FRAME LOSS VISIBILITY MODELING OF STEREOSCOPIC VIDEO FOR H.264/AVC-MVC

Arash Vosoughi and Pamela C. Cosman

ECE, University of California, San Diego

ABSTRACT

We develop a visibility model which can predict the visibility of frame losses in compressed 3D video. The 3D video is encoded using the MVC (multiview coding) extension of the H.264/AVC standard. Frame losses both in the left view (base view) and right view (enhancement view) of the stereoscopic video are considered. A subjective test is conducted to identify which types of frame losses are perceptually noticeable. Several features are extracted from the encoded frames, and then support vector machines are employed to build visibility models based on these features. Results show that our model can predict the visibility of frame losses in stereoscopic video with good accuracy.

***Index Terms*—** Stereoscopic video, multi view coding, H.264/AVC, packet loss, subjective testing, error concealment, support vector machine.

1. INTRODUCTION

Three-dimensional video has become very popular in recent years. 3D video is expected to have important applications over IP networks such as 3D TV on demand, 3D TV broadcasting, and 3D video conferencing. Transmission of compressed 3D video over IP networks is inevitably subject to packet losses, and thus, characterizing the impact of packet losses on the quality of the video is important. Recent research has introduced objective metrics for evaluating the quality of 3D video (for example see [1], [2], [3], and [4]). Developing an objective quality metric, which incorporates the complex perceptual attributes of 3D such as depth, overall image quality, presence, naturalness, and visual comfort, is a major challenge. Thus, subjective tests are needed to understand how the human visual system perceives different kinds of frame losses in 3D stereoscopic video.

In this paper, our goal is to develop a robust predictor, which can predict the visibility of the lost frames in 3D stereoscopic video using information extracted from the encoded video, where the 3D video is encoded by an MVC [5] encoder. We make use of support vector machines (SVMs) to build this predictor. We conduct an experimental subjective test to collect data for training the SVM-based visibility model. Similar visibility models for two-dimensional video were introduced in [6] and [7], where comparable subjective tests are designed as well.

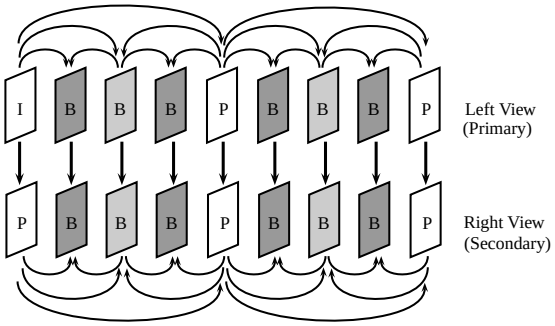

Fig. 1. Typical MVC structure for stereoscopic video.

The organization of this paper is as follows: in Section 2, we briefly overview multiview coding. In Section 3, we describe the design of the subjective test. In Section 4 we describe feature extraction and the modeling process, and also present the results.

2. MULTI VIEW CODING (MVC)

Fig. 1 shows a typical prediction structure of MVC used to encode a stereo video. Arrows indicate which frames are used as the reference frames for predictive encoding of the other frames. Frames of the left view are coded with a typical hierarchical B frame GOP (group of pictures) structure as provided by H.264/AVC: I frames are coded without reference to any other pictures, and temporal prediction and bipredictionare used for encoding the P frames and B frames, respectively. Temporal prediction is also used for predictive encoding of the frames of the right view. However, to improve the compression efficiency, MVC also exploits the inherent similarities between the pictures of the left view and right view by enabling interview prediction in which the pictures of the left view are used as reference pictures for encoding the frames of the right view.

3. SUBJECTIVE TEST

We conducted a subjective test in which the observers responded to a number of impairments they saw in a 3D video.

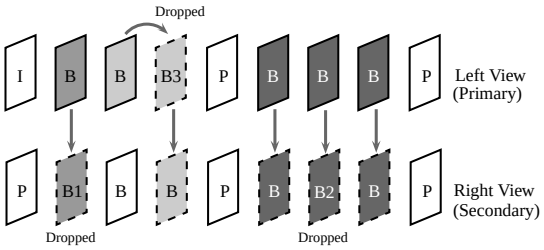

Fig. 2. Error concealment for three types of B frames.

These impairments were caused by dropped frames injected into the original tested video, and viewers responded to each impairment by clicking the space-bar of a keyboard. Frames were dropped randomly in time such that the minimum, maximum, and average time duration between two successive drops are 4, 8, and 6 seconds, respectively. 50 people participated in the test. The total number of 840 impairments were assessed by the viewers, where each impairment was assessed by 10 viewers. A pilot training 3D video was displayed to the viewers to train them about what kinds of artifacts they could expect to see in the test and how they should respond to them. The video sequences contained various types of motion (low, high, and medium motion scenes) and textures. Video contents were extracted from some 3D movies such as "Avatar" and "Priest"; they had the resolution of 1920×1080 pixels and had 23.976 frames per second. The length of the tested video for this experiment was 17 minutes. The test was conducted in a well lit room using a 47" LG 3D television. The observers were students with ages ranging from 21 to 31. All participants had normal or corrected to normal vision, and they also had good stereo vision (as tested with the fly stereo test).

We used the JMVC 8.2 (Joint Multiview Video Coding) reference software for encoding the tested stereo video. The GOP structure is IBBBPBBBP…, and the GOP size is 16. Three different types of frames are lost in the tested stereo video: (1) B frames in the right view which are not used for temporal prediction (we name these frames $B1$, see Fig. 2), (2) B frames in the right view which are used for temporal prediction (we name these frames $B2$), (3) B frames in the left view which are not used for temporal prediction but are used for interview prediction (we name these frames $B3$). The videos were compressed at very high bit rate, allowing essentially lossless compression, so that the viewers see frame drop and concealment artifacts in the absence of any compression artifacts.

Errors due to frame losses are usually concealed by an error concealment method before the lossy stereo video is displayed on the TV screen. In our subjective test, we employed an error concealment approach similar to the one suggested in [8]. Fig. 2 shows how this error concealment approach functions on the three different types of dropped frames we considered in the subjective test. According to Fig. 2, when a $B1$ frame is dropped in the right view, it is concealed by frame copying from its corresponding frame in the left view. This equates to switching to a 2D view, since the images of the left and right views become the same. On the other hand, when a $B2$ frame is dropped in the right view, the two neighboring frames, which are subject to error propagation due to the dropped $B2$ frame, are also concealed by frame copying from their corresponding frames in the left view. According to Fig. 2, a $B3$ frame is first concealed by frame copying from its reference B frame in the left view, and then the error propagated to its corresponding B frame in the right view is concealed by switching to the 2D view.

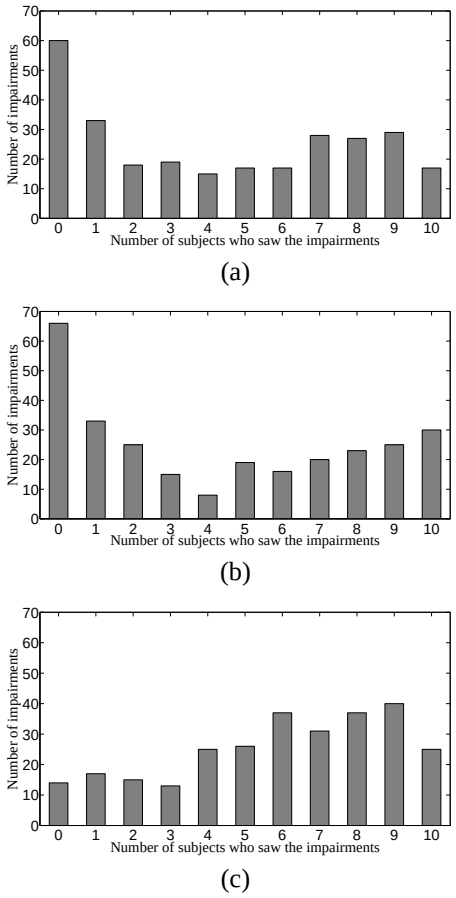

Fig. 3. Histograms of the number of impairments seen by the viewers. (a) Secondary view, non-reference B frames ($B1$), (b) Secondary view, reference B frames ($B2$), (c) Primary view, B frames not used for temporal prediction ($B3$).

Figures 3 (a), (b), and (c) show the histograms of the number of impairments observed by the viewers for the three frame types $B1$, $B2$, and $B3$, respectively. We assign a visibility score to each frame that is defined as the fraction of subjects who saw the impairment caused by dropping that frame. The profiles of visibility scores for $B1$ and $B2$ frame losses

are very similar. This is mainly because no frame copying (i.e., frame freezing) is performed in the primary view if a $B1$ or $B2$ frame is dropped in the secondary view, and thus, motion is preserved after error concealment even though the 3D view is switched to the 2D view. In contrast, we see in Fig. 3 (c) that the visibility of impairments is considerably higher for $B3$ frames compared to the $B1$ and $B2$ frames. This may be because when a $B3$ frame is dropped it generates two different types of artifacts simultaneously, i.e., frame freezing in the left view and switching to the 2D view.

4. VISIBILITY MODELING

4.1. Feature Extraction

For each frame we extract several features, which are used for predicting the visibility of that frame if it is dropped. Some of these features (which include some description of motion) are extracted from the encoded frame in the left view, and some of them (which include some description of disparity between the views) are extracted from the encoded frame in the right view. Features obtained from the encoded frame in the left view include the number of macroblocks (MBs) which are signaled as intra ($NumIntraMode$), skip ($NumSkipMode$), or direct ($NumDirectMode$) mode, sum of the average of the magnitude of the motion vectors in forward and backward directions ($AvgMagMV_{BW}$ + $AvgMagMV_{FW}$), X and Y components of the average of the motion vectors in forward and backward directions ($AvgMVX_{BW}$, $AvgMVY_{BW}$, $AvgMVX_{FW}$, $AvgMVY_{FW}$), and sum of the variances of the X and Y components of the motion vectors in forward and backward directions ($VarMVX_{BW}$ + $VarMVX_{FW}$, $VarMVY_{BW}$ + $VarMVY_{FW}$). Features obtained from the encoded frame in the right view include the average of the magnitude of disparity vectors ($AvgMagDV$), X and Y components of the average of the disparity vectors ($AvgDVX$, $AvgDVY$), the total area of the right frame which is predicted using interview prediction ($AreaInterViewPred$), the total area of the right frame which is predicted using temporal prediction ($AreaTempPred$), the total area of the right frame which is predicted using both the temporal and interview prediction ($AreaTempInterViewPred$), and variances of the X and Y components of the disparity vectors ($VarDVX$, $VarDVY$). A categorical feature ($FrameType$) is also considered, which takes three different values corresponding to the three frame types $B1$, $B2$, and $B3$.

4.2. Visibility model results

The proposed visibility model gets the vector of features extracted from the encoded frame as the input, and provides a predicted visibility score at the output. We employ SVMs to solve this regression problem, where the data collected in the subjective test is used for training and testing the SVM. We use a radial basis function (RBF) kernel for training the SVM. By selecting the RBF kernel, we must find only two parameters, C and γ [9]. We performed a grid search approach using 5-fold cross validation to find the best values of C and γ. Since a complete grid search is time-consuming, we followed the approach recommended in [9], where we first performed a coarse grid search to find a better region in the grid, and then a finer grid search on that region. Cross validation is utilized so that the model does not overfit the training data. We used LIBSVM software for our simulations [10].

Table 1. The most important features in terms of the MSE.

1	$NumIntraMode$
2	$AvgMagMV_{BW}$ + $AvgMagMV_{FW}$
3	$FrameType$
4	$VarDVX$
5	$NumDirectMode$
6	$AvgDVX$
7	$NumSkipMode$
8	$AreaTempPred$
9	$AvgMVX_{BW}$
10	$VarDVY$
11	$AvgDVY$
12	$VarMVY_{BW}$ + $VarMVY_{FW}$

We use two different metrics to measure the accuracy of the proposed visibility model: mean square error (MSE) and correlation coefficient. Both of these metrics are calculated between the true visibility scores of the data obtained in the subjective test, and the corresponding visibility scores predicted by the model. We obtain the MSE and correlation coefficient via 5-fold cross validation over the data.

Table 1 shows the list of the most important features. In this table, features are ranked in descending order of importance; they provide the largest drop in MSE when they are added to the model one after the other. Fig. 4 (a) illustrates the model accuracy in terms of the MSE versus the number of features utilized to train the model. It is observed that the MSE decreases from 0.12 (for the case where no features are used) to 0.03 (for the case where 12 features are exploited). We also observe that increasing the number of features more than 12 does not improve the accuracy of the model. Fig. 4 (b) shows the accuracy of the proposed visibility model in terms of the correlation coefficient metric. We observe that the correlation coefficient reaches 0.86 when 12 features are used. The MSE and correlation coefficient values confirm that the proposed visibility model performs well in predicting the true visibility scores obtained in the subjective test.

5. CONCLUSIONS

We considered the problem of predicting the visibility of frame losses in 3D stereoscopic video when it is coded using

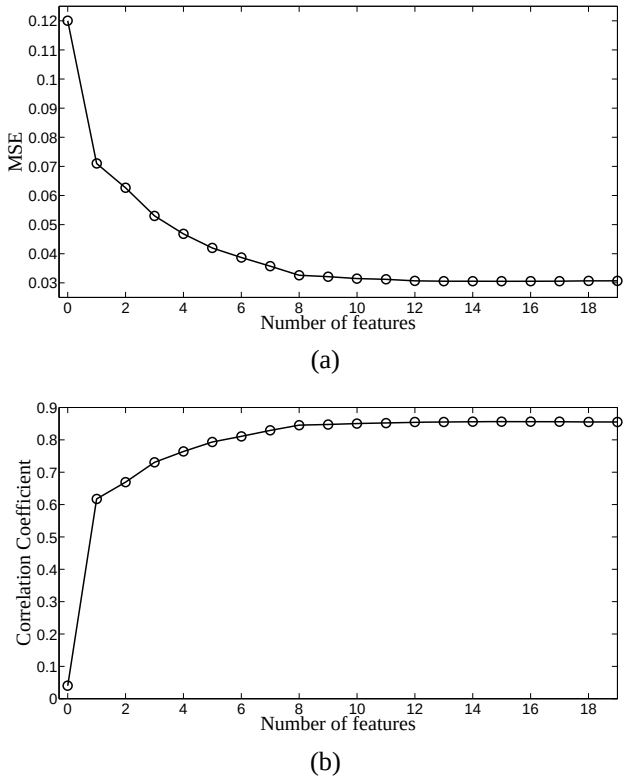

Fig. 4. (a) MSE between the true visibility scores and predictions of the model, (b) correlation coefficient between the true visibility scores and predictions of the model.

the H.264/AVC-MVC encoder. At first, we set up a subjective test to see which types of lost frames are perceived by people. Statistical results of our subjective test showed that the visibility profiles of $B1$ and $B2$ frames, which are non-reference and reference B frames in the secondary view, are interestingly very similar. Results also indicated that $B3$ frames (which are B frames in the primary view used only for inter-view prediction) are more visible than $B1$ and $B2$ frames. We extracted several simple features from the encoded frames based on the syntax elements of the H.264/AVC-MVC (such as the motion/disparity vectors, prediction mode of MBs, and etc.) for visibility prediction. We then developed a visibility model by means of SVMs using the data obtained in the subjective test. Results show that the developed visibility model is able to predict the visibility of the lost frames based on the simple exploited features with good accuracy in terms of both the MSE and correlation coefficient metrics.

The visibility model has two important potential applications in video transmission: intelligent packet dropping and unequal error protection. When an intermediate router in the network becomes congested, intelligent packet dropping can be better than random packet dropping. Once we have a visibility model which can predict the visibility of dropped frames (which are considered as packets), an intelligent strategy for frame dropping can be used such that the minimal video quality degradation is generated by the router. Unequal error protection (UEP) means that stronger protection is applied to the more important packets by allocating more FEC (forward error correction) to them. Our visibility model can also be applied for UEP, since it can predict the importance of the frames in terms of their visibilities.

6. REFERENCES

[1] H.-T. Quan, P. Le Callet, and M. Barkowsky, "Video quality assessment: from 2D to 3D- challenges and future trends," in *Int. Conf. on Image Proc. (ICIP)*. IEEE, Sept. 2010, pp. 4025–4028.

[2] J. Starch, J. Kilner, and A. Hilton, "Objective quality assessment in free-viewpoint video production," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2008, pp. 225–228.

[3] Y. Zhang, P. An, Y. Wu, and Z. Zhang, "A multiview video quality assessment method based on disparity and SSIM," in *Int. Conf. Signal Processing (ICSP)*, Oct. 2010, pp. 1044–1047.

[4] Z. Zhu, Y. Wang, Y. Bai, and Q. Shi, "New metric for stereo video quality assessment," in *Symposium Photonics and Optoelectronics (SOPO)*, Aug. 2009, pp. 1–4.

[5] A. Vetro, T. Wiegand, and G.J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," IEEE, Apr. 2011, pp. 626–642.

[6] T.-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. Cosman, and A. Reibman, "A versatile model for packet loss visibility and its application to packet prioritization," *IEEE Trans. on Image Proc.*, vol. 19, no. 3, pp. 722–735, Mar. 2010.

[7] T.-L. Lin, Y.-L. Chang, and P. Cosman, "Subjective experiment and modeling of whole frame packet loss visibility for H.264," in *Packet Video Workshop*. IEEE, Dec. 2010, pp. 186–192.

[8] M. Barkowsky et al., "Subjective quality assessment of error concealment strategies for 3DTV in the presence of asymmetric transmission errors," in *Int. Packet Video Workshop*, Dec. 2010, pp. 193–200.

[9] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classication*, Department of Computer Science, National Taiwan University, Apr. 2010.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm, official web site of LIBSVM software.