OBJECTIVE QUALITY EVALUATION OF VIDEO SERVICES

Marco Montenovo¹, Alessandro Perot¹, Marco Carli², Paolo Cicchetti², Alessandro Neri²

¹ HP C&I Italy, Roma, Italy
² University of Roma "Roma TRE", Roma, Italy

ABSTRACT

In this paper an objective metric for assessing video quality communication is presented. The proposed approach is based on the combination of two metrics: a Reduced Reference video quality metric for assessing the annoyance of analog noise and compression generated artifacts; a No Reference video quality metric for video corruption caused by packet loss in a video stream. Experimental results show the effectiveness of the proposed algorithm.

1. INTRODUCTION

The fast market penetration of new multimedia services like UMTS videotelephony and multimedia messaging, video on demand over Internet, satellite, terrestrial, and mobile digital video broadcasting, experienced during the last year, increases the relevance of real time video quality assessment in the design and provisioning of new multimedia services.

In fact, objective and reliable Quality of Service (QoS) assessment is a mandatory task for a proper regulation of the interaction among wired and wireless operators and service and content providers.

The experience collected during the last decades evidenced that data quality assessment can not be performed by monitoring just a few simple indicators like the classical Peak to Signal Noise Ratio (PSNR). The understanding and effective assessment of image/video quality is, in fact, an interdisciplinary task that requires the interaction among engineers, psychologists and physiologists.

Since the final user of the transmitted video is a person, the best way for assessing the quality of the received data is performing a subjective test to obtain the Mean Opinion Score (MOS).

Unfortunately, such a test is time consuming, highly costly and in general unsuitable for real time video delivering. In the recent past, many efforts have been focused in the construction of objective metrics that capture the human judgment.

These metrics can be classified in full-reference, reduced-reference and no-reference metrics according to the amount of knowledge about the original data required at the receiver side to perform the quality evaluation. Despite their accuracy and reliability, full reference methods exploit high implementation complexity when real time assessment of video quality at the end user premises is required.

To overcome this burden, “Reduced Reference” (RR) metrics require only a reduced version of the original or a description of it. “No-Reference” (NR) metrics are based on computation of parameters based only on the received videostream.

For instance, use of estimates of edge blurriness, random noise level and amount of artifacts introduced by the compression system as ringing and blocking, is proposed in [4] for NR image quality assessment. The reason behind this choice is that those features can be easily described in mathematical terms and evaluated in real-time.

However, several factors contribute to hardener the task of designing a reliable NR-metric: among them we can cite the limited knowledge of the Human Vision System and how the human brain processes the visual stimulus. Furthermore, the perceived video quality highly depends on the subject: his a priori knowledge of the topic, his interest in the content, the level of attention while looking at the video and so on.

In this contribution we propose a combination of a reduced reference metric and a no-reference method to assess the quality of a video transmitted over an IP network. The remaining part of the paper is organized as follows: in Section 2 the structure of the proposed metric is described. In Section 3 some results of the performed experiment are presented; finally, in Section 4 the conclusions are drawn.

2. PROPOSED METRICS

The system we are dealing with is presented in Figure 1. Basically, an Internet content provider delivers videos
to several clients in two ways: downloading and streaming the data over the IP network.

Fig. 1: Streaming system model.

Starting from [4], in this work we combine the evaluation of Gaussian and impulse noise levels, blockiness and edge blurriness, with factors that depend on the particular content we are dealing with, the compression systems usually adopted (MPEG-2 coders[5-7]) and the information losses introduced by the adopted communication system (e.g. streaming over IP network).

The Reduced Reference indicator is obtained by comparing the results of a set of video quality tests performed on both a reference video sequence (supposed to be of high quality) and the video sequence whose quality has to be estimated. Since the quality reduction due to the mentioned degradations is performed by comparing individual indicators respectively computed on the original and on the received sequence, we propose to embed the statistics computed on the original sequence in the coded bitstream by means of data hiding techniques.

It is worth to mention that the choice of the algorithm for computing each parameter has been driven by the ‘almost-real-time’ constraint. As well known, videos present huge amount of data to be process, and QoS constraints force strict thresholds for delaying the delivery of the video to the final user.

The basic indicators include:

- **Estimated Additive Gaussian Noise power level** based on Wiener filtering. Wiener filter in fact performs very well in presence of additive white noise and therefore the difference between input and output of the filter can be used to estimate the noise level [1];
- **Impulsive noise power level estimation** based on a median filtering. This filter is very efficient in eliminating pixels that violate uniformity bonds with the surrounding pixels. The exit of the filter is compared with the non-filtered image;
- **Blocking and blurring artifacts characteristics.** A combined index is used to evaluate the degradation induced by blocking and blurring; blocking figure of merit is calculated based on the average of pixel to pixel differences on block edges. Evaluation of blurring consistency combines activity inside the image (average of pixel to pixel differences) and zero crossing rate. To better approximate MOS, a combination of these two indicators is performed as described in [2][3]; an example of blocking artifact and its evaluation is shown in Figure2.
- **Statistics of ringing artifacts** detected by means of a Perona-Malik filter. The effectiveness of this filter as de-ringing filter makes it a good choice to evaluate the consistency of this kind of artifact. The spectrum of the resulting noise (from the difference of the original and filtered image) will be colored on an affected image and mainly white on a clean image [4].

Figure 2: Blocking artifact evaluation. top row original and distorted image; bottom the Spectrum Power Density of the two images

The no reference component of the quality metric mainly refers to the impact of temporal resolution reduction, packet losses, latency, and delay jitter.

We observe that perception of quality degradation strictly depends on the algorithms employed by the receiver to conceal the errors as well as on the intraframe temporal redundancy reduction strategies (intraframe coding).

On the other hand, although information on packet loss and out of sequence ratios can be derived by gathering the communication channel output, here we assume that only decoded output can be employed for quality assessment.

Thus, detection of blocks partially repaired by the error concealment procedure is performed by computing for each block the correlation coefficients time series. Based on the correlation coefficient two basic factors are tested: presence of blocking effect in small areas and edge continuity. The combination of these two estimates gives a good approximation of the MOS regardless of intrinsic video characteristic (slow or fast moving video, local image activity factor, etc...) and compression parameters (spatial and temporal resolution).

Presence of P-frames and B-frame in video streams is also taken into account. Scene and I-frame detection as well as methods to assign weights to different age
corruptions are used to evaluate the effects of the presence of predictive frames on corrupted video.

The parameters characterizing the overall reduced reference metric have been determined by means of subjective tests. A second set of subjective experiments has then been employed to verify the performance of the proposed method.

3. JERKINESS ASSESSMENT SUBJECTIVE TEST AND EXPERIMENTAL RESULTS

Adaptation of the video format to the terminal capabilities and to the capacity of the access channel often requires reduction of both spatial and temporal resolution. In particular, temporal subsampling can induce an unnatural jerky motion. Here we propose a jerkiness objective evaluation based on the analysis of the time series of the normalized intraframe correlation $\rho[k]$.

Let us denote with $\rho^0_i$ and $\rho^p_i$ the correlations between the $i$-th and $(i+1)$-th frame of the original and processed sequence, respectively. In addition we denote with

$$\Delta \rho_i = \frac{\rho^p_i - \rho^0_i}{\rho^0_i} \times 100 \quad i=1,\ldots,N-1$$

the normalized difference.

Without loss of generality, here we assume that the video is rendered using the same frame-rate of the original sequence. However, to account for the temporal subsampling associated to codecs and other ancillary functions, we introduce a sampling-and-holding mechanism between the video decoder and the rendering. As illustrated in Figure 3 where the interframe correlation time series for different subsampling factors of the “2001_a” video is reported, a jerky video is characterized by a burst correlation with many coefficients equal to 1, compared to the more homogeneous behavior of the coefficients extracted from the full temporal resolution sequence.

Consequently the normalized correlation differences $\Delta \rho_i$, corresponding to repeated frames present positive values with magnitude in the range $2 \div 4\%$, while negative values below $-4\%$ are observed in correspondence of the frame update. We note that, when the subsampling factor is increased, the correlation between extracted samples decreases so that the magnitude of negative normalized correlation differences also increases.

In principle one could expect that jerkiness could be detected by properly thresholding the normalized correlation differences $\Delta \rho_i$.

Unfortunately, the subjective experiments evidenced that it not possible to consider only the percentage of differences $\Delta \rho_i$ exceeding a given threshold, since these values are highly content dependent and thresholds cannot be generalized. Some considerations help in defining a generalized jerkiness metric. Let us denote with $k$ the first frame of a set of $n$ identical frames in a video. Two possible states occur:

- The $k$-th frame of the processed video is the first one after an abrupt change: a set of identical frames is played before this abrupt transition causing $\Delta \rho_k < 0$.
- The $k$-th frame is not the first one after an abrupt change: in this case $\Delta \rho_k = 0$.

Since $n$ consecutive frames are identical, the correlation among them is 1 and coefficients $\Delta \rho_{k+1}, \Delta \rho_{k+2}, \ldots, \Delta \rho_{k+n}$ are greater than $\Delta \rho_k$. When a new different frame is rendered, $\Delta \rho_k$ has a negative value. After that, we have a set of $n$ identical frames, so that $\Delta \rho_k$ will be equal to 0 once again. The described behaviour is pictured in Figure 4. It can be noticed that the group of identical frames present always a negative $\Delta \rho_k$ followed by $n$ coefficients $\Delta \rho_i \geq 0$ (top: $n=1$, bottom: $n=3$).

In state-of-the-art applications, where bit-rate is adapted to the available resources and to the channel congestion, the temporal resolution varies accordingly and the jerky motion is caused by the coder in a random...
fashion. This behavior can not be model as a sequence of repeated frames.

In our system, to cope with this feature, we use other two coefficients: the number $c$ of perceived abrupt changes and the average length of repeated frames $l$.

To determine the frame repeated number a threshold $Th_1$ is computed as follows:

$$Th_1 = 0.4 \times \max_i \{\Delta \rho_i\}.$$  

Let us denote with $L_{\text{near still}}$ the list of the indexes of the first frame of a near-still subsequence and with $R_{\text{near still}}$ the list of the lengths of the runs of those subsequences. Initially both lists are empty and the current index $k$ is set to 1. Then, for each $k$ in $[1, N-1]$, the algorithm compares the correlation difference with threshold $Th_1$. If $\Delta \rho_k > Th_1$, then the index $k$ is inserted in the list $L_{\text{near still}}$ and the corresponding run length is set to 1. The length of the near still sequence is incremented until a frame $(k+n)$ for which $\Delta \rho_{k+n} < 0$ is found. In that case the run length is inserted in the list $R_{\text{near still}}$ and the current frame is set to $(k+n)$. Let us denote with $|.|$ the cardinality of a set. The RR estimate of the jerkiness of the video is based on the following indicators:

- $c$: number of near-still subsequences, corresponding to the cardinality $|L_{\text{near still}}|$ of $L_{\text{near still}}$;
- $m$: percentage of near-still frames:
  $$m = \frac{|\{i | \Delta \rho_i > Th_1\}|}{N} \times 100$$
- average length of a near still subsequence
  $$f = \frac{1}{|L_{\text{near still}}|} \sum_i R_i$$

The knowledge of $f$ allows us to determine how much the jerky motion is perceived. The average of $\Delta \rho_i$ allows to discriminate when the correlation coefficients vary due to jerkiness or due to compression and noise. Its value will be lower when affected by jerky motion.

In these considerations should also be taken into account that compression and noise give different and opposite impact on the frame-to-frame correlation. The blocking artifacts caused by compression cause the frames to be more homogenous, while a random noise can increase the differences among frames. By analyzing these three parameters is possible to rate the amount of jerkiness in the video.

To calibrate the jerkiness assessment vs. the subjectively perceived annoyance, a subjective test has been performed.

We used ten subjects drawn from a pool of students at the University of Roma TRE. The students are thought to be relatively naive concerning video artifacts and the associated terminology. They are asked to wear any vision correcting devices (glasses or contacts) that they usually wear to watch television. A common computer monitor is used to display the test video sequences. The experiment is run with one subject at a time. Each subject is seated straight ahead of the monitor, and they are positioned at a distance of 80 cm from the video monitor.

The experimental session consists of four phases. In the first phase, the subject is verbally given instructions. In the second phase, training sequences are shown to the subject. The training sequences represent the impairment extremes for the experiment and are used to establish the annoyance value range. In the third phase, the test subjects runs through several practice trials. The practice trials are identical to the experimental trials and are used to familiarize the test subject with the experiment. Finally, the actual experiment is performed with the complete set of test sequences.

To each subject are presented in random order 70 test sequences of variable length between 7 and 10 seconds. Among the 70 sequences:

- 14 originals without any artifacts caused by compression or transmission;
- 56 distorted sequences obtained by the originals after decimation with 4 different steps.

After each video is displayed, the subject is asked to enter a value between 1 and 5, representing how annoying the impairment is, compared to the worst defect present in the training sequences. The value 1 corresponding to the minima (no artifact is perceived); the value 5 is the maximum (very annoying artifact).

Based on the subjective experiments we propose to evaluate the jerkiness effect based on the following logic:

1) **Jerkiness free:**
   $$m < 0.5 \text{ OR } f < 3.5 \text{ OR } c < 2$$

2) **Medium jerkiness**
   $$m > 0.5 \text{ AND } 3.5 \leq f < 4.5 \text{ AND } c \geq 3$$
   OR
   $$m > 0.5 \text{ AND } 4.5 \leq f < 5.5 \text{ AND } c \geq 2$$

3) **Heavy jerkiness:**
   $$m > 0.5 \text{ AND } f > 5.5 \text{ AND } c \geq 2.$$
In the following some experimental results are presented. In Figures 5 and 6 are presented two sequences used in the experiments and the corresponding algorithm outputs.

The figures show a set of the images of each video, a graph corresponding to the Mean Opinion Score (obtained by the subjective experiment) versus the decimation factor, and the outputs of the proposed algorithm for some decimation steps (2, 4, 5 and 6). The green color represents a no-jerkiness indicator for the corresponding subsequence, while yellow and red colors are the output if a mild or severe jerkiness has been detected.

When the decimation factor is equal to 2, no alarms are generated; when the subsampling factor is bigger than 4 an alarm is generated. In the ‘2001_a’ sequence the output 2 is generated (green) for all the sub sequences; this is not true for the ‘interview’ one. When the sequence is decimated at step 6, almost all the subsequences are classified as jerky. Same results have been also obtained for the interview sequence.

The MOS, obtained by subjective test described in the previous paragraph, confirm the results.

5. CONCLUSIONS

We have introduced a novel method for RR evaluation of video quality based on HVS characteristics. For the created set of test images the advantages of the proposed metrics are demonstrated in comparison to conventional PSNR and state-of-the-art UQI and MSSIM.

REFERENCES

http://anchovy.ece.utexas.edu/~zwang/research/nr_jpeg_quality/index.html