Using metadata in video quality assessment based on the structural similarity (SSIM) index metric

Tiago Gonçalves  
Instituto Universitário de Lisboa (ISCTE-IUL) – Instituto de Telecomunicações  
Lisboa, Portugal  
a24548@iscte.pt

Rui J. Lopes  
Instituto Universitário de Lisboa (ISCTE-IUL) – Instituto de Telecomunicações  
Lisboa, Portugal  
Rui.Lopes@iscte.pt

Paulo Nunes  
Instituto Universitário de Lisboa (ISCTE-IUL) – Instituto de Telecomunicações  
Lisboa, Portugal  
Paulo.Nunes@iscte.pt

ABSTRACT

In this paper, we propose the use of metadata to define the perceptual weight of semantically meaningful regions in video frames for video quality assessment. The proposed approach is evaluated using the structural similarity (SSIM) index metric and a comparison is made with the unweighted approach and the luminance-based weighted approach. Experimental results demonstrate that the proposed approach leads to significantly different mean SSIM values relatively to the other methods under comparison, suggesting a better correlation with the human perceived visual quality.

Categories and Subject Descriptors

E.4 [Coding and Information Theory]: Data compaction and compression;  

General Terms


Keywords

Video Quality Assessment, Visual Metadata, Structural Similarity (SSIM), Quality of Experience (QoE), EvalVid

1. INTRODUCTION

As more and more multimedia content is being delivered through best-effort IP networks, the video quality assessment methodologies become more relevant. In addition to the video encoding process and the terminal characteristics, the IP network also introduces impairments that should be considered.

In this context, the ultimate goal of quality assessment is to emulate the human’s perceived quality of the ‘consumed’ multimedia signals, also known as Quality of Experience (QoE) evaluation.

Subjective QoE evaluation, which is, therefore, concerned with the end-user experience, is usually accomplished by measuring the Mean Opinion Score (MOS) [1]. MOS is obtained from a large group of individuals that is asked to watch a set of videos and to rate their quality continuously in real-time, using a 1–5 scale. The subjective QoE assessment must follow a precise methodology and use a well-controlled environment [7], which makes the process very expensive and cumbersome.

On the other hand, the objective video quality assessment goal is to obtain a MOS, not from individuals’ opinions, but from measurable characteristics of the multimedia signals or from network measurements, such as average packet loss rate or sustainable bit rate. Peak Signal-to-Noise Ratio (PSNR), and the related quantity Mean Squared Error (MSE), are the most widely and well-known used Full Reference (FR) quality metrics used for image and video quality assessment. Nevertheless, PSNR has a limited range of validity as it measures essentially the sample-wise distortion between a reference signal and its ‘impaired’ version. The Structural Similarity (SSIM) index follows a different approach for video quality assessment [8], as it is based on the idea that the human visual perception is highly adapted for extracting structural information from a scene. Structures of the objects in a scene are independent of the influence of the luminance and contrast. Thus, luminance, contrast and structure are the separated components that are measured and compared. The SSIM indexing metric uses an 8x8 pixels sliding window approach, where the sliding windows moves pixel-by-pixel from the top-left corner to the bottom-right corner of the image. The overall quality value is defined as the average of the quality map, also known as the Mean SSIM (MSSIM) index.

A straightforward extension of the SSIM metric for video data, by simply averaging its value over entire frames and over time does not take into account that neither all regions in a visual scene nor all frames are equally important in terms of human visual perception. Therefore, the MSSIM averaging computed for each video frame should be modified to a weighted average, given the different importance of the various frame regions to the human observers. In addition, the same principle should be applied over time. In [9], the authors follow a simple approach of weighting differently the various sliding windows for each video frame according to their average luminance value. It is recommended to increase the weight of the windows with higher luminance, and weighting each frame in the video sequence according to its motion activity.

Other aspects are, however, involved in the perceptual relevance of the various scene regions, besides the local average luminance and motion activity values, which could be considered in weight assignment. While PSNR and SSIM are only based on the video signal, it is plausible that the semantic information has an important role on human perception of content. Therefore, in this paper, the proposed approach for quality evaluation is to use semantically meaningful metadata to describe the subjectively relevant regions in each video frame and to establish a weight mapping function according to the subjective relevance of each region. This approach constitutes a potential improvement for SSIM-based quality evaluation, but can also be used with other FR video quality metrics.
2. WEIGHTED MSSIM VIDEO QUALITY ASSESSMENT

The Structural Similarity (SSIM) index between signals $x$ and $y$ is

$$SSIM(x,y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{\mu_x^2 + \mu_y^2 + C_1(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where $\mu_x$ and $\mu_y$ are, respectively, the $x$ and $y$ averages, $\sigma_x^2$ and $\sigma_y^2$ their corresponding variance, $\sigma_{xy}$ is the covariance between $x$ and $y$, $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where $L$ is the pixel dynamic range (255 in our case), and $K_1$ and $K_2$ are two constants for stabilization purposes (in this paper $K_1 = 0.01$ and $K_2 = 0.03$, which are the typically recommended values [8]).

Notice that SSIM equals 1 when the images that are being compared are equal, and is less than 1 when they are not equal.

Since the SSIM is obtained for $8 \times 8$ windows, computing the MSSIM metrics involves averaging over these windows. Equally averaging all the windows on a single frame is the simplest way of computing MSSIM. Nevertheless, images are meaningful to humans, both in terms of pixels values, activity and relevant semantics, suggesting that a different weighting should be used.

2.1 Weights Based on Luminance

The approach proposed in [9] is based on the assumption that dark regions do not attract attention, therefore should be weighted with smaller values. Being $\mu_i$ the mean luminance for a given window and $w_{ij}$ the weight of the $j$-th window in the $i$-th frame, the local weighting is advised in [9] to be adjusted as

$$w_{ij} = \begin{cases} 0 & \text{if } \mu_i \leq 40 \\ \frac{\mu_i - 40}{10} & 40 < \mu_i \leq 50 \\ 1 & \mu_i > 50 \end{cases},$$

giving lower importance to low luminance.

2.2 Weights Based on Metadata

The alternative video quality evaluation approach that is proposed in this paper is to associate metadata information with the video in order that the importance of the video objects is defined a priori. How the metadata is obtained is not in the scope of this work.

The required elements for applying this method are: i) the semantic description of the sequence; ii) the definition of the frame regions associated with the semantic content; and iii) a mapping function of the descriptions in weights, for every region.

In this paper we have manually assigned weights to the regions that subjectively seemed to capture more visual attention. The regions were simply cataloged into three groups: i) high-significant regions (weight of 1); ii) medium-significant regions (weight of 0.5); and iii) low-significant regions (weight of 0.1). Figure 1 illustrates the regions used for a sample video frame and their associated weights. This particular video sequence consists of an aircraft slowly moving from left to right, but not traveling the entire image. The aircraft region is considered the most significant, followed by its boundaries and the control tower, which are half-significant. The rest of the scene almost does not capture visual attention, therefore is considered insignificant.

This technique can be used in any scenarios where the three previously defined elements can be defined.

3. ARCHITECTURE AND TEST METHODOLOGY

3.1 EvalVid

The core architecture used for video transmission quality evaluation is an extension to the EvalVid framework. EvalVid is a toolset for video quality evaluation transmitted over real or simulated communication networks [3][4] using a FR scheme. The transmission description is accomplished by two packet traces, one from the sender and another from the receiver.

Two main extensions were made to EvalVid. The first extended to change the error model of the network simulation, from a bit error rate model to a Packet Loss Rate (PLR) model, which is more relevant in our assessment. The second main extension was to add MSSIM measurement to the platform (since only PSNR and computed MOS from PSNR were originally implemented), and to support different weighting methods: equal, luminance-based, and metadata-based.

3.2 Test Sequences and Conditions

For the tests in this paper, we have used 3 video sequences from the LIVE Wireless Video Quality Assessment Database [5] representing airport hangar activity: vid1_hp.yuv (Sequence 1), vid6_hp.yuv (Sequence 2) and vid12_hp.yuv (Sequence 3). Each video has 768×480 pixels with 300 frames at 25 fps.

We simulated the transmission of the encoded video sequences (VBR H.264 coded with a target bit rate of 1 Mbit/s and a bit rate tolerance of 100 kbit/s) with five different uniform PLR conditions: 0%, 1%, 5%, 10% and 20% (50 simulations each), for each video sequence. The target bit rate was chosen in order to focus the video quality assessment on the network impairments due to packet losses and not on the video coding impairments.

Then, the equal weight, luminance-based weight and metadata-based weight methods were applied. For the metadata semantic descriptions of the Sequence 2 and 3, the same criteria illustrated in Figure 1 for Sequence 1 was used.

4. TEST RESULTS

Figures 2-4 show the results of the MSSIM measurements for each video sequence. Each group of columns refers to a different PLR, being each column associated with a weighting strategy (equal, luminance-based and metadata-based). The results are the average of MSSIM over the 50 simulation runs for each strategy.
and the error bars are defined for a 95% confidence interval, with a standard normal distribution.

Figure 2. Sequence 1 MSSIM for each PLR and weighting strategy.

Figure 3. Sequence 2 MSSIM for each PLR and weighting strategy.

Figure 4. Sequence 3 MSSIM for each PLR and weighting strategy.

For all test sequences (Figure 2-4) the MSSIM values are similar for a PLR of 0%, with the average and the confidence interval not differing too much. These values although very close, are not equal to 1 because the SSIM index has been measured between the original reference video sequence and the coded video sequence. Notice that for a PLR 0% only coding impairments are taken into account, while for remain PLRs measure the network impairments are also considered.

Figure 2 shows that, the luminance-based approach always increases the MSSIM value, while the metadata-based approach always decreases it. Figures 3 and 4 do not follow this trend. Although the luminance-based approach values are always higher than the metadata-based approach, the latter is not always lower than the equal average case, namely in the worst-case transmission conditions (PLR of 10% and 20%).

5. DISCUSSIONS AND CONCLUSION

We propose the use of metadata to enhance video quality assessment based on the SSIM index metric by balancing the relevance (weight) of the various regions of a video frame. Preliminary experimental results shown that the proposed approach significantly changes the obtained MSSIM values in the presence of packet losses indicating that this strategy deserves further investigation, namely, how the MSSIM values obtained with the proposed approach correlate with MOS values obtained through subjective testing. Effective non-ad-hoc ways to define the weight mapping function and the computational cost assessment are other possible topics to be explored. The weighting over time is another issue that should be studied, possibly taking into account end-user visual memory perceptual characteristics such as sensitivity to motion, and cinematographic effects like fade in/out. Future work will include also the extension of this approach to Reduced Reference (RR) metrics.

6. REFERENCES


