a) Color frame (b) Depth frame

Fig. 4. Error concealment for depth frame

depth macroblock to zero.

We generated 5 versions of the lossy video. Each version includes 300 packet losses. These losses are divided equally and randomly among each type (color and depth, each has 6 types). Each version of the lossy video was evaluated by 12 viewers. All of the viewers have normal or corrected-to-normal vision, and have good stereo vision (tested by the stereo fly test). Before the experiment, a 3-min pilot training video was shown so that the viewers could get a sense of the artifacts they were going to see. The lossy videos also include some intervals without any loss so that we can measure the false positive rate caused by factors other than the packet losses, such as view synthesis artifacts.

B. Experimental Results

We define the visibility score of each packet as the number of viewers who saw its loss divided by the total number of viewers who assessed that version of lossy video. Fig. 5 shows the mean visibility score of each type of loss. The visibility of losses in color frames is generally higher than losses in depth. One reason is that a color packet loss would affect the left color view itself and the right color view rendered from it. However, if a depth packet is lost, only the right color view would be affected. Another reason is that color packet losses usually cause blocky artifacts, which are probably more likely to be seen than the geometric distortion caused by depth losses (Fig. 6 and 7).

Among the color packet losses, losses in P frames are the most visible. One might have expected that losses in I frames should be the most damaging since they have the longest error propagation, and this in fact has been true in the case where packets hold a fixed number of macroblocks. However, as we fix the size in bytes of each packet, the spatial area affected by a packet loss in an I frame is usually less than the area affected in a P frame which is less than the area affected in a B frame. Table II shows the average number of packets included in each type of frame. One packet in a color I frame covers on average 1.97% of the spatial area of a frame, while a packet in a color P frame covers 2.55% area. So under the interaction of time duration and spatial area affected, losses in P frames have the highest visibility score. This is consistent with the previous work [17].

TABLE II. AVERAGE NUMBER OF PACKETS INCLUDED IN A FRAME

Video	I frame	P frame	B frame
Color	50.8	39.3	20.4
Depth	2.3	2.0	1.5

For the depth packet losses, it turns out that losses in I frames are the most prominent. While a depth I frame packet does cover slightly less spatial area than a depth P frame packet, the visibility scores for depth packets do not follow the same trend as color packets because the error concealment for depth is quite different. Losses in depth P and B frames can be concealed better than losses in I frames. Motion in the color frame and the depth frame is highly correlated. Besides, depth frames include very little texture, so the residual energy after motion compensation is usually small. Therefore, copying the motion vectors of the corresponding color frames is very helpful to recover the corrupted macroblocks. There are no motion vectors in I frames, and the spatial interpolation often yields an unsatisfying result when the corrupted area is large.

In the experiment, there are twenty 4-sec intervals without any loss in each version of the video. We collected the viewers' responses in those intervals to measure the false positive rate. False positive responses may be due to compression artifacts, view synthesis artifacts, or just inattention. The false positive rate is 4.17%, which is well below the mean visibility score of losses in all the color frame types and in depth I and P frames. However, the mean visibility scores of packet losses in depth B1, B2 and b frames are 0.0560, 0.0787 and 0.0333 respectively, which are close to the false positive rate. This suggests that some or most of the responses counted for these losses may not actually be due to the losses. It would be wrong to conclude however that all depth B frame losses can be assumed to be unimportant visually, because different packets of the same type sometimes have very different visibility scores. For example, some losses in depth b frames have visibility score as high as 0.75, though most losses in that frame type were not perceived by any viewers. The mean visibility score of losses in color P1 frames is 0.6787 and 30.4% of that type of losses were seen by all the viewers, but some other packets of that type have zero visibility score. So the mean value may not well represent the visibility score of each loss. Therefore, we aim to investigate the features of each packet and use them to predict the visibility score.

IV. VISIBILITY MODEL

Since the main use of the predictions of the visibility score would be for unequal error protection, we want to make the prediction at the encoder side. That means we have access to the original video, the compressed bitstream and the reconstructed video. To predict the visibility score, we extract features from the videos and bitstream, and then utilize

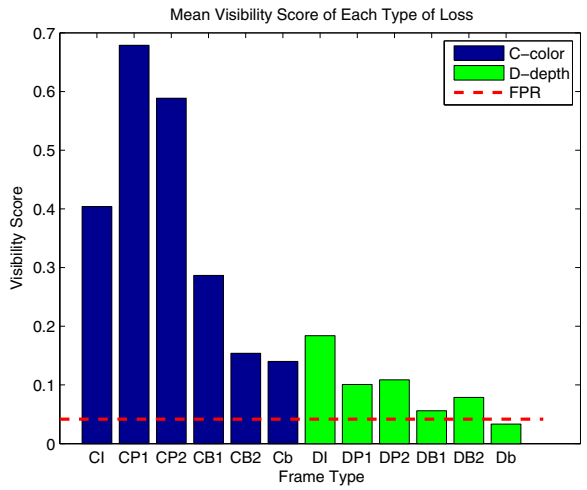


Fig. 5. Mean visibility score of each type of loss. The dash line shows the false positive rate, which is 0.0417.

(a) Left Color View

(b) Right Color View

Fig. 6. Packet loss in a color frame. The blocky artifacts appear on the woman's face in both views.

those features to build a visibility model. We first describe the features in this section, then explain the modeling approach and the results.

A. *Feature Extraction*

The extracted features are grouped into two categories: content independent features and content dependent features. The feature abbreviations and brief descriptions are given in Table III and IV.

Content independent features, such as the frame type determined by the GOP structure and the spatial location of

(a) Left Color View

(b) Right Color View

Fig. 7. Packet loss in a depth frame. The artifacts appear only in the right view around the man's leg.

TABLE III. CONTENT INDEPENDENT FEATURES

Feature	Abbreviation	Description
IsColor	IsColor	Packet is in color frame
Time Duration	TMDR	Maximal number of frames affected
Deviation from Border	DevFromBorder	floor(N/2) − |Height-floor(N/2)|, N is number of rows of macroblocks
Frame Type	IsCIframe	Packet is in color I frame
	IsCPframe	Packet is in color P frame
	IsCBframe	Packet is in color B frame
	IsDIframe	Packet is in depth I frame
	IsDPframe	Packet is in depth P frame
	IsDBframe	Packet is in depth B frame

the packet, do not depend on the content of the video. The following features are considered:

1) IsColor: a boolean factor which is set to 1 if the packet is in a color frame, and is set to 0 if it is in a depth frame.
2) Time Duration (TMDR): the maximal number of frames affected by the loss, which is completely determined by the type of frame that includes the packet. (Table I)
3) Deviation from Border (DevFromBorder) = floor(N/2) − |Height-floor(N/2)|, where Height is the vertical location of the packet center, N is the number of rows of macroblocks in one frame. N = 68 in this experiment since we use HD video.
4) IsCIframe, IsCPframe, IsCBframe, IsDIframe, IsDPframe and IsDBframe are boolean factors denoting the frame type. IsCPframe means the packet is in a color P frame. We do not specify P1 and P2 so that the prediction model can be used for other GOP structures and I frame periods.

Content dependent features are those related to the content of the video, such as motion complexity. We extract some of them from both color and depth videos. If the lost packet is in a color frame, the features of the macroblocks contained in the packet and features of the co-located depth macroblocks

TABLE IV. CONTENT DEPENDENT FEATURES

Feature	Abbreviation	Description
Number of Macroblocks	NumMB	Number of macroblocks(MBs) in packet if IsColor = 1, 4 times number of MBs in packet if IsColor = 0
Packet Size	PktSize	Number of bytes in packet
Number of Macroblocks Coded in a Certain Mode	CNumIntra	Number of color MBs in affected area which are intra coded
	CNumInter	As above, inter coded
	CNum(Skip/Direct)	As above, skip or direct coded
	DNumIntra	Number of depth MBs in affected area which are intra coded
	DNumInter	As above, inter coded
	DNum(Skip/Direct)	As above, skip or direct coded
Ratio of Macroblocks Coded in a Certain Mode	CIntraRatio	CNumIntra / NumMB
	CInterRatio	CNumInter / NumMB
	C(Skip/Direct)Ratio	CNum(Skip/Direct) / NumMB
	DIntraRatio	DNumIntra / (NumMB / 4)
	DInterRatio	DNumInter / (NumMB / 4)
	D(Skip/Direct)Ratio	DNum(Skip/Direct) / (NumMB / 4)
Maximal Sub-partitions	CMaxInterparts	Maximal sub-partitions in affected color MBs
	DMaxInterparts	Maximal sub-partitions in affected depth MBs
Motion Vector	CMaxMotX, CMeanMotX, CVarMotX	Maximum of absolute value, mean and variance of horizontal motion vectors(MVs) of affected color MBs
	CMaxMotY, CMeanMotY, CVarMotY	Maximum of absolute value, mean and variance of vertical MVs of affected color MBs
	CMaxMotM, CMeanMotM, CVarMotM	Maximum, mean and variance of MV magnitude of affected color MBs
	CMaxMotA, CMeanMotA, CVarMotA	Maximum of absolute value, mean and variance of motion direction of affected color MBs
	DMaxMotX, DMeanMotX, DVarMotX	Maximum of absolute value, mean and variance of horizontal MVs of affected depth MBs
	DMaxMotY, DMeanMotY, DVarMotY	Maximum of absolute value, mean and variance of vertical MVs of affected depth MBs
	DMaxMotM, DMeanMotM, DVarMotM	Maximum, mean and variance of MV magnitude of affected depth MBs
	DMaxMotA, DMeanMotA, DVarMotA	Maximum of absolute value, mean and variance of motion direction of affected depth MBs
Residual Energy	CMaxRSENGY	Maximum of residual energy of affected color MBs after motion compensation
	DVarRSENGY	Variance of residual energy of affected depth MBs after motion compensation
MSE	MaxMSE, MeanMSE, VarMSE	Maximum, mean and variance of MSE per MB
SSIM	MinSSIM, MeanSSIM, VarSSIM	Minimum, mean and variance of SSIM per MB
Foreground Macroblocks	FGNum	Number of foreground MBs in packet
	FGRatio	FGNum / NumMB

are extracted. Likewise for a loss in a depth frame, we extract information from itself and the co-located color macroblocks.

1) Number of macroblocks affected by the packet loss (NumMB). It denotes the area in the frame affected by the loss. For packets in color frames, NumMB is the number of macroblocks in the packet. For packets in depth, NumMB equals 4 times the number of macroblocks in the lost packet, since one macroblock in the depth map corresponds to 4 macroblocks in the right color view synthesized from it. This feature relates to the frame type, the spatial correlation and the motion complexity. For example, in a P or B frame, if the motion is complicated, the residual energy after motion compensation would be high, then more bits would be allocated to code the macroblocks and a packet would include fewer macroblocks than would one which contains macroblocks from a low motion frame.
2) Packet Size (PktSize). Since each packet contains at most one frame, some packets could be less than 1316 bytes. Most of the values of PktSize are around 1316 as we fix the length of the packet. In the following two scenarios, the packet can be much less than 1316 bytes: (1) the whole frame is included in one packet and (2) the packet is the last one in that frame. So this feature may relate to spatial location and motion complexity. In the videos we use in this experiment, only a small number of color B frames, and some depth P and depth B frames are packetized into one packet, as the videos are HDTV.
3) Number of macroblocks coded in intra, inter and skip/direct mode (CNumIntra, CNumInter, CNum(Skip/Direct), DNumIntra, DNumInter and DNum(Skip/Direct)). Once we get the location of a lost color packet, we extract the mode of macroblocks in the lost color packet and the mode of co-located macroblocks in the depth frame. Similarly, for a depth packet loss, we extract the mode of macroblocks in the packet and the mode of the co-located color macroblocks. CNumIntra denotes the number of macroblocks located in the affected area in the color frame which are coded in intra mode; and DNum(Skip/Direct) denotes the number of macroblocks in the affected area in the depth frame which are coded in skip or direct mode.
4) Ratio of macroblocks coded in intra, inter and skip/direct mode (CIntraRatio, CInterRatio, C(Skip/Direct)Ratio, DIntraRatio, DInterRatio and D(Skip/Direct)Ratio) is the number of macroblocks coded in that mode divided by the number of macroblocks in the affected area. These features relate to the motion of the affected area. For example, if the packet is in a P frame and the IntraRatio is very high, that probably means the motion is complicated and the error could be hard to conceal.
5) CMaxInterparts and DMaxInterparts are the maximal number of sub-partitions in the color and depth macroblocks lying in the affected area respectively. If the MaxInterparts is large, it probably also implies complicated motion.
6) MotX and MotY are the motion vector components in the horizontal and vertical directions. MotM is the magnitude of the motion vector ($MotM = \sqrt{MotX^2 + MotY^2}$). MotA is the direction of the motion ($MotA = \arctan(MotY/MotX)$). We compute the maximum of the absolute value, mean, and variance of MotX, MotY, MotM and MotA of both color and depth macroblocks in the affected area. If all the macroblocks in the area are coded in intra mode, then all those values are set to 0. We use DMeanMotM to denote the mean value of the motion vector magnitude of the affected depth macroblocks.
7) RSENGY is the residual energy per pixel after motion compensation of the macroblock. We compute the maximum of the residual energy of color macroblocks (CMaxRSENGY) and the variance of the residual energy of depth macroblocks (DVarRSENGY) in the affected area. The residual energy of depth macroblocks is usually small as they include little texture. But it can have a large value when the object is moving in the z direction. Its

variance over the affected macroblocks can also relate to the motion complexity.

8) For each packet loss, MSE and SSIM (Structural Similarity Index) [18] per macroblock are computed between the compressed (error-free) video and the decompressed video with that one packet loss. We do not compute those values between the original raw video and the decompressed video with the packet loss because we are only interested in the quality degradation caused by the packet loss, not by compression artifacts. We compute only the initial error caused by the packet loss within the frame where the loss occurs, instead of computing cumulative error over all the frames affected. This helps to reduce computational complexity. We then take the maximum, mean and variance of MSE per MB (MaxMSE, MeanMSE, VarMSE), and minimum, mean and variance of SSIM per MB (MinSSIM, MeanSSIM, VarSSIM). A large value of MSE and a small value of SSIM indicate large degradation in quality.
9) Viewers are usually attracted by foreground objects which may have different motion from the background. The cameras may also focus on those objects so the background may be blurry. So errors in the background can usually be concealed better than errors in the foreground. If most of the affected area is background, it may be less likely for the viewers to notice the packet loss. With depth maps, it is easy to extract foreground pixels from the frame. Pixels with depth deeper than some threshold are considered as background. To find a good threshold, we first plot the histogram of the depth values in that frame. Then we pick the minimum between the two non-neighboring highest bins as the threshold. In each macroblock, if over half of the pixels are foreground, we consider it as a foreground macroblock. FGNum is the number of foreground macroblocks in the packet. FGRatio is the portion of foreground macroblocks in the affected area, which is equal to FGNum divided by the total number of macroblocks in the packet.

B. Modeling Approach

We employ the generalized linear model (GLM) with logit as the link function for binomial distribution to build the prediction model. The inputs of the model are the features of a packet, and the output is the prediction of the packet's visibility score. The model is

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

where p is the visibility score, x_j is a feature, β_j is its coefficient, and γ is a constant term.

The whole dataset includes 1500 samples. We use 5-fold cross validation to select the most important features and prevent overfitting. The whole data is divided into 5 partitions, 4 of which are used to train the model and the one left out is used to test the performance. The procedure is repeated 5 times. Different partitions are used as test data each time. The optimal feature is selected in each step to minimize the mean squared error (MSE) between the predicted visibility score and the ground truth. A fixed set of features is used to train only one model.

C. Performance

We use mean squared error (MSE) and correlation coefficient to measure the performance of the model. We compute the two metrics between the prediction and the ground truth via 5-fold cross validation. Fig. 8 shows the performance vs. the number of features added into the model. The correlation coefficient reaches 0.75 when 30 features are added into the model, 0.72 with 20 features, 0.70 with 10 features and 0.67 with only 5 features.

Table V shows the ten most important features in the prediction model, where $\times$ means multiplication of the two single features. IsColor plays a key role since color packet losses are generally more visible than depth packet losses. Three out of the top ten features relate to IsColor, and their coefficients all have positive signs. The most important one is IsColor $\times$ CIntraRatio. It indicates that if the packet is in a color frame, and more macroblocks are coded in intra mode, the packet is more likely to be seen. That is because those corrupted macroblocks are not likely to be concealed well.

The spatial location of the packet is also critical to the visibility. Viewers are usually attracted by the objects at the center of the screen, both because the camera location is often chosen to place interesting objects at the center, and also because the large screen sizes of HDTV mean the viewer is often less aware of the periphery. A large value of IsColor $\times$ DevFromBorder means the loss affects both views and appears near the center.

The objective quality metrics are also helpful. TMDR $\times$ MaxMSE is the second most important feature. It implies that a big distortion which lasts for a long time is very likely to be seen. The feature with MinSSIM carries a negative coefficient, as smaller value of SSIM indicates worse quality.

The frame type is another important factor in the model. The features related to IsCBframe and IsDBframe have negative coefficients as would be expected, since losses in B frames are less visible than average losses. IsCPframe $\times$ CNumIntra has a positive impact on the visibility. A large value of intra-coded macroblocks implies the motion is complicated or there is a scene cut. Then zero motion copy would probably not yield a good result.

The single term PktSize carries a positive sign. Most of the values of PktSize are around 1316. The packet size can be well below 1316 bytes if the whole frame is included in the packet or if the packet is the last one in that frame. In the first scenario, it may imply the residual energy is small and motion is not very complicated. In the second scenario, it means the loss is far away from the center, thus less visible.

V. Conclusion

We present a human observer experiment on fixed-sized packet loss visibility of view+depth compressed 3D video. Losses in color frames are generally more likely to be seen than losses in depth frames, probably due to the different types of artifacts they cause and the number of views affected by the loss. Losses in color P frames are the most damaging, even worse than losses in color I frames. Among losses in depth frames, I frames are the most difficult to conceal thus are the most visible. We build an encoder-based model to predict the

TABLE V. THE TEN MOST IMPORTANT FEATURES OF THE PREDICTION MODEL

Feature #	Feature	Coefficient
γ	1	-3.2515
1	IsColor × CIntraRatio	0.0328
2	TMDR × MaxMSE	2.3362e-6
3	IsColor × DevFromBorder	0.0634
4	IsCBframe × CMaxMotA	-0.5752
5	IsColor × CMaxMotM	0.0047
6	IsCPframe × CNumIntra	0.0031
7	D(Skip/Direct)Ratio × MinSSIM	-1.7107
8	PktSize	0.0012
9	DInterRatio × DVarMotA	-7.1154
10	IsDBframe × DMaxMotX	-0.0113

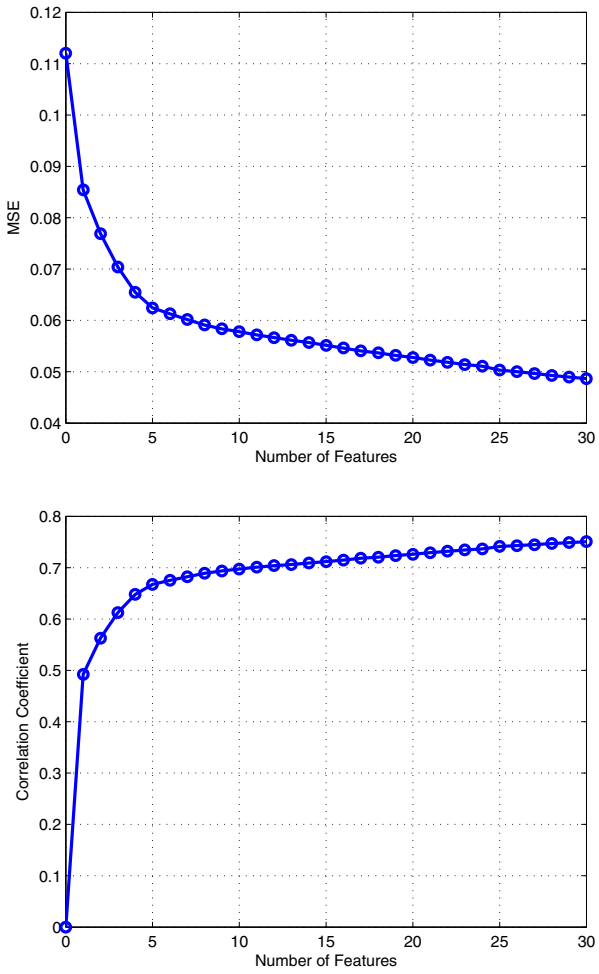


Fig. 8. Performance of the Prediction Model

visibility of packet losses with features related to frame type, spatial location of the loss and motion complexity. The model shows good performance in terms of MSE and correlation coefficient.

ACKNOWLEDGMENT

This research was supported in part by the Intel/Cisco Video Aware Wireless Network (VAWN) program, and by the National Science Foundation under grant CCF-1160832.

REFERENCES

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, 2004, pp. 93–104.

[2] C. T. E. R. Hewage, S. Worrall, S. Dogan, H. Kodikaraarachchi, and A. Kondoz, "Stereoscopic TV over IP," in *4th European Conference on Visual Media Production*, 2007, pp. 1–7.

[3] C. T. E. R. Hewage, S. Worrall, S. Dogan, S. Villette, and A. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304–318, 2009.

[4] L. Pinto, J. Carreira, S. Faria, N. Rodrigues, and P. Assuncao, "Subjective quality factors in packet 3D video," in *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 149–154.

[5] D. Scharstein and R. Szeliski. Middlebury stereo evaluation. [Online]. Available: http://vision.middlebury.edu/stereo/

[6] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[7] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, *Reference Softwares for Depth Estimation and View Synthesis*, ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15377, April 2008.

[8] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *IEEE International Conference on Image Processing*, vol. 1, 2007, pp. I – 201–I – 204.

[9] E. Bosc, R. Pépion, P. Le Callet, M. Köppel, P. Ndjiki-Nya, L. Morin, and M. Pressigout, "Perceived quality of dibr-based synthesized views," in *SPIE Applications of Digital Image Processing XXXIV*, vol. 8135, 2011.

[10] A. Jain, L. Tran, R. Khoshabeh, and T. Nguyen, "Efficient stereo-to-multiview synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 889–892.

[11] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004.

[12] H.264/AVC JM reference software. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[13] *Common Test Conditions for 3DV experimentation*, ISO/IEC JTC1/SC29/WG11 MPEG 2011/N12745, May 2012.

[14] *DSL Forum Technical Report TR-126: Triple-play Services Quality of Experience (QoE) Requirements*, December 2006.

[15] W.-M. Lam, A. Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 1993, pp. 417–420.

[16] Y. Liu, J. Wang, and H. Zhang, "Depth image-based temporal error concealment for 3-D video transmission," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 600–604, 2010.

[17] Y.-L. Chang, T.-L. Lin, and P. Cosman, "Network-based IP packet loss importance model for H.264 SD videos," in *IEEE 18th International Packet Video Workshop (PV)*, 2010, pp. 178–185.

[18] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.