SELECTIVE FREEZING OF IMPAIRED VIDEO FRAMES USING LOW-RATE SHIFT-INVARIANT HINT

Mina Makar[1,2] *and Wai-Tian Tan*[2]

[1]Department of Electrical Engineering, Stanford University
[2]Mobile and Immersive Experience Laboratory, HP Labs

ABSTRACT

Irrecoverable data loss may be unavoidable for real-time video communication over common best-effort networks. Rather than always display impaired pictures or always "freeze" the last good picture, it is preferable to transmit additional hints to support selective freezing of heavily damaged pictures only. In particular, errors in impaired pictures tend to be localized, and often manifest themselves as small spatial shifts that are visually preferable to freezes. Therefore, it is essential that the hints can identify localized error and do not penalize small shifts. We show two ways such "shift-invariant" hints for detecting concealment error can be constructed to less than 1% of video rate. Experiments using 720p sequences achieve recall and precision of 90% and 75%, respectively, with respect to a shift-invariant PSNR measure. We also present an adaptive decision rule to obtain shorter and less frequent freezes.

Index Terms— video streaming, video quality monitoring

1. INTRODUCTION

Picture freezes and breakups are two notorious yet distinct artifacts of streaming video. After best-effort loss recovery and concealment have been performed, an important task for a receiver is to decide whether an impaired picture should be displayed. Many impaired pictures are preferable to freezes, yet some may be highly objectionable. It is therefore important to have methods to distinguish the acceptable pictures from the objectionable ones.

Determining visual quality without other hints is challenging. In contrast, there are existing reduced-reference methods that use low-rate hints instead of source pictures to determine received video quality. In particular, [1] proposed a low-rate method based on spread-spectrum ideas that can accurately approximate the PSNR for video degraded with both compression and losses. Parts of the techniques in [1] are adopted into ITU J.240 standard [2] for video quality monitoring. Nevertheless, it is well known that PSNR does not accurately capture visual quality. In particular, even well error-concealed pictures that are otherwise visually pleasing may contain slight perturbations that yield low PSNR, especially when contrast is large. Therefore, it is essential to develop low-rate hints that are invariant or insensitive to small perturbations or shifts to properly guide picture freezing decisions.

Shift-invariant metrics have been studied [3]. In particular, the SSIM scheme admits both an shift-invariant extension [3] and a reduced-reference implementation [4]. Nevertheless, as we will discuss later, this scheme is designed for spread errors and fails to recognize localized errors common in error-concealed pictures.

In this paper, we develop and compare two shift-invariant hints that work well for localized errors, and characterize their performance against PSNR. The rest of the paper is organized as follows. Section 2 provides a summary of our design. We first introduce a suitable target metric that is shift-invariant, and present two reduced-reference methods for its estimation. We then describe how picture freeze decisions can be adapted over time. This is followed by experimental results in Section 3, followed by a conclusion.

2. SHIFT-INVARIANT QUALITY ESTIMATION

Fig. 1 shows two error-concealed pictures when compressed H.264 video is subjected to loss. It is visually obvious that the top picture suffers from significant breakup and probably should not be shown. In contrast, the bottom picture only suffers minor visual distortion, and should be preferentially displayed. One possible way to establish preference for the bottom picture is to employ a shift-invariant metric. This is because an acceptable concealed picture typically differs from its loss-free counter-part via minor perturbations that can be approximated by small shifts. We next define a shift-invariant metric to serve as target for evaluating the effectiveness of our subsequent reduced-reference hints.

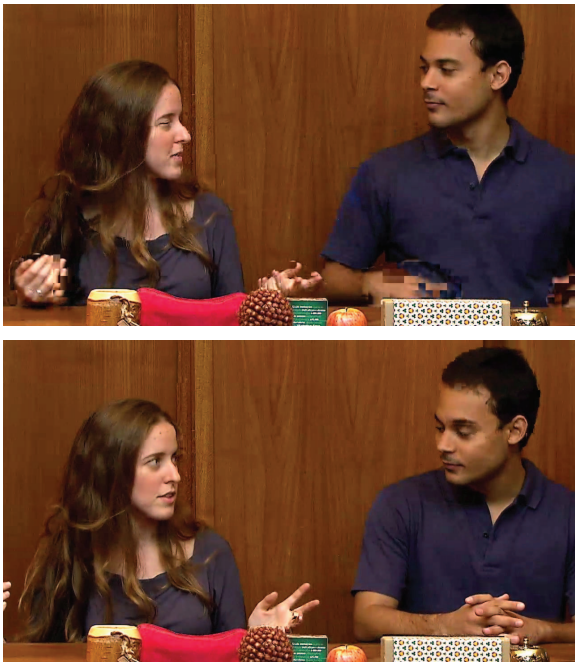


Fig. 1. Two loss-impaired pictures. The top picture, with PSNR of 28.07 dB compared to loss-free transmitted picture, is more objectionable than the bottom picture with a lower PSNR of 25.40 dB.

2.1. A Target Shift-Invariant Metric

There are many possible ways to compute distortion between a received picture R and sent picture S that exhibit some degree of shift-invariance. We adopt a direct approach in this paper. Instead of computing PSNR between R and S, our chosen target metric achieves shift-invariance by performing motion compensation (MC) over a small range, then compute the PSNR between R and MC(S). We call this the motion-compensated PSNR (MC-PSNR). In Fig. 1, MC-PSNR succeeds in capturing human visual preference where PSNR fails, with a MC-PSNR of 30.17 dB and 34.67 dB for the top and bottom pictures, respectively. In the results section, MC-PSNR is computed using 16×16 blocks with a search range of 4 or about 0.5% picture height for employed 720p content.

We choose this metric over the shift-invariant extension of SSIM [3] due to the latter's inability to handle localized errors. For example, with a sizable impaired region that is 10% picture width and height, 99% of the image would have no distortion, giving a score over 0.99 out of 1, regardless of how distorted that region is. In contrast, squared-error based measures such as MC-PSNR can better represent large errors with small spatial support. Nevertheless, the purpose of the target metric is for benchmarking only. Other choices may be equally appropriate.

2.2. System Overview

Given a sent picture S and a received picture R, our goal is to design a hint $h(S)$ whose compressed version $\hat{h}(S)$ can be used in place of S at the receiver for estimating a shift-invariant distortion D. A picture is displayed if the computed distortion D is smaller than a threshold T:

$$D(\hat{h}(S), R) < T.$$

In practice, it is common to compute $h(R)$, and determine D simply as the mean square error between $\hat{h}(S)$ and $h(R)$. Thus, we have:

$$\text{display if:} \qquad \text{MSE}(\hat{h}(S), h(R)) < T \tag{1}$$

which is the approach we adopt in this paper. The procedures are outlined in Fig. 2. We next describe two variants of h that can be used with (1) to achieve shift-invariance.

2.3. Shift-invariant Hints via Picture Resize

A well-established but shift-sensitive quality metric is PSNR. One way to reduce such sensitivity is through picture downscaling. Specifically, instead of PSNR between S and R, we seek to determine the PSNR between their respective downscaled versions s and r. While a receiver can compute r, sending s as hint represents an impractical overhead. Instead, we employ the technique of [1] as follows to achieve low-rate computation of PSNR between r and s.

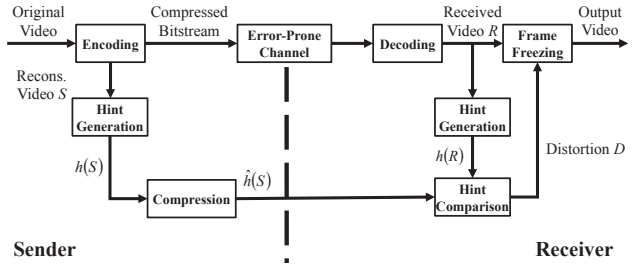


Fig. 2. A compressed hint $\hat{h}(S)$ is designed to allow distortion computation in a shift-invariant manner.

$$h_{resize} = J240^{+}(s) \tag{2}$$

where $J240^{+}(x)$ is a vector obtained by first performing a pixel-wise multiplication of image x with a pseudo-random sequence of ± 1, followed by a Walsh-Hadamard transform (WHT), then pixel-wise multiplication with a second pseudo-random sequence of ± 1, followed by an inverse WHT, and then sampling. This method in [1] is denoted by $J240^{+}$ since it is an extension of the methods in ITU J.240 to handle localized errors due to losses.

For 720p pictures in the results section, s and r are downscaled by 8 in both dimensions from S and R, respectively, and the hint employs 80 coefficients. We refer to this scheme as *ResizePSNR*. For comparison purposes, we also employ the similar scheme *FullPSNR* but without resizing, i.e., $h(S) = J240^{+}(S)$.

2.4. Shift-invariant Hints via Frequency Dropping

Another method to realize shift invariance is to explicitly discard high frequency components of the signals as follows:

$$h_{DCT} = Select(\text{DCT}(S)) \tag{3}$$

where DCT is the 2-D discrete cosine transform, and *Select* retains a small set of low-frequency coefficients. In the results section, 80 low frequency DCT coefficients are used as hint, and this scheme is denoted by *SelectDCT*. Both *ResizePSNR* and *SelectDCT* perform low-pass filtering. The key difference is that *ResizePSNR* capture more high frequencies but with less precision due to sampling.

2.5. Temporal Adjustment of Threshold

So far, we have discussed how distortions can be computed for each frame independently. Nevertheless, we know that freezing 20 consecutive pictures is disruptive to viewing while freezing the same number of pictures every other frame yields acceptable quality. Clearly, we should adapt the threshold in (1) based on past frame freezing decisions.

It has been shown in [5] that viewer mean opinion score (MOS) can be accurately modeled by the number of freeze episodes and their durations in the last 10 seconds. Following their findings, we simplify their proposed model to avoid per-sequence training for fitting MOS. For a picture freeze of duration τ, we compute the degradation $e(\tau)$ as:

$$e(\tau) = \frac{53.03}{1 + \left(\frac{562}{\tau}\right)^{1.01}} \tag{4}$$

We propose to change the threshold T for each video frame based on the total quality degradation as follows:

$$T = c_1 + c_2 \sqrt{\sum_i e^2(\tau_i)} \tag{5}$$

where the summation is over all freeze episodes, each with duration τ_i, in the last 10 seconds, and c_1 and c_2 are positive constants where c_1 represents the threshold used to judge each frame independently and c_2 controls how much we increase the threshold to freeze fewer frames in case of burst errors. In the results section, c_1 and c_2 are empirically chosen to be 10 and 1, respectively.

3. EXPERIMENTAL RESULTS

We next present results for two 30 fps, 720p sequences, *Shields* and *Conference*, with distinctly different characteristics. *Shields* exhibits slow and predominantly panning motion, yielding concealment artifacts that fit our shift-invariant assumption. In contrast, *Conference*, as shown in Fig. 1, contains stationary background and human subjects with complex motion. This yields many concealment artifacts that differ from small shifts. *Conference* is cropped from 1080p source from the Federal University of Rio de Janeiro. Both sequences contain 300 frames, and are repeated in a loop for 50 times to obtain results in this section.

Fig. 3-(a) shows a segment of the PSNR trace when *Shields*, encoded using H.264, is subjected to simulated packet loss ratio of 0.5% with average burst length of 3, and a round-trip time (RTT) of 200 ms. Losses are generally corrected within one RTT using reference picture selection. All loss impaired pictures are marked in cyan, with the pictures A and B having the highest and lowest PSNR, respectively. This means freezing decisions made using PSNR would likely display A but not B. The freezing decisions achieved using our shift-invariant target measure of Section 2.1 are marked in red, where B is displayed but not A, indicating that extreme preference reversal is possible. In other words, using PSNR to guide freezing decisions is likely to unnecessarily omit good pictures while inadvertently display pictures with breakup.

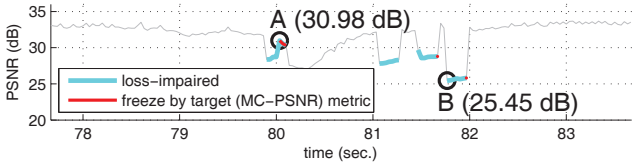
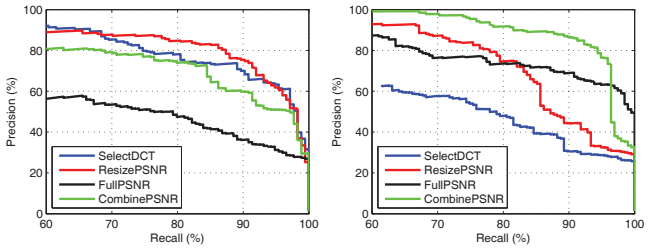


Fig. 3. PSNR trace for *Shields* showing extreme preference reversal between pictures A and B by PSNR and our target shift-invariant metric MC-PSNR.

We next characterize the decision quality of the proposed shift-invariant hints *ResizePSNR* and *SelectDCT* with respect to our target shift-invariant metric MC-PSNR that has access to the sent picture. Using simulation with the same loss and delay as before, we first examine all loss-impaired pictures to find the set of *bad* pictures that should not be shown according to our target MC-PSNR metric. There is no optimal bad set since it depends on various factors such as viewer preference and viewing distance. Instead, for the sake of comparison, we choose the worst 10% as the bad set. The bad set of MC-PSNR is generally different from the bad set of other metrics. We then examine the precision and recall achieved by different hints as freeze decision threshold in (1) is varied. Precision of a hint is the percentage of pictures in its bad set that is also in the bad set of MC-PSNR, and recall is the detection percentage of the bad set of MC-PSNR. The temporal threshold adjustment of Section 2.5 is not applied, since we are only interested in how well the different hints approximate our target MC-PSNR metric.

The results for *Shields* are shown in Fig. 4-(a). We see that *ResizePSNR* performs marginally better than *SelectDCT*, with both shift-invariant hints significantly outperforming *FullPSNR*, which estimates full frame PSNR. Since imperfections in the loss-impaired pictures of *Shields* are dominated by small shifts, this shows that both shift-invariant hints are effective in approximating MC-PSNR while PSNR based metric fails. Specifically, at 90% recall, i.e., when we are willing to accept 10% undetection rate of bad pictures, *FullPSNR* has precision of 36%, meaning for every bad frame detected,



(a) Shields (b) Conference

Fig. 4. Precision and recall for the frame freezing decisions generated by various hints with respect to our target metric MC-PSNR.

1.77 good frames is inadvertently mis-classified. In contrast, the corresponding mis-classification rate for *ResizePSNR* is 0.32 frame, a reduction of 82%.

For *Conference*, we can see that *ResizePSNR* again outperforms *SelectDCT*, indicating that the latter is too aggressive in discounting higher frequencies. More interestingly, unlike in *Shields*, *FullPSNR* which estimates full frame PSNR outperforms both *ResizePSNR* and *SelectDCT* at higher recalls above 85%. This is explained by the complex motions in *Conference*, which yields imperfections that are picture breakups rather than shifts in the worst impaired pictures. For those worst pictures, motion compensation is unlikely to help and MC-PSNR is essentially simple PSNR. In contrast, by de-emphasizing higher frequencies, both *ResizePSNR* and *SelectDCT* introduce more deviation from MC-PSNR.

Our shift-invariant hints yield good agreement with MC-PSNR for translation motions, but PSNR yields better agreement under complex motions. It is natural to seek hybrid hints that combine shift-invariance and full frame PSNR. There are many possible hybrids, and one simple choice is given by *CombinePSNR*, which spends half its hint bit-budget on a *ResizePSNR* hint, and the other half on a *FullPSNR* hint. The resulting estimated MSE for *CombinePSNR* is simply taken as the geometric mean of the two noisier MSE according to *ResizePSNR* and *FullPSNR*. The results are shown in green in Fig. 4. For *Shields*, its performance remains close to *ResizePSNR*, but generally lies between *ResizePSNR* and *FullPSNR*. For *Conference* however, *CombinePSNR* significantly outperforms both *ResizePSNR* and *FullPSNR*. This suggests that a shift-invariant component in the hint can significantly improve agreement with target MC-PSNR even for content without significant panning.

It is perhaps surprising that the hybrid scheme *CombinePSNR* outperforms both its constituents *ResizePSNR* and *FullPSNR* for *Conference*. This phenomenon is best explained using Fig. 6, where the PSNR between the sent and loss-impaired received pictures are shown sorted in descending MC-PSNR for various hints. A hint in perfect agreement with MC-PSNR would rank the loss-impaired pictures in the exact same order, yielding a monotonically decreasing curve. We already explained why *FullPSNR* is in agreement with MC-PSNR for the worst pictures in *Conference*. This is shown by the near monotonic behavior for the *FullPSNR* curve in Fig. 6 beyond rank 1400. Nevertheless, the significant variation between the ranks of 600 and 1400 causes general inability to distinguish the good from the bad. Variations in the same range are also present for the *ResizePSNR* curve but at a much lower degree. More importantly, the variations in *ResizePSNR* and *FullPSNR* are likely to be independent. Since MSE of *CombinePSNR* is formed by the geometric mean of its constituents' MSE, its PNSR is their average PSNR, which will show smaller variation. As a result, the *CombinePSNR* curve is more monotonic than either *ResizePSNR* or *FullPSNR*. The

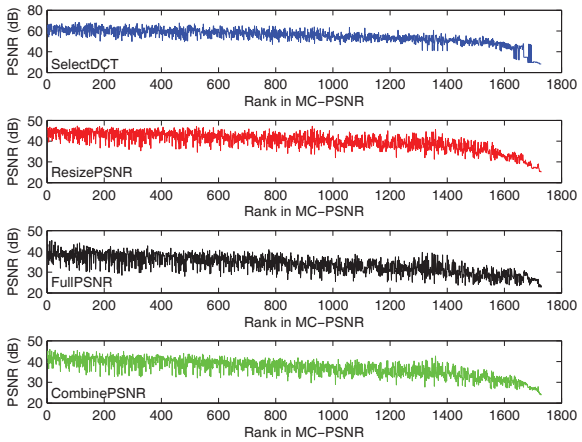


Fig. 5. PSNR between the sent and loss-impaired received pictures according to various hints sorted in descending MC-PSNR for *Shields*.

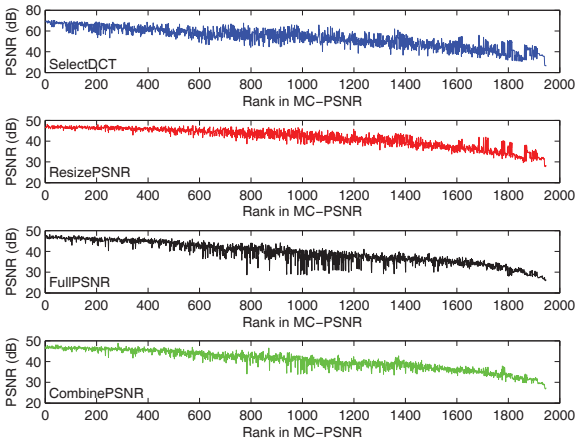


Fig. 6. PSNR between the sent and loss-impaired received pictures according to various hints sorted in descending MC-PSNR for *Conference*.

corresponding results for *Shields* are shown in Fig. 5.

Fig. 7 presents the results of the adaptive temporal adjustment of threshold T for frequent errors. Fig. 7-(a) shows a PSNR trace for *Conference* with loss impaired pictures marked in bold. The temporal evolution in T is shown in Fig. 7-(b) where we see that raises in threshold last at least 10 seconds. The frame freezing decisions of *ResizePSNR* using constant threshold (by setting $c_2 = 0$ in (5)) and adaptive threshold are shown in Fig. 7-(c). We see that the use of adaptive threshold successfully suppresses close cluster of freeze episodes and reduces the duration of some freeze episodes.

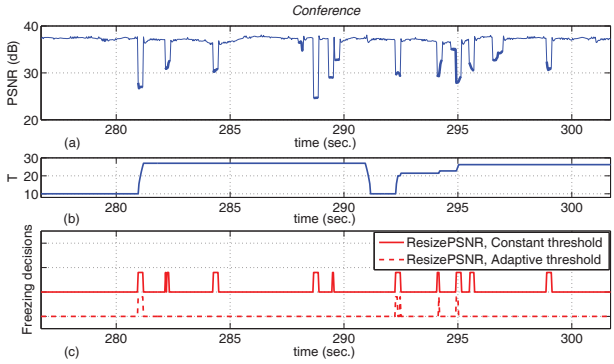


Fig. 7. Adaptive temporal adjustment of T. (a) PSNR trace, (b) adaptive threshold values and (c) frame freezing decisions with constant and adaptive thresholds (high: freeze, low: display)

Finally, we discuss how negligible bit-rate overhead of 1% is achieved for sending the hints. We encode the video sequences at 2 Mbps, 1% of which is 667 bits per frame. With 80 coefficients per frame, it suffices to quantize each coefficient to 8 bits without further entropy coding. In this paper, we use only 7 bits per coefficient to leave room for headers for $J240^+$ based hints. For *SelectDCT*, different number of bits are used for different frequency components according to their dynamic ranges. The overhead can be further decreased by entropy coding, e.g., by using distributed source coding ideas as in [6] to exploit correlation between $h(R)$ and $\hat{h}(S)$.

4. CONCLUSIONS

For the purpose of determining frame freeze decisions, we show that shift-invariant metric such as MC-PSNR can be approximated in a reduced reference framework by resizing the picture or dropping DCT coefficients. For a panning sequence, these methods produce superior precision and recall compared to PSNR based methods. For a sequence without significant panning, we show that these schemes can be combined with PSNR measures to improve precision and recall with respect to MC-PSNR. We also propose an adaptive thresholding technique to account for a viewer's increased dissatisfaction when freeze episodes are clustered.

By allowing selective freezing of visually unpleasant pictures only, our technique significantly improves the visual quality of the resulting video at a lower than 1% increase in bit-rate.

5. REFERENCES

[1] R. Kawada, O. Sugimoto, A. Koike, M. Wada, and S. Matsumoto, "Highly precise estimation scheme for remote video PSNR using spread spectrum and extraction of orthogonal transform coefficients," *Electronics and Communications in Japan, Part 1*, vol. 89, no. 6, 2006.

[2] *Framework for remote monitoring of transmitted picture signal-to-noise ratio using spread-spectrum and orthogonal transform*, Std., ITU-T Recommendation J.240, 2004.

[3] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. on Image Processing*, vol. 18, no. 11, pp. 2385-2401, Nov. 2009.

[4] A. Rehman and Z. Wang, "Reduced-reference SSIM estimation," *Proc. International Conference on Image Processing (ICIP)*, Hong Kong, Sept. 2010.

[5] R. R. Pastrana-Vidal, J.-C. Gicquel, C. Colomes, and H. Cherifi, "Frame dropping effects on user quality perception," *Proc. 5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, April 2004.

[6] K. Chono, Y.-C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod, "Reduced-reference image quality estimation using distributed source coding," *Proc. International Conference on Multimedia and Expo*, Hannover, Germany, June 2008.