NO-REFERENCE QUALITY METRIC FOR COLOR VIDEO COMMUNICATION

F. Battisti, M. Carli, A. Neri

Università degli Studi “Roma TRE”
Dipartimento di Elettronica Applicata
Via della Vasca Navale, 84, 00146, Roma, Italy

ABSTRACT

In this paper an objective No-Reference metric for assessing the quality degradations of color videos introduced by transmission over a heterogeneous IP network is presented. The proposed approach is based on the analysis of the interframe correlation measured at the output of the rendering application. It does not require information about the kind of errors, delays and latencies that affected the link and countermeasures introduced by decoders in order to face the potential quality loss. Experimental results show the effectiveness of the proposed algorithm in approximating the assessments obtained with Full Reference metrics.

1. INTRODUCTION

In the last few years, a fast market penetration of new multimedia services has been experienced. This spreading is accompanied by an increasing need for assessing the perceived quality of those services. This is important for evaluating the effectiveness of the offered services and customer satisfaction, supporting, for instance, service plan options selection and price policy. During the last decade, research on video quality has been mainly focused of the development of objective video quality metrics that should be able to reproduce the subjective evaluation of the observers and that are based on the knowledge of both the original and the received video stream (Full Reference quality metrics - FR).

In [4] a color video quality assessment based on edge-color distortion is presented. This FR metric is a modification of state of the art quality metrics based on edge and color distortion models. The response functions of HVS to the two models are obtained by analyzing relation between predicted value of the two models and the subjective quality estimation. Experimental tests show a good correlation between the proposed metric and the perceived video quality.

The main drawback of the FR techniques is that they cannot be used in real time applications because the original video is not available at the receiver side. To partially overcome this problem, reduced reference (RR) quality metrics as well as no-reference (NR) metrics have been devised. RR methods are based on the extraction of representative features from the content to be evaluated that are transmitted to the receiver with a negligible bandwidth increment. The quality assessment is driven by the received side information [6].

Recently, a RR metric [1] based on interest points extracted from regions with high or low activity is presented. The proposed metric has a short computational time and it is then suitable for real time usage. The obtained results are promising and they show a good correlation with the MOS (Mean Opinion Score).

In the field of 3D media, the authors of [3] present a RR quality metric based on edge detection for compressed depth maps associated with color plus depth 3D video. In this paper edge information is used as side information. The obtained results are compared with a FR metric and the proposed method shows a good correspondence between the FR and the RR method.

In order to overcome the modification of bitstream formats and communication protocols introduced by RR metrics, NR metrics, that are based on the received video-stream only, are the only applicable solution. Although less accurate than FR metric, NR metrics can be applied without the need of extra information.

In [2], a NR video quality metric based on the HVS (Human Visual System) is presented. The proposed metric relies on a 3D multi-spectral wavelet transform and takes into account the sensitivity of the HVS to luminance, contrast and activity. The experimental results show a good correlation of the achieved results with subjective quality assessment.

In this contribution we propose a No-Reference method that is able to assess the degradations of a color video introduced by transmission over an heterogeneous IP network.

Decrease of channel reliability as well as sudden increase of the offered traffic can lead to loss of both data and temporal integrity that, depending on the characteristics of the adopted protocols (e.d. TCP vs. UDP), and play-out temporal constraints (e.g. maximum temporal delay) can appear at the decoder input as losses. This could lead to the impossibility of decoding isolated as well as clustered blocks, tiles, and even entire frames. Considering the continuous increase of computing power of both mobile and wired terminals, we can expect a wide spread of error concealment techniques aimed at increasing the perceived quality. In this work it is assumed that the output of the rendering algorithm is observed. There is no knowledge about the kind of errors, delays and latencies that affected the link and countermeasures introduced by decoders in order to face the potential quality loss. The work is based in the gray scale NR quality metric developed in [7]. Here we extend the work to color videos by considering the impairing caused in each color component.

The rest of the paper is organized as follows: in Section 2 the proposed metric is described, in Section 3 some results of the performed experiment are presented and in Section 4 the conclusions are drawn.

2. PROPOSED APPROACH

The proposed approach for evaluating the video quality relies on the observation that channel errors and end-to-end jittering delays can produce different artifacts on the received video including the loss of one or more consecutive frames or the loss of isolated or clustered blocks.

In the proposed approach the authors deal with the most commonly used strategy for error concealing that is based on the study of frame repetition. In this case, the concealed frames are characterized by a high temporal correlation; to deal with this problem, the rendered sequence is first partitioned into static and dynamic shots. Next, the shots classified as static, are evaluated in order to detect if the small amount of changes in the shot corresponds to the loss of consecutive frames or to a static scene. At the same time, the dynamic shots are tested to verify the presence of isolated and clustered corrupted blocks. These analysis result in a temporal and distortion maps.

2.1. Frame segmentation in dynamic and static shots based on a global temporal analysis

The first step of the NR procedure is the grouping of frames in dynamic or static shots based on a temporal analysis. In order to perform this operation, for each color component (R, G, B) the following procedure is performed.

Let $\mathbf{F} = \{\mathrm{F_k}, \mathrm{k} = 1, ..., \mathrm{L}\}$ denote one color component of a video sequence composed by L frames of size $m \times n$ pixels. The generic k^{th} frame can be partitioned in $N_r \times N_c$ blocks $\mathbf{B}_k^{(i,j)}$, whose top-left corner is located in (i, j), of size $r \times c$ pixels. Let $\bar{F}_k$ be the mean luminance value for the k^{th} frame and $\bar{B}_k^{(i,j)}$ the mean luminance value of block $\mathbf{B}_k^{(i,j)}$. Let $\Delta\mathbf{F}_k = \mathbf{F}_k - \bar{F}_k$ and $\Delta\mathbf{B}_k^{(i,j)} = \mathbf{B}_k^{(i,j)} - \bar{B}_k^{(i,j)}$ denote the deviation of the luminance of the k^{th} frame and of the block $\mathbf{B}_k^{(i,j)}$ from the corresponding mean values.

The normalized inter-frame correlation coefficient ρ_k between the k^{th} and the $(k-1)^{th}$ frames is defined as:

$$\rho_k = \frac{\langle \Delta\mathbf{F}_k, \Delta\mathbf{F}_{k-1} \rangle}{\|\Delta\mathbf{F}_k\|_{L^2} \, \|\Delta\mathbf{F}_{k-1}\|_{L^2}}, \tag{1}$$

where $< \bullet, \bullet >$ denotes the inner product and $\|\bullet\|_{L^2}$ the L^2 norm. Similarly, the inter-block correlation $\rho_k^{B(i,j)}$ can be computed as:

$$\rho_k^{\mathbf{B}(i,j)} = \frac{\left\langle \Delta\mathbf{B}_k^{(i,j)}, \Delta\mathbf{B}_{k-1}^{(i,j)} \right\rangle}{\left\|\Delta\mathbf{B}_k^{(i,j)}\right\|_{L^2} \left\|\Delta\mathbf{B}_{k-1}^{(i,j)}\right\|_{L^2}}. \tag{2}$$

It is possible to group the frames into static and dynamic shots by comparing the inter-frame correlation $\rho_k, k = 1, ..., L$, with a threshold λ_s:

$$\begin{cases} \rho_{\mathrm{k}} < \lambda_S : & \textit{dynamic shot} \\ \rho_{\mathrm{k}} > \lambda_S : & \textit{static shot} \end{cases} \tag{3}$$

Usually when frames are lost, the receiver plays the last decoded frame until a new correctly decoded one is received. In this case the inter-frame correlation presents a spiky behavior with values close to one in correspondence of the lost frames. It is important to underline that detection of such a behavior is not sufficient to identify a partial or total frame loss. In fact in the case of static scenes, consecutive frames present a high inter-frame correlation.

It is also important to be able to distinguish between frames that are affected by errors and the ones belonging to a static scene. This can be achieved by using a system for assessing the presence of jerkiness, that is the phenomenon that leads to perceive individual still images in a video and can be useful for identifying a set of frames with the same characteristics. After this analysis is performed, it is possible to create a degradation map computed by means of a two-step procedure based on a temporal and a spatial degradation analysis.

2.2. Local temporal analysis

The local temporal analysis is performed in two stages. The aim of the first one is to identify and to extract from each frame the blocks that are potentially affected by artifacts. This classification analysis is performed by classifying the blocks as:

- with medium content variations,
- affected by large temporal variations,
- with small content variations.

by comparing their temporal correlation $\rho_k^{B^{(i,j)}}$ with two thresholds, θ_l and θ_h.

Then, observing that correlations close to 1 may correspond either to a repeated block or to a block belonging to a static region, while values close to zero can occur in presence of either sudden content changes (usually after shot boundaries) or errors, while occurs. As can be noted in Eq. 4, the highest distortion value is assigned to blocks considered as *unchanged* from the previous frame, while zero distortion is assigned to blocks with *medium* content variation.

The corresponding temporal variability map $\mathbf{\Gamma}_k^C = \{\Gamma_k^{CB^{(i,j)}}\}$ is computed by comparing the inter-frame correlation of each block with two thresholds θ_l and θ_h:

$$\Gamma_k^{CB^{(i,j)}} = \begin{cases} 1, & if\ \rho_k^{B^{(i,j)}} < \theta_l \\ 0, & if\ \theta_l \le \rho_k^{B^{(i,j)}} \le \theta_h \\ 2, & if\ \rho_k^{B^{(i,j)}} > \theta_h \end{cases} \tag{4}$$

The selection of the two thresholds, θ_l and θ_h, is based on the following considerations. If the correlation between corresponding blocks belonging to consecutive frames is close to one, it may correspond either to a repeated block or to a block belonging to a static region. On the other hand, a correlation value close to zero is obtained when completely different content blocks are under test: this situation can occur when there is a sudden content change (usually found after shot boundaries) or if error occurs. As can be noted in Eq. 4, the highest distortion index is assigned to blocks considered as *unchanged* from the previous frame, while zero distortion index is assigned to blocks with *medium* content variation.

2.3. Spatial analysis

The blocks that in the temporal analysis are classified as potentially affected by packet loss undergo a spatial analysis. This phase consists in:

- static regions detection: it aims at verifying whether a high correlation between the current block $\mathbf{B}_k^{(i,j)}$ and the previous one $\mathbf{B}_{k-1}^{(i,j)}$ is due to the loss of a block or to the presence of a static region. To perform this task, for each block with $\Gamma_k^{CB^{(i,j)}} = 2$, it is checked if at least v among the surrounding blocks have been classified as unchanged. In case of positive result, the block is classified as belonging to a static region and its potential distortion index $\Gamma_k^{CB^{(i,j)}}$ is set to zero; $if|\{(p,q)|(p,q) \in N(i,j), and\Gamma_k^{CB^{(p,q)}} = 2\}| > vthen\Gamma_k^{CB^{(i,j)}} = 0$

- edge consistency check: it is used to verify the presence of edge discontinuities in block boundaries. Let E_l and E_r be the L^1 norms of the vertical edges respectively on the left and on the right boundary of the block, and with A_c, A_l and A_r the average values of the L^1 norms of the vertical edges inside the current block and of the left and right adjacent blocks. A block with $\Gamma_k^{CB^{(i,j)}} \neq 1$ is classified as affected by visible distortion if:

  $$\left|E_l - \frac{(A_c + A_l)}{2}\right| > \theta \quad or \quad \left|E_r - \frac{(A_c + A_r)}{2}\right| > \theta \tag{5}$$

  The same procedure is applied to the horizontal direction;

- repeated lines test: it is performed to detect frames that have been partially correctly decoded. Usually, when the packet loss affects an intra-frame encoded image, a portion of the frame is properly decoded while the remaining part is replaced with the last row correctly decoded.

  Let $\mathrm{f}_\mathrm{k}[i]$ be the i^{th} row of the k^{th} frame. Starting from the m^{th} line of the frame, the L^1 norm of the horizontal gradient component is computed and compared to a threshold λ_H. If

  $$\|\Delta f_k[i]\|_{\mathrm{L}^1} > \lambda_{\mathrm{H}} \tag{6}$$

  , the procedure is repeated on the previous line $(i-1)$ to check if consecutive lines are identical by comparing the L^1 norm of their difference with a threshold λ_V

  $$\|f_k[i] - f_k[i-1]\|_{\mathrm{L}^1} < \lambda_{\mathrm{V}}. \tag{7}$$

  This procedure is iterated until the test fails.

  After the repeated lines test has been performed, a binary map $\Gamma_k^{RLB^{(i,j)}}$ of [0,1] entries is created where '1' corresponds to a block belonging to a vertical stripes filled region and '0' otherwise.

2.4. Reference frame detection

The previous procedure allows to assess the presence of blocks belonging to the current frame, which are affected by distortions caused by packet loss. Nevertheless, due to error propagation, the impairment can propagate until an intra-frame encoded image (I-frame), is received. An I-frame is usually characterized by a low correlation with the previous frame and a high correlation with the following frame.

For this reason the k^{th} frame is classified as an I-frame if

$$\rho_{k-1} - \rho_k > 2\eta_P \quad and \quad \rho_{k+1} - \rho_k > 2\eta_S \tag{8}$$

and no more than P out of Q among the previous or the following frames are characterized with an overall distortion Γ_k greater than a threshold λ_I, that is:

$$\Gamma_k > Pin[k-P,k]and[k,k+Q]. \tag{9}$$

The decision thresholds are adapted to the current video content. In particular, η_P and η_S are proportional to the mean absolute differences of the correlation coefficients in the intervals $[k - M_l, k]$ and $[k, k + M_h]$, i.e.:

$$\eta_P = \frac{1}{M_l} \sum_{h=k-M_l+1}^{k} |\rho_h - \rho_{h-1}| \tag{10}$$

and

$$\eta_S = \frac{1}{M_h} \sum_{n=k+1}^{k+M_h} |\rho_n - \rho_{n-1}| \tag{11}$$

M_l is selected to guarantee that the time interval needed for the adaptation of η_P starts at the frame following the last detected I-frame. When processing the k^{th} frame, no information about location of next I-frames is available and the length of the interval employed for the adaptation of η_S is considered constant. When the time interval between to I-frames is less than M_h, only the I-frame with the lowest correlation with the previous frame is retained.

3. EXPERIMENTAL RESULTS

To assess the effectiveness of the proposed quality metric, several tests have been performed. In more details, a set of color videos has been transmitted over an IP channel affected by increasing values of packet loss: 0.1%, 0.5%, 0.9%, 1.3%, 3% and 5%.

The quality of the received videos has been compared to the quality of the original videos by means of the proposed NR metrics and of the NTIA-VQM FR video quality metric. The considered database of video test is composed by three color videos at standard definition resolution (720 x 544 pixels) freely available in [8]. The selected videos present different content to allow the verification of the effectiveness of the quality metric depending on the video semantic. Figure 1 shows sample frames extracted from the videos in the database.

In order to perform the mentioned analysis, an experimental setup has been developed. This is constituted by one streaming source, a network segment, and one receiver. The network segment has been modeled by means of an open-source network emulator: NETEM (NETwork EMulator) [9]. The emulator has been used for introducing packet losses in the incoming streams in order to simulate the behavior of a real network based on the best-effort paradigm. Each considered media stream has been processed in order to simulate a set of increasing Packet Loss Rates.

The parameters of NR metric have been identified by means of non linear regression between the FR metric and the number of estimated corrupted blocks.

Specifically, let:

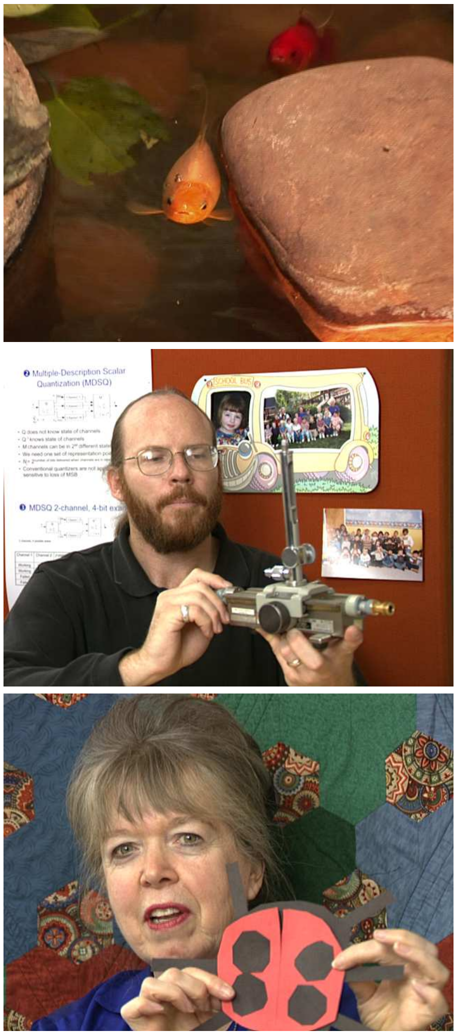
Multiple-Description Scalar Quantization (MDSQ)
SCHOOL BUS
MDSQ 2-channel, 4-bit exa

Fig. 1. Sample frames from the videos used in the experimental tests.

- $\mathbf{N}^{CCB} = [N_R^{CCB}, N_G^{CCB}, N_B^{CCB}]^T$ be the column vector of the number of clustered corrupted blocks for the three color components,
- $\mathbf{N}^{ICB} = [N_R^{ICB}, N_G^{ICB}, N_B^{ICB}]^T$ be the column vector of the number of isolated corrupted blocks for the three color components,
- $\mathbf{N}^{RL} = [N_R^{RL}, N_G^{RL}, N_B^{RL}]^T$ be the column vector of the number of repeated lines for the three color components,
- ρ^{LOSS} be the packet loss rate;

the NR metric VQM_{NR} is computed as follows:

$$VQM_{NR} = \left[\beta^{CCB}\mathbf{N}^{CCB} + \beta^{ICB}\mathbf{N}^{ICB} + \beta^{RL}\mathbf{N}^{RL} + \beta^{LOSS}\rho^{LOSS}\right]^{1/2} \quad (12)$$

where

$$\begin{aligned} \beta^{CCB} &= [-0.0002, -0.0013, 0.0014], \\ \beta^{ICB} &= [0.0043, 0.0202, 0.0144], \\ \beta^{RL} &= [0.0068, 0.0051, 0.0002], \\ \beta^{LOSS} &= 0.0489 \end{aligned}$$

are the (row vector) regression coefficients.

In Figure 2 the plot of NR-metric vs. the FR-metric VQM are reported for the selected videos. The processed video sequences have been concatenated with increasing PLRs.

4. CONCLUSIONS

In this paper a No Reference metric for assessing the quality of color video transmission has been presented. Preliminary tests show the effectiveness of the proposed system.

The scores collected by this metric, in evaluating impaired videos, have been compared with the ones gathered with the Full Reference NTIA-VQM. As demonstrated by the comparison, there are still many open issues to be investigated. In comparing NR and FR metrics one key factor is represented by the temporal realignment algorithm. In fact, in presence of highly textured backgrounds, severe frame losses, and medium to high compression ratios, at least our implementation of the NTIA-VQM algorithm does not provide reliable estimates of the variable delay among the original and processed video. This in turn implies a bias in the estimated FR metrics induced by the wrong selection of the reference frame to be used for the comparison. Since, in Fig.2 processed video sequences have been concatenated with increasing PLR, this effects is visible for frame number greater than 500.

As a general remark, the influence of the adopted keyframe detection algorithm should be investigated. Moreover the choice of the parameter λ_s should be based on the fact that many studies show that human attention is attracted by objects whose movement is relevant with respect to the other points in the scene. λ_s should probably be adapted to the relative motion of the surrounding areas.

Another aspect to be further investigated is the influence of the adopted error concealment technique implemented in the decoder. We noticed that the newer version of VLC is able to mask in a more effective way some transmission errors like the presence of isolated blocks. This means that in the future the parameters we considered in the proposed metric may be different according to the improvements achieved in the field of error concealment techniques.

5. REFERENCES

[1] M. Nauge, M. C. Larabi and C. Fernandez, ”A reduced-reference metric based on the interest points in color images”, *Procs. of Picture Coding Symposium (PCS)*, pp. 610-613, 2010.

[2] A. Maalouf and M. C. Larabi, ”A no-reference color video quality metric based on a 3D multispectral wavelet transform”, *Procs. of Second International Workshop on Quality of Multimedia Experience*, pp. 11-16, 2010.

[3] C.T.E.R. Hewage and M.G. Martini, ”Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3D video”, *Procs. of 17th IEEE International Conference onImage Processing (ICIP)*, pp. 4017 - 4020, 2010.

[4] X. Jianhua, L. Junli, W. Yongsheng, J. Gangyi, and C. Gang, ”A Color Video Quality Assessment Based on Edge-Color Distortion”, *Procs. of Symposium on Photonics and Optoelectronic*, pp. 1-4, 2010.

[5] Z. Fuqiang, L. Junli, C. Gang, and M. Jiaju, ”Assessment of Color Video Quality Based on Quaternion Singular Value Decomposition”, *Procs. of Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, pp. 7-10, 2009.

[6] Z. Wang and E.P. Simoncelli, ”Reduced-reference image quality assessment using a wavelet-domain natural image statistic model”, *Procs. of Human Vision and Electronic Imaging X*, vol. 5666, 2005.

[7] A. Neri, M. Carli, M. Montenovo, A. Perrot, and F. Comi - No reference quality assessment of Internet multimedia services,” *Proc. 14th European Signal Processing Conference (EUSIPCO-2006)*, 2006.

[8] ”The Consumer Digital Video Library”, available at *URL:http://www.cdvl.org/*

[9] ”Network Emulation with NetEm”, available at *URL:http://www.linuxfoundation.org/en/Net:Netem*, 2005.

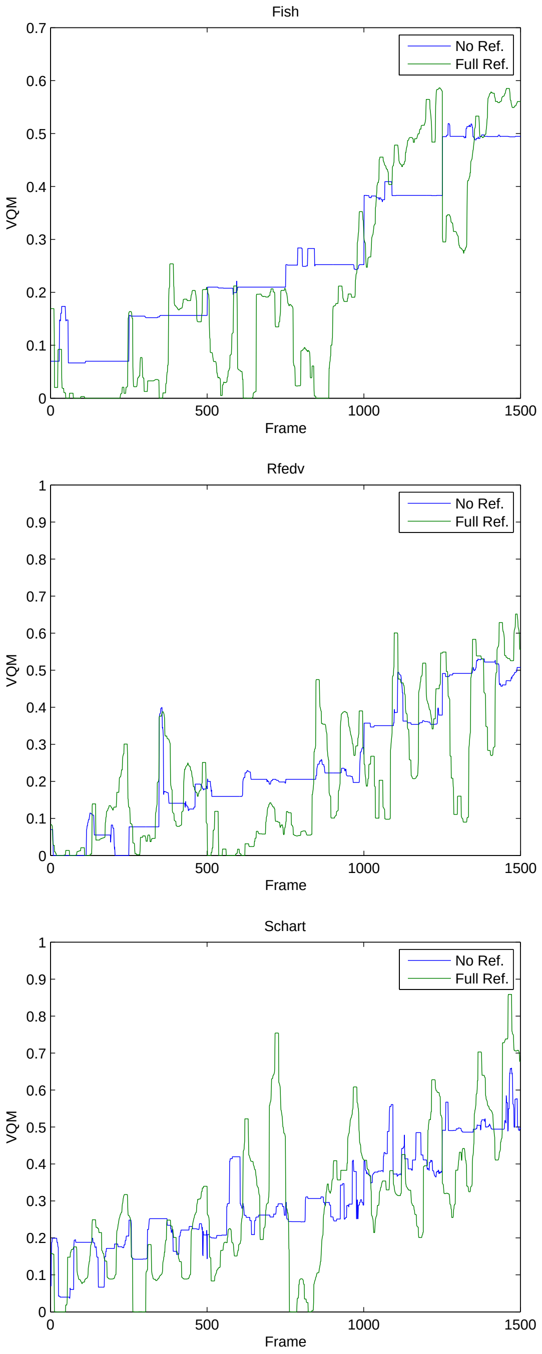

Fig. 2. NR-metric vs. the FR-metric VQM.