NO-REFERENCE QUALITY ASSESSMENT OF HEVC VIDEOS IN LOSS-PRONE NETWORKS

Mohammed A. Aabed and Ghassan AlRegib

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332, U.S.A.
{maabed, alregib}@gatech.edu

ABSTRACT

In this paper, we propose a no-reference quality assessment measure for high efficiency video coding (HEVC). We analyze the impact of network losses on HEVC videos and the resulting error propagation. We estimate channel-induced distortion in the video assuming we have access to the decoded video only without access to the bitstream or the decoder. Our model does not make any assumptions on the coding conditions, network loss patterns or error concealment techniques. The proposed approach relies only on the temporal variations of the power spectrum across the decoded frames. We validate our proposed quality measure by testing it on a variety of HEVC coded videos subject to network losses. Our simulation results show that the proposed model accurately captures channel-induced distortions. For the test videos, the correlation coefficients between the proposed measure and the full-reference SSIM values range between 0.70 and 0.80.

Index Terms— video quality monitoring, high efficiency video coding (HEVC), temporal distortion propagation, video streaming, network losses

1. INTRODUCTION

The Joint Collaborative Team on Video Coding (JCT-VC) earlier this year completed the final draft for the new standard for video coding, high efficiency video coding (HEVC) [1]. Furthermore, the telecommunication standardization sector of the International Telecommunication Union (ITU-T) has approved HEVC as one of its standards (H.265) [2]. HEVC has double the coding efficiency of H.264/MPEG-4 AVC and supports up to 8K ultra high definition (UHD) videos [3, 2]. Moreover, HEVC introduces new coding tools and features to facilitate higher compression gain.

The paramount coding performance of HEVC comes at the expense of a more complex encoding operation compared with AVC. HEVC introduces the coding unit tree (CTU) structure which allows more flexibility for coding, transform, and prediction modes [3]. Furthermore, HEVC employs an open Group of Picture (GOP) format in which inter-coded pictures are used more frequently than AVC to allow higher compression gain. These features, however, make the bitstream and the decoded sequence more sensitive to errors and losses due to the higher level of data dependency. This, in turn, introduces more challenges in terms of video quality assessment and monitoring, error concealment, etc. To this end, we investigate in this work the impact of channel errors or losses on the fidelity of the decoded HEVC video by estimating the channel-induced distortion.

The problem of quality assessment for streamed video sequences has been recently addressed in several papers in the literature [4, 5, 6, 7, 8]. In [4], the authors measure the subjective score of video quality by proposing a video quality metric based on features obtained from the packet-headers of the bitstream. Staelens *et al.* [5] use genetic programming symbolic regression to formulate a no-reference bitstream-based video quality metric. De Simone *et al.* [6] report the performance of their subjective quality assessment campaign of the HEVC standard involving 494 test subjects. The authors in [7] test the performance of various full-reference quality metrics on 4k UHD videos. This work shows that PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM metrics were accurate in distinguishing different quality levels for the same content. In [8], the feasibility of the HEVC standard for UHD broadcasting services is examined. The authors report their results and analysis of subjective quality assessment of 4k-UHD HEVC videos. The work herein addresses the objective quality assessment of streamed HEVC videos subject to network losses with access only to the decoded videos.

In this paper, we propose a no-reference video quality measure for HEVC videos. We begin by examining the coding conditions in HEVC and the impact of network losses on the decoded video. We show that network losses has a more severe impact on HEVC videos compared with AVC videos. We then introduce a no-reference distortion measure, which exploits only the temporal variation of the spectral density between the frames. One of the contributions of this work is that the proposed approach does not make any assumptions on the concealment technique, network conditions or coding

parameters. It blindly operates on the decoded video after the decoder. We argue that the change in the spectral density between frames can pinpoint the amount of distortion in the frames.

The rest of this paper is organized as follows. In section 2, we illustrate the significant impact of network losses on an open GOP. We then explain our mathematical model to estimate the channel-induced distortions, which operates in the frequency domain. Section 3 details the simulations setup and test sequences used in the experiments herein, followed by the results and analysis of the model validation experiments. Finally, section 4 concludes the paper and outlines future directions of this work.

2. NO-REFERENCE VIDEO QUALITY ASSESSMENT

In this section, we begin by explaining the new coding structure in HEVC and the impact of network errors or losses under these coding conditions. Next, we illustrate our proposed no-reference video quality metric and the intuition behind it. We note that our approach operates only on the decoded video without making any assumptions about the encoding configurations, error concealment strategy or network conditions.

2.1. Error Propagation in an Open GOP Structure

The design of HEVC standard included many new features to efficiently enable random access and bitstream splicing. Many functionalities such as channel switching, seeking operations, and dynamic streaming services require a good support of random access. In contrast to H.246/MPEG-4 AVC, HEVC employs an open GOP operation. In this format, a new clean random access (CRA) picture syntax is used wherein an intra-coded picture is used at the location of random access point (RAP) to facilitate efficient temporal coding [3]. The intra period varies depending on the frame rate to introduce higher compression gain [9]. This coding structure is shown in Fig. 1. In this figure, frames are represented using circles and the order at the bottom of the figure is the picture order count (POC). The sequence starts with an I-frame (POC 0) which is followed by a P-frame (POC 8) and 7 B-frames (POCs 2 through 7) to form an open GOP of size 8. The next open GOP starts with the P-frame (POC 8) from the previous GOP (frames 8-16 in Fig. 1). This pattern continues until the end of the intra period. The arrows in the figure represent decoding dependencies.

In HEVC, favouring inter-coding over intra-coding is more subtle than in AVC. As a result, HEVC imposes a very high data dependency between the frames. Henceforth, the impact of channel-induced errors on certain frames that potentially propagate to the end of the GOP is more significant in HEVC than in AVC. Fig. 2 shows an example of the impact of loosing the Network Abstraction Layer (NAL) unit corresponding to frame 8 and replacing it with the temporally

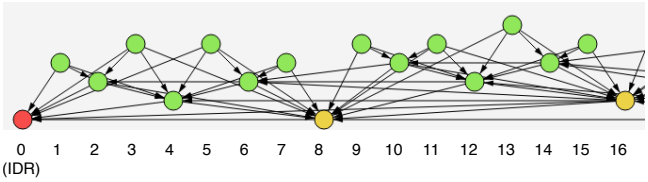

Fig. 1. The open GOP structure in HEVC coded videos [10]

.

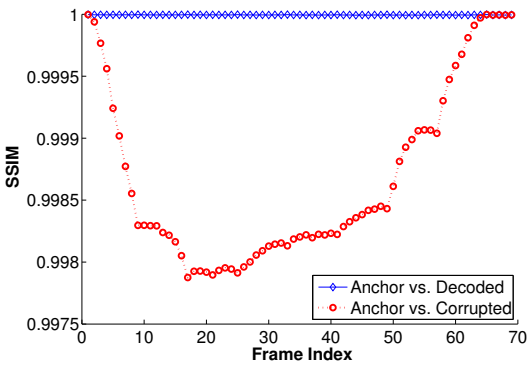

Fig. 2. The impact of loosing frame 8 on the SSIM values of the GOP for `BQMall` sequence; frame rate is 60 frames per second.

closest available frame at the decoder, which is frame 0 in this example (See Fig. 1). In our simulations and tests, we abide by the recommended encoding format wherein every frame is taken as a single slice which is encapsulated in a separate NAL unit [11]. Fig. 2 shows that the channel loss under these coding conditions propagates until a new I-frame is encountered, which is frame 64 in this example.

Under the assumption that we do not have access to the decoder and we only have access to the decoded sequences as explained in section 1, we do not have knowledge of how losses have propagated to other frames. Hence, in order to estimate these distortions without any reduced or full reference information, we can only rely on the spatial and temporal features of the decoded video.

2.2. No-Reference Distortion Estimation

In this section, we explain our proposed no-reference video quality assessment metric. The proposed approach relies on the fact that any channel-induced distortion will result in a temporal inconsistency between frames within a GOP. We measure this inconsistency through the temporal variation of the Power Spectral Density (PSD) across frames. Let f_k and f_{k-1} be the frame of interest and previous frame, respectively. Furthermore, let P_k and P_{k-1} denote their respective PSDs:

$$P_k[v,u] = \frac{1}{MN}\left|\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f_k[m,n]\, e^{-j2\pi(um+vn)}\right|^2 \quad (1)$$

where k is the temporal index of the frame in the received video, $M \times N$ is the resolution of the video, and v and u are the discrete frequencies. We next divide the PSD, P_k, into non-overlapping blocks of size $L \times L$. We refer to the PSD of block i in frame f_k as $B_k(i)$. Similarly, $B_{k-1}(i)$ is the PSD of block i in frame f_{k-1}. For every block, we estimate the channel-induced distortion by measuring the energy difference in the temporal domain as follows:

$$\Delta B_k(i) = B_k(i) - B_{k-1}(i). \tag{2a}$$

We next measure the variation of the energy differences within block i in frame f_k as follows:

$$G_k(i) = \frac{\max\left[\Delta B_k(i)\right]}{\sqrt{\text{Var}\left[\Delta B_k(i)\right]}} \tag{2b}$$

where $\max[\cdot]$ is the maximum value in block $\Delta B_k(i)$, $\text{Var}[\cdot]$ is the variance of the values in block $\Delta B_k(i)$, and $G_k(i)$ is the ratio of the maximum PSD value in block i to the standard deviation of the PSD of the block. Next, we compute the negative mean of $G_k(i)$, denoted by D_k, taken over all the spatial indices i in frame k as follows:

$$D_k = -\text{E}\left[G_k(i)\right] \tag{2c}$$

where $\text{E}[\cdot]$ is the expectation operation taken over the spatial indices, i's, for all the blocks. It should be noted that while $B_k(i)$ and $\Delta B_k(i)$ are square matrices, $D_k(i)$ and D_k are scalars. Furthermore, the obtained vector for the whole sequence of D_k values is normalized to obtained $\tilde{D}_k$. Finally, we amplify the the estimated distortion as follows:

$$\hat{D}_k = \tilde{D}_k \cdot \sigma_s(k) \tag{2d}$$

where $\sigma_s(k)$ is the standard deviation of the vector $\left[\tilde{D}_{k-s}, \cdots \tilde{D}_k, \cdots, \tilde{D}_{k+s}\right]$. s is the window size, which is determined empirically.

The goal of the operation in (2d) is to scale the measured distortion in (2c) within the context of its neighbouring frames. If the variance of the measured quantity in (2c) is high, this indicates high variations in the PSD levels from one frame to another, which indicates higher error likelihood within the GOP. In our experiments, $s = 5$ and the block size is $L \times L = 16 \times 16$ pixels.

Let us consider a scenario where a frame, k, has been lost and replaced by its predecessor in display order. For this particular frame, (2c) produces $D_k = 0$. Since $-\infty < D_k \leq 0$, the normalized value will have values $0 \leq \tilde{D}_k \leq 1$.

3. EXPERIMENTS AND RESULTS

In this paper, all the experiments and tests follow the recommendations published by JCT-VC for common test conditions for HEVC [9]. We use a subset of six difference video sequences in our experiments. All the video sequences were coded using the HEVC standard using the test model version (HM 12.0) [11]. The coding was done using the main random access profile. Next we detail the coding parameters and the obtained results.

Sequence	Resolution	Intra Period	FPS	Number of Frames
RaceHorses	832x480	24	30	300
BasketballDrill	832x480	48	50	500
PartyScene	832x480	48	50	500
BQMall	832x480	64	60	600
BasketballDrive	1920x1080	48	50	500
ParkScene	1920x1080	24	24	240

Table 1. Test Video Sequences

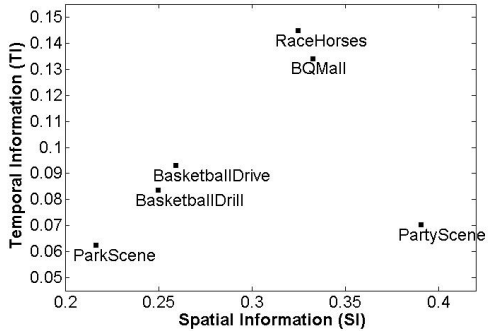

Fig. 3. Spatial information (SI) versus temporal infomration (TI) indices for the selected sequences [13].

3.1. Coding Conditions and Simulations Parameters

Table 1 summarizes the sequences used in our experiments and the encoding parameters. We fix the initial Quantization Parameters (QPs) value to 32. For the error patterns, we use the the loss patterns in the proposed NAL unit loss software [12]. The results shown in this paper are performed with the 10% loss pattern, which results in 5%-7% loss rate in the tested sequences. In our experiments, only inter-coded frames are subject to losses. Furthermore, Fig. 3 shows the spatial information (SI) and temporal information (TI) indices on the luminance channel for the selected sequences, as per the recommendation in [13]. The higher the score on the SI or the TI scale, the more complex the spatial and temporal features of the test sequence. In this context, we diversify the selection of sequences to validate our model under different temporal and spatial features.

3.2. Results and Analysis

Figs. 4 and 5 show the calculated measures for RaceHorses and PartyScene sequences, respectively. From the two plots, we notice that the value of $\hat{D}_k$ peaks at the location of lowest SSIM score. These points correspond to the lost frames, which were replaced by previous frames during the

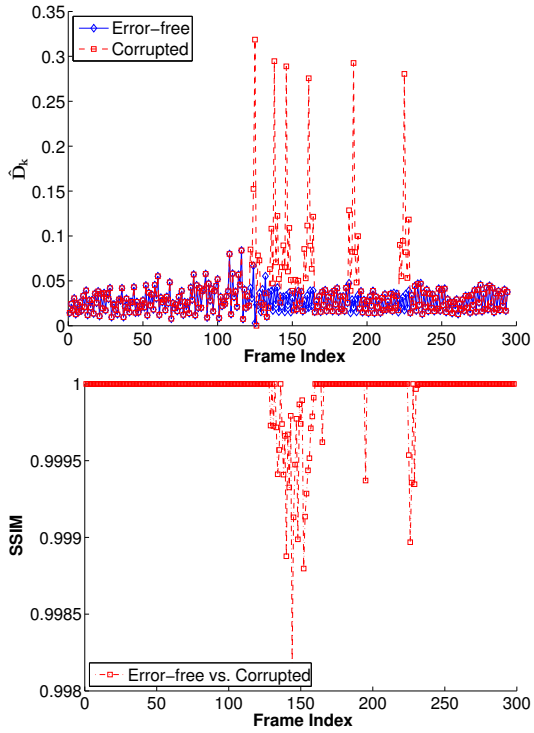

Fig. 4. The proposed no-reference quality measure compared with the obtained SSIM for the corrupted and error-free `RaceHorses` sequences.

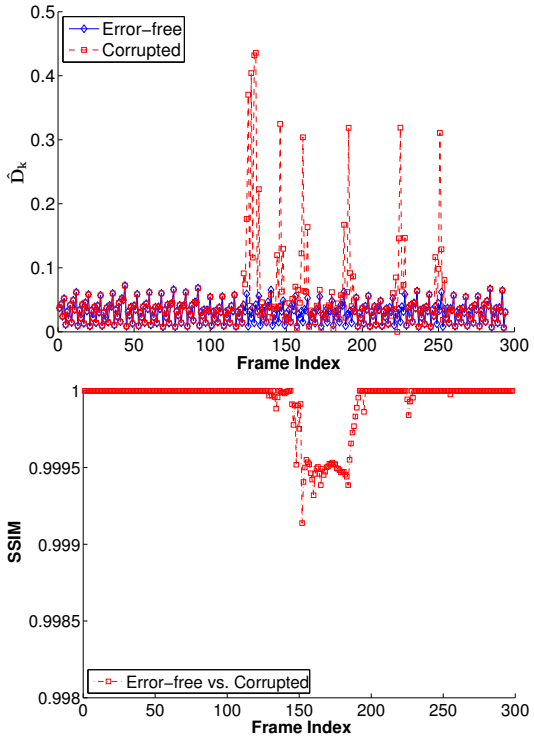

Fig. 5. The proposed no-reference quality measure compared with the obtained SSIM for the corrupted and error-free `PartyScene` sequences.

concealment process. In this case, $D_k \approx 0$, as alluded in Section 2.2. This value decreases for the following dependent frames since only a subset of the CTUs in these frames depend on the lost frames.

Sequences	Correlation Coefficients
`RaceHorses`	0.79
`BasketballDrill`	0.76
`PartyScene`	0.77
`BQMall`	0.70
`BasketballDrive`	0.80
`ParkScene`	0.77

Table 2. Correlation between the estimated frame distortion, D_k, and the full-reference SSIM values.

In order to validate the proposed distortion model, we calculate the correlation coefficients between the estimated distortion and the measured SSIM of the corrupted sequence compared with the error-free one. Table 2 summarizes the experimental results for all the tested sequences. Note that the proposed model correlates well with the SSIM values. The correlation coefficients for all test sequences range between 0.70 and 0.80. In particular, the proposed approach works well for the sequences with low temporal complexity such as the `ParkScene` video sequence. In this case, the majority of the changes in the PSDs between consecutive frames is due to the channel-induced distortion. Furthermore, our distortion measure works well for sequences with medium or low temporal complexity, such as `BasketballDrive` and `BasketballDrill`.

The correlation, however, tends to drop for the case of `BQMall` due to the complex nature of localized motion in the video, as can be observed from the TI index in Fig. 3. Nonetheless, this problem can be overcome by incorporating spatial inconsistency, which is beyond the scope of this paper. Our approach still performs fairly well for the `RaceHorses` sequence, which is close the `BQMall` in term of spatial and temporal features.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a new no-reference video quality measure to estimate the channel-induced distortion due to network losses. The proposed technique does not make any assumption about the coding conditions or video sequence. It rather explores the temporal changes between the frames, in the frequency domain, to estimate the the visual inconsistencies. We validate our approach by testing the proposed technique on various sequences and calculate the correlation coefficients with the full-reference SSIM values. Our experiments show that the proposed technique captures the erroneous frames due to both network losses and error propagation. In future work, we plan to improve the accuracy of the distortion estimation by including other features.

5. REFERENCES

[1] Benjamin Bross, Woo-Jin Han, Jens-Rainer Ohm, Gary J. Sullivan, Ye-Kui Wang, and Thomas Wiegand, "High efficiency video coding (hevc) text specification draft 10 (for fdis & final call)," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-L1003.v34*, jan 2013.

[2] ITU-T, "H.265: High efficiency video coding," Tech. Rep., ITU Telecommunication Standardization Sector, april 2013.

[3] G.J. Sullivan, J. Ohm, Woo-Jin Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] J. Ascenso, H. Cruz, and P. Dias, "Packet-header based no-reference quality metrics for h.264/avc video transmission," in *2012 International Conference on Telecommunications and Multimedia (TEMU)*, 2012, pp. 174–151.

[5] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference h.264/avc bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 8, pp. 1322–1333, 2013.

[6] Francesca De Simone, Lutz Goldmann, Jong-Seok Lee, and Touradj Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *J. Vis. Comun. Image Represent.*, vol. 22, no. 8, pp. 734–748, Nov. 2011.

[7] Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *2013 18th International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–8.

[8] Sung-Ho Bae, Jaeil Kim, Munchurl Kim, Sukhee Cho, and Jin Soo Choi, "Assessments of subjective video quality on hevc-encoded 4k-uhd video for beyond-hdtv broadcasting services," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 209–222, 2013.

[9] Frank Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-K1100*, Shanghai, China, 11th meeting, oct 2012.

[10] Parabola Research, *Parabola Explorer Software, Version 2.5*, University of Southampton Science Park, UK, 2013.

[11] Frank Bossen, David Flynn, and Karsten Sühring, *HM 12.1 Software Manual*, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-Software Manual, may 2013.

[12] Stephan Wenger, "Nal unit loss software," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG1, JCTVC-H0072*, feb 2012.

[13] ITU-T, "P.910: Subjective video quality assessment methods for multimedia applications," Tech. Rep., ITU Telecommunication Standardization Sector, 2008.