EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

No reference quality assessment for MPEG video delivery over IP

Federica Battisti*, Marco Carli and Alessandro Neri

Abstract

Video delivering over Internet protocol (IP)-based communication networks is widely used in the actual information sharing scenario. As well known, the best-effort Internet architecture cannot guarantee an errorless data delivering. In this paper, an objective no-reference video quality metric for assessing the impact of the degradations introduced by video transmission over heterogeneous IP networks is presented. The proposed approach is based on the analysis of the inter-frame correlation measured at the output of the rendering application. It does not require any information on errors, delays, and latencies affecting the links and on the 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 by using human visual system (HVS)-inspired full reference metrics.

1 Introduction

In the last decade, a fast market penetration of new multimedia services has been experienced. UMTS/CDMA2000-based videotelephony, multimedia messaging, video on demand over Internet, and digital video broadcasting are a growing share of nowadays' economy. The large-scale spreading of portable media players indicates the increasing end user demand for portability and mobility. At the same time, media sharing portals and social networks are mostly based on user-contributed video content, showing a different perspective of user relation with digital media: from being media consumers to being part of content creation, distribution, and sharing. This information sharing evolution presents many challenges; in particular, Internet protocol (IP)-based multimedia services require a transparent delivery of media resources to end users that has to be independent on the network access, type of connectivity, or current network conditions [1]. Furthermore, the quality of the rendered media should be adapted to the end users' system capabilities and preferences. An overview of the future challenges of video communication are addressed in [2] and in [3]. As can be noted, there are many factors that can prejudice the quality of the delivered video.

For this reason, the effectiveness of each video service must be monitored and measured for verifying its compliance with the system performance requirements, for benchmarking competing service providers, for service monitoring and automatic parameters setting, network optimization, evaluation of customer satisfaction, and adequate price policies setting.

To this aim, *ad hoc* designed tools have to be employed. In fact, the systems adopted for assessing the performances of traditional voice-based transmission systems are usually inadequate for evaluating the quality of multimedia data. In fact, a single training sequence comparison or the bit-wise error collection, which is used for measuring the quality of the received signal in the voice-based communication model, is not able to catch the dynamic feature of a video stream and its correlation with the overall experience of the final user.

To cope with this task, the research on video quality assessment has mainly been focused on the development of objective video quality metrics able to mimic the average subjective judgment. This task is of difficult solution being dependent on many factors as:

- Video characteristics and content: size, smoothness, amount of motion, of sharp details, and the spatial and temporal resolution
- Actual network conditions: congestion, packet loss, bit error rate, time delay, time delay variation

*Correspondence: federica.battisti@uniroma3.it
Department of Engineering, Universita' degli Studi Roma TRE, Rome 00146, Italy

- Viewer's condition: display size and contrast, processing power, available memory, viewing distance
- Viewer's status: feeling, expectations, experience, involvement

These factors are often difficult or impossible to be measured, especially in real-time communication services. Furthermore, each factor has a different impact on the overall perceived distortion whose visibility strongly depends on the content (e.g., salt and pepper noise can be not noticeable in highly textured areas of a frame while it can become highly visible in the next uniform frame).

In this contribution, a no-reference metric for assessing the degradations introduced by transmission of an original video, encapsulated in a MPEG2 TS, over a heterogeneous IP network is presented. Variable channel conditions may cause both isolated and clustered packet losses resulting in data and temporal integrity loss at the decoder side. This could lead to the impossibility of decoding isolated or clustered blocks, tiles, and even entire frames. Considering the continuous increase of computing power of both mobile and wired terminals, a wide spread of error concealment techniques aimed at increasing the perceived quality can be expected. The proposed system can be considered blind with respect to errors, such as delays and latencies that affected the link, and to the concealment strategies implemented at the decoders to face the potential quality loss. The proposed method is based on the evaluation of perceived quality of video delivered by a packet switched network. These networks are characterized by loss, delayed, or out-of-sequence delivery of packets. The proposed metric is therefore valid for mobile networks exploiting IP, i.e., UMTS, long-term evolution (LTE), and LTE Advanced.

The rest of the paper is organized as follows. In Section 2, the state-of-the-art research addressing video quality metrics is addressed, while in Section 3, the model of the proposed metric is presented. In Section 4, the evaluation of the video distortion is described. In Section 5, the results of the performed experiments are presented based on two types of evaluation: comparison with the National Telecommunications and Information Administration (NTIA) scores and comparison with subjective tests. Finally, in Section 6, the conclusions are drawn.

2 State of the art

A set of parameters describing the possible distortion in a video has been defined and classified in 'ITU-T SG9 for RRNR project' [4]. Here, temporal distortion, temporal complexity, blockiness, blurring, image properties, activity, and structure distortion are independently evaluated and then linearly combined with the aim of reliably fitting the measured mean opinion score (MOS) with the calculated MOS. This general model has been recently improved in [5] taking into account the dynamic range distortion. Objective quality metrics can be classified according to different criteria [6-8]. One of this is the amount of side information required to compute a given quality measurement; depending on this, three classes of objective metrics can be described:

- *Full reference metrics (FR)*: the evaluation system has access to the original media. Therefore, a reliable measure for video fidelity is usually provided. A drawback presented by these metrics is that they require the knowledge of the original signal at the receiver side.
- *Reduced reference metrics (RR)*: the evaluation system has access to a small amount of side information regarding the original media. In general, certain features or physical measures are extracted from the reference and transmitted to the receiver as side information to help the evaluation of the video quality. The metrics belonging to this class may be less accurate than the FR metrics, but they are also less complex, and make real-time implementations more affordable.
- *No-reference metrics (NR)*: the evaluation system has no knowledge of the original media. This class of metrics is the most promising in the context of video broadcast scenario since the original images or videos are in practice not accessible to end users.

The FR class includes all methods based on pixel-wise comparison between the original and the received frame. Among them, the most relevant example of objective metric is the peak signal-to-noise ratio (PSNR), which is widely used to perform a fast and simple quality evaluation. It is based on the computation of the ratio between the mean square error (MSE) between the image to be evaluated and the reference image, and the maximum range of pixel values. Even if the metrics belonging to this class are easy to compute, they do not always well correlate with quality as perceived by human observers. In fact, these metrics do not consider the masking effects of the human visual system (HVS) and each pixel degradation contributes to the overall error score even if the error is not perceived by a human subject.

A novel and effective approach has been proposed in [9] with the NTIA-video quality metric (VQM) which combines in a single score the perceptual impact of different video artifacts (block distortion, noise, jerkiness, blurring, and color modifications). The NTIA-VQM is a general purpose video quality model designed and tested in a wide range of quality and bit rates. It is based on a preprocessing consisting in spatial and temporal alignment of reference and impaired sequences, region extractions, gain, and offset correction. Following, feature extraction, spatio-temporal parameter estimation, and local indexes polling are performed for computing the overall quality score. The features considered in VQM are extracted from the spatial gradients of the luminance and chrominance components and from measuring contrast and temporal information extracted from the luminance component only. It has been validated by exploiting extensive subjective and objective tests. The high correlation with MOS shown in the performed subjective tests is the reason for the wide use of this metric as a standard tool (ANSI) for FR video assessment[10].

Similarly, moving picture quality metric (MPQM) [11] and its colored version color moving picture quality metric [12] are based on the assumption that the degradation of the quality of a video is strictly related to the visibility of the artifacts. In more detail, it is based on the decomposition of the original and of the impaired videos in visual channel, and on the distortion computation performed by considering the masking effect and the sensitivity contrast function. The main limitation of this, and similar methods based on error sensitivity model, is in the simple (often linear) model adopted for the HVS, which badly approximates the complex, non-linear, and still partially disclosed vision system.

A different approach is based on the hypothesis that the human brain is very efficient in extracting the structural information of the scene rather than the error information. Therefore, as proposed by Wang et al. in [13], a perceptual-based metric should be able to extract information about structural distortions. The SSIM metric, proposed in [13], shows a good correlation with the subjective scores obtained by campaigns of subjective tests. Other classical FR metrics inspired by the HVS are in the works by Wolf et al. [14] and Watson et al. [15], while a survey of available FR video quality metrics can be found in [16]. It is worth noticing that for an effective frame-to-frame comparison, both the original video and the one under test must be synchronized.

Recently, Shnayderman et al. [17] compared the singular value decomposition coefficients of the original and the coded signal, while in [18], the authors computed the correlation between the original and the impaired images after a 2D Gabor-based filter bank, based on the consideration that the cell of the visual cortex can be modeled as 2D Gabor functions. As can be noticed, these metrics can also be applied to videos by computing a frame by frame quality evaluation.

As previously stated, the usability of FR metrics is often limited in real scenarios due to the need for availability of the original video. Nevertheless, being the perceived quality dependent on the content, it cannot be directly inferred from the knowledge of parameters such as channel reliability and temporal integrity. To partially overcome this problem, RR and NR quality metrics have been devised.

With respect to FR metrics, only few attempts of RR and NR metrics have been presented in literature. Among RR ones, Carnec et al. in [19] present a still image metric based on color perception and masking effects, resulting in a small overhead. Blocking, blurring, ringing, masking, and lost blocks are linearly combined in [20] for a frame by frame comparison. Wang and Simoncelli in [21], based on the frequency distribution characteristics of natural images, proposed to use the statistics of the coded image to predict the visual quality.

Different approaches are proposed by Kanumuri et al. in [22]: the RR metric is based on a two-step approach. The information gathered from the original, received, and decoded video are used in a classifier whose output will be used in the evaluation of artifact visibility performed on a decision tree trained by subjective tests. Similarly, in [23], a general linear model (GLM) is adopted for estimating the visibility threshold of packet loss in H.264 video streaming. In [24], GLM is modified by computing a saliency map, for weighting the pixel-wise errors, and by taking into account the influence of the temporal variation of saliency map and packet loss. The results show that if the HVS features are considered, the prediction of subjective scores is improved.

Finally, a novel approach is represented using different communication systems for delivering information on the original media. In this classification are, for example, the data hiding-based RR metrics. In these approaches, a thumbnail of the original frame [25-28], a perceptually weight watermark [29], or a particular image projection [30] are used in the quality evaluation as fingerprint of the original frame quality. The main vulnerability of these methods is in the robustness of the watermarking method. In fact, any alteration, wanted or not, of the inserted data may strongly affect the objective assessment.

The need for the reference video or for partial information about it is a considerable drawback in many real-time applications. For this reason, the design of effective NR metrics is a big challenge. In fact, although human observers are able to assess the quality of a video without using the reference, the creation of a metric that could mimic such a task is difficult and, most frequently, it results in a loss of performances in comparison to the FR/RR approaches. To achieve effective evaluations, many existing NR metrics estimate the annoyance by detecting and estimating the strength of common artifacts (e.g., blocking and ringing).

NR techniques are the most promising because their final score can be considered, for an ideally perfect metric, as an absolute quality value, independent from the knowledge of the original content. Few metrics have been designed for the evaluation of impairments due to single artifacts as blockiness [31], blurriness [32], and jerkiness [33].

Different strategies have been proposed to evaluate the impact of impairments caused by compression algorithms and transmission over noisy channels. These can be classified according to the parts of the communication channel that are involved:

- Source coder errors: MSE estimation due to compression [34], example-based objective reference in [35], motion-compensation edge artifacts in [36];
- Variable delay [37,38];
- Packet loss effects [39-41]. In [41], the NR metric is based on the estimation of mean square error propagation among the group of picture (GOP) in motion compensation-based encoded videos. The idea is to consider the motion activity in the block as initial guess of the distortion caused by the initial packet loss;
- Bitstream-based video quality metric: in these systems, several bitstream parameters, such as motion vector length or number of slice losses, are used for predicting the impairments visibility in MPEG-2, SVC [42], or H.264 [43,44] and HD H.264 [45] videos. Recently, the bitstream metric proposed in [46] has been modified with pixel-based features to cope with HDTV stream [47];
- Rendering system errors: [48-50].

Other examples include the works presented by Webster et al. [51] and Brétillon et al. [52]. The estimation of the pattern of lost macroblocks based on the knowledge of the decoded pixels is used as input to a no-reference quality metrics for noisy channel transmission. The metrics by Wu and Yuen and Wang et al. estimate quality based on blockiness measurements [31,53], while the metric by Caviedes and Jung takes into account measurements of five types of artifacts [54]. Recently, in [55], a methodology for fusing metrics feature for assessing video quality has been presented. This work has also been adopted in the ITU-T Recommendation P.1202.2.

3 The NR procedure

In the following, the motivations behind each step of the NR procedure are briefly described and then detailed in Subsections 3.1 to 3.4. As previously stated, channel errors and end-to-end jitter delays can produce different artifacts on the received video. The effect of these errors can have a dramatic impact on the quality perceived by users since the loss of a single packet can result in a corrupted macroblock. Corrupted information can affect both spatial (to neighboring blocks) and temporal (over adjacent frames) quality due to the predictive, motion-compensated coding scheme adopted by most of existing video coders. The visual impact of these errors strictly depends on the effectiveness of the decoder scheme and on the concealment strategy that is adopted.

In order to recover transmission errors, decoders can exploit several strategies depending on the error resilience or concealment techniques adopted in the communication scheme. Error resilience is based on the addition of redundant information at the encoder side for allowing the decoder to recover some transmission errors: the drawback is the increase in the amount of transmitted data. On the other hand, error concealment is a post-processing technique in which the decoder tries to mask the impairments caused by packet losses and bit stream errors that have been detected but not corrected. In this case, even if the quality of the recovered data is usually lower than the original one, the system does not require encoder/decoder modification or extra information delivering. Several concealment techniques have been proposed in literature whose effectiveness increases with complexity. The simplest proposed strategy consists in filling the missing areas with a constant value or with information extrapolated by considering the last correctly decoded block. More sophisticated techniques apply prediction/interpolation of the lost block(s) by exploiting spatial and temporal redundancy [28]. Concealment effectiveness is largely affected by the spatial and temporal extension of the missing information, with best performances obtained in the case of small clusters and isolated blocks. An example of visual artifacts on a test sequence 'Brick', when transmitted on a noisy channel affected by a 15% packet loss, is shown in Figure 1. When an error affects the entire frame or a large part of it, the decoder may decide to completely drop the corrupted frame and to *freeze* the last error-free video frame until a new valid frame is correctly decoded. In this case, the perceived quality of the played video will depend on the dynamics of the scene. In fact, although only error-free frames are played, the motion of objects composing the scene may appear unnatural, due to its stepwise behavior (jerkiness effect). The dropping mechanism can also be caused by a playback system that is not fast enough to decode and display each video frame at full nominal speed. It is worth noticing that the same experience is perceived in the presence of *frame freezing* artifacts or by repetition of the same frame.

The NR metric proposed in this paper is independent from the error concealment techniques implemented in the video player; however, since frame repetition is a very common concealment method, here, the assessment of the quality loss produced by *freezing* the video in correspondence of frame losses is specifically addressed. More in details, before applying the NR jerkiness metric proposed by the authors in [56], the played sequence is analyzed in order to detect the presence of repeated frames.

To this aim, the rendered sequence is first partitioned into static and dynamical shots, on the basis of the amount

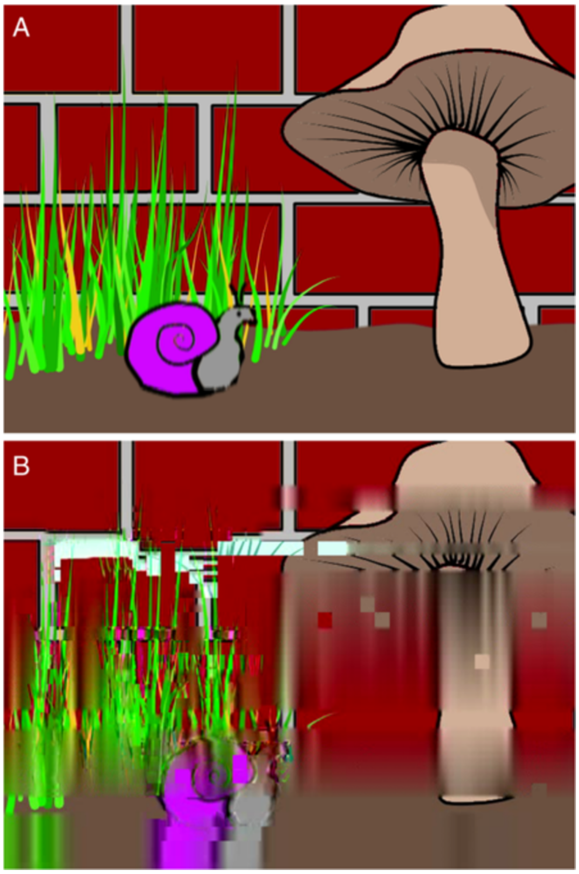

Figure 1 Impact of a noisy channel (15% PLR) on the transmission of a test video sequence. In particular, by considering the rendered frame **(A)** versus the original one **(B)**, several artifacts can be noted: isolated blocks, repeated lines, blurring and wrong color reconstruction.

of changes between consecutive frames. Next, the shots classified as static are evaluated in order to detect if the identified small amount of changes corresponds to a real static scene or to the freeze of entire frames or part of them. At the same time, the dynamical shots are tested to verify the presence of isolated and clustered corrupted blocks. These analyses result in temporal variability and spatial degradation maps that are used to assess the video quality by evaluating the overall distortion as shown in Figure 2. In the following, the details of the proposed system are presented.

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

As previously described, the first step of the NR procedure is the grouping of frames in dynamic or static shots based on a temporal analysis.

Let $\mathbf{F} = \{\mathbf{F}_k, k = 1, \ldots, L\}$ denote a video sequence composed by L frames of $m \times n$ pixels.

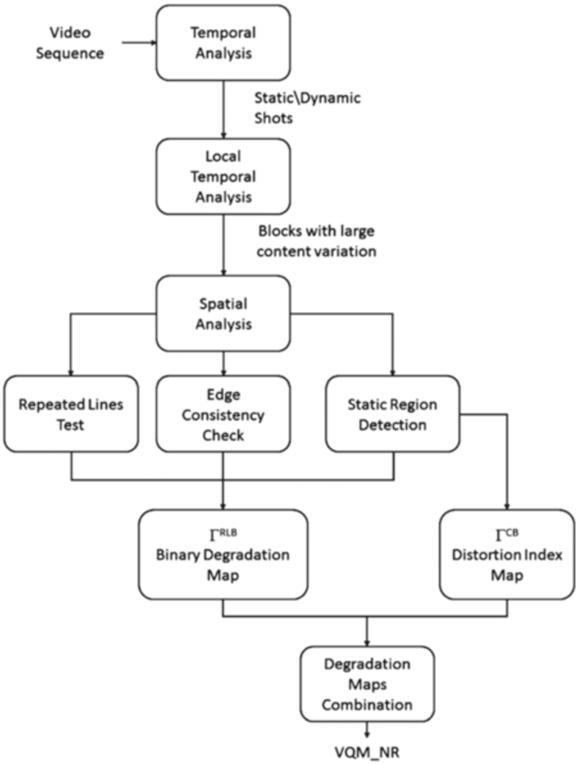

Figure 2 Block diagram of the proposed metric.

The generic kth frame can be partitioned in $N_r \times N_c$ blocks $\mathbf{B}_k^{(i,j)}$ of $r \times c$ pixels with top left corner located in (i,j). Let $\bar{\mathbf{F}}_k$ be the mean luminance value for the kth frame and $\bar{\mathbf{B}}_k^{(i,j)}$ the mean luminance value of block $\mathbf{B}_k^{(i,j)}$. Let $\Delta\mathbf{F}_k = \mathbf{F}_k - \bar{\mathbf{F}}_k$ and $\Delta\mathbf{B}_k^{(i,j)} = \mathbf{B}_k^{(i,j)} - \bar{\mathbf{B}}_k^{(i,j)}$ denote the deviation of the luminance of the kth 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 kth and the $(k-1)$th frames is defined as:

$$\rho_k = \frac{\left\langle \Delta\mathbf{F}_k, \Delta\mathbf{F}_{k-1} \right\rangle}{\|\Delta\mathbf{F}_k\|_{L_2} \left\|\Delta\mathbf{F}_{k-1}\right\|_{L_2}}, \tag{1}$$

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 dynamical shots by comparing the inter-frame correlation $\rho_k, k = 1, \ldots, L$, with a threshold λ_s:

$$\begin{cases} \rho_{\mathrm{k}} < \lambda_S : \text{ dynamical shot} \\ \rho_{\mathrm{k}} > \lambda_S : \quad \text{static shot} \end{cases} \tag{3}$$

where the threshold λ_s is set to the equal error rate (EER) between the classification of a static block as *dynamic* and vice versa.

As illustrated in Figure 3, the inter-frame correlation presents a spiky behavior with values close to one in correspondence of *frozen* frames. It is important to underline that the 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.

Therefore, it is 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. In fact, jerkiness is the phenomenon that leads to perceive a video as consisting of a sequence of individual still images. In this contribution, we adopt the approach that has been presented in [56].

After the segmentation into dynamic and static shots, the task of quality evaluation gets easier. In fact, for static shot sequences, it is possible to evaluate the quality of the first frame and to extend the obtained score to the frames belonging to the static cluster. In this way, a degradation map is computed for the first frame (that can still be affected by artifacts) and is inherited by the frames belonging to the same static shot. When dealing with the distortion associated to isolated and clustered impaired blocks, it is estimated by means of a two-step procedure

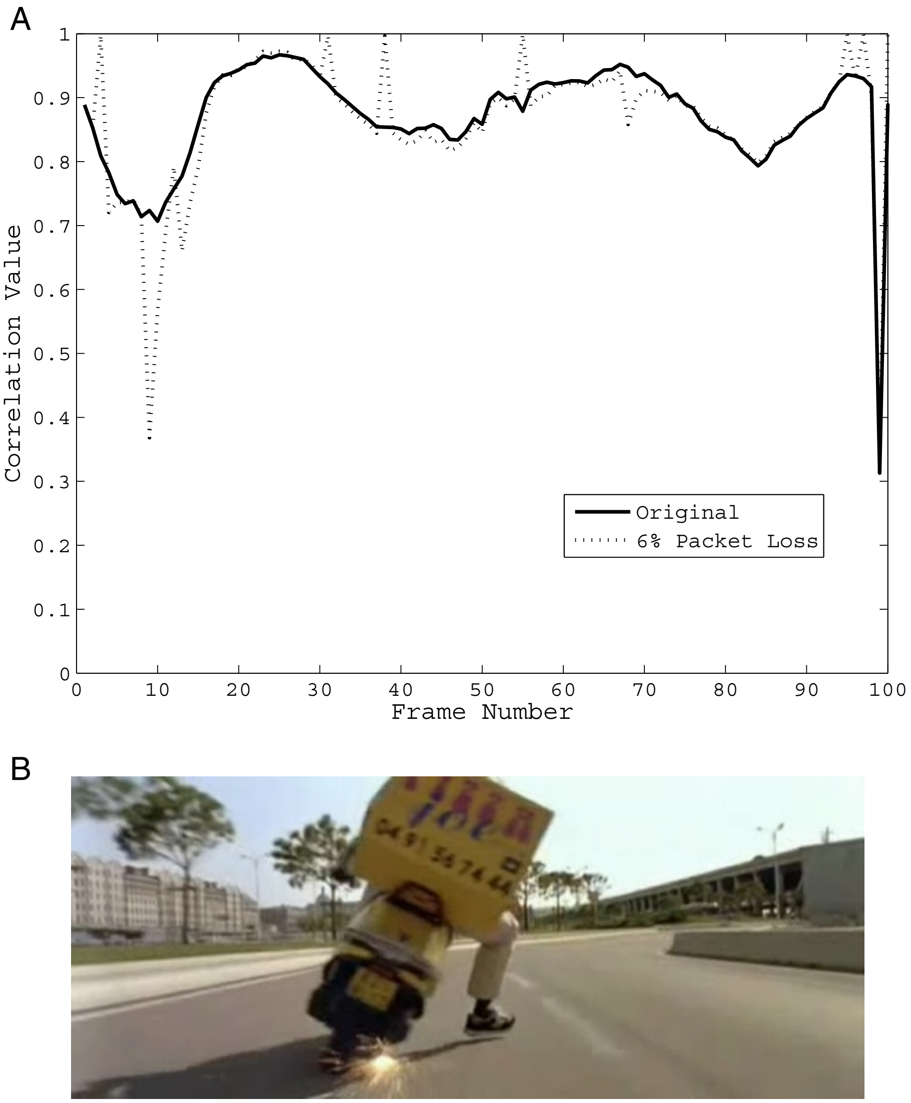

Figure 3 Test video sequence affected by 6% packet loss. (A) The normalized interframe correlation among the first 100 frames extracted from the original video sequence *Taxi* and the first 100 frames extracted from the same sequence affected by 6% packet loss. **(B)** One frame extracted by the video sequence.

based on temporal and spatial degradation analysis. In the following, this will be referred to as 'degradation map' computation.

3.2 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 analysis is performed by classifying the blocks as:

- With medium content variations
- Affected by large temporal variations
- With small content variations

depending on their temporal correlation $\rho_k^{B^{(i,j)}}$.

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

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

The selection of the two thresholds, θ_l and θ_h, is performed based on the assumption that:

- The correlation, between corresponding blocks belonging to consecutive frames, is close to one in the presence of a repeated block or of a block belonging to a static region.
- The correlation value is close to zero in case of a sudden content change (usually occurring after shot boundaries) or in the presence of an error.

For this reason, as can be noted in Equation 4, the highest temporal variability index is assigned to blocks considered as *unchanged* from the previous frame, while zero distortion index is assigned to blocks with *medium* content variation. In more details, let us define with probability of false alarm (P_{fa}) the probability of detecting the repeated blocks as affected by errors in the absence of distortion and with probability of miss detection (P_{md}) the probability of considering as unaltered a frame in the presence of errors. The two thresholds, θ_l and θ_h have been selected in order to grant

$$|P_{\mathrm{fa}} - P_{\mathrm{md}}| < \varepsilon_1$$

where ε_1 has been experimentally determined, during the training phase, by comparing the performances achieved by the temporal analysis algorithm with the scores provided by a group of video quality experts in an informal subjective test.

3.3 Spatial analysis

The blocks classified as potentially affected by packet loss during the temporal analysis phase undergo a spatial analysis. The spatial analysis is performed in several steps:

- 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 single or multiple blocks or to a static region. To perform this task, for each block with $\Gamma_k^{\mathrm{VB}^{(i,j)}} = 2$, it is checked if at least v among the surrounding blocks present a strong temporal correlation. In case of positive result, the block is classified as belonging to a static region and its potential distortion index $\Gamma_k^{\mathrm{CB}^{(i,j)}}$ is set to zero. The parameter v has been identified by experimental test. Practically, a set of expert viewers has been presented with a set of short videos presenting different content situations affected by increasing blocking artifacts. The parameter v has been selected as the one resulting in the highest correlation between the people score and the algorithmically performed spatial analysis block. That is:

$$\Gamma_k^{\mathrm{CB}^{(i,j)}} = \begin{cases} 0 & \text{if } |\{(p,q)| \ (p,q) \in N(i,j), \text{and } \Gamma_k^{\mathrm{VB}^{(p,q)}} = 2\}| > \nu \\ \Gamma_k^{\mathrm{VB}^{(i,j)}} : & \text{otherwise} \end{cases} \tag{5}$$

- Edge consistency check: the presence of edge discontinuities in block boundaries can be used as an evidence of distortions. For the sake of simplicity, we detail the procedure for the case of gray scale images. It can be easily extended to the color case by evaluating separately the edge consistency for each color component. 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 $\Gamma_k^{\mathrm{CB}^{(i,j)}} \neq 0$ is classified as affected by visible distortion if:

$$\left| E_{\mathrm{l}} - \frac{(A_{\mathrm{c}} + A_{\mathrm{l}})}{2} \right| > \theta \quad \text{or} \quad \left| E_{\mathrm{r}} - \frac{(A_{\mathrm{c}} + A_{\mathrm{r}})}{2} \right| > \theta \tag{6}$$

where the threshold θ has been defined on the basis of experimental trials. In particular, it corresponds to just noticeable distortion collected evaluated for the 90% of subjects. The same procedure is then applied to the horizontal direction. If the block edges are consistent (i.e., no visible distortion has been detected along horizontal and vertical directions), $\Gamma_k^{\mathrm{CB}^{(i,j)}}$ is reset to 0.

- Repeated lines test: it is performed to detect frames that have been partially correctly decoded. A very common concealment strategy is based on the fact that when the packet loss affects an intra-frame encoded image, and a portion of the frame is properly decoded, the remaining part is replaced with the last row correctly decoded. As can be noted in Figure 4, the procedure results in a region containing vertical stripes.

  Let $f_k[i]$ be the ith row of the kth frame. Starting from the mth line of the frame, the L_1 norm of the horizontal gradient component is computed and compared to a threshold λ_H. If

  $$\left\| \Delta f_k[i] \right\|_{\mathrm{L}_1} > \lambda_{\mathrm{H}}, \tag{7}$$

  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

  $$\left\| f_k[i] - f_k[i-1] \right\|_{\mathrm{L}_1} < \lambda_{\mathrm{V}}. \tag{8}$$

  This procedure is iterated until the test fails, thus meaning that there is a different information carried out by consecutive lines. After the repeated lines test has been performed, a binary spatial degradation map, $\Gamma_k^{\mathrm{RLB}^{(i,j)}}$ of [0,1] entries, is created where '1' corresponds to a block belonging to a vertical stripes region and '0' otherwise. The two thresholds, λ_{V} and λ_{H} have been set after a training process with a pool of experts trying to match the subjective impression of repeated lines.

3.4 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. Figure 5 shows the normalized inter-frame correlation of a sequence extracted from an action movie. As can be noted, an I-frame is usually characterized by a low correlation with the previous frame and a high correlation with the next frame. This behavior is always verified unless the same scene is shown for a long period.

Let us denote with ν_k^{CB} the number of corrupted blocks, i.e.,

$$\nu_k^{\mathrm{CB}} = \quad |\{\Gamma_k^{\mathrm{CB}^{(i,j)}} \neq 0\}|. \tag{9}$$

Then, the kth frame is classified as an I-frame if

$$\rho_{k-1} - \rho_k > 2\eta_P \quad \text{and} \quad \rho_{k+1} - \rho_k > 2\eta_S \tag{10}$$

and no more than P_p out of the Q_p previous frames and no more than P_n out of the Q_n following frames are characterized by a number ν_k^{CB} of blocks with inconsistent edges exceeding a threshold λ_I.

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.:

$$\eta_P = \frac{1}{M_l} \sum_{h=k-M_l+1}^{k} \left| \rho_h - \rho_{h-1} \right| \tag{11}$$

and

$$\eta_S = \frac{1}{M_h} \sum_{n=k+1}^{k+M_h} |\rho_n - \rho_{n-1}|. \tag{12}$$

The value M_l is selected to guarantee that the time interval needed for the adaptation of η_P starts at the frame following the last correctly detected I-frame. When processing the kth 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 to be constant. When the time interval between two I-frames is less than M_h, only the I-frame with the lowest correlation with the previous frame is retained.

4 Distortion map evaluation

The evaluation of the video quality metric VQM$_{\mathrm{NR}}$ is based on the degradation index maps $\Gamma_k^{\mathrm{RLB}^{(i,j)}}$ and $\Gamma_k^{\mathrm{CB}^{(i,j)}}$

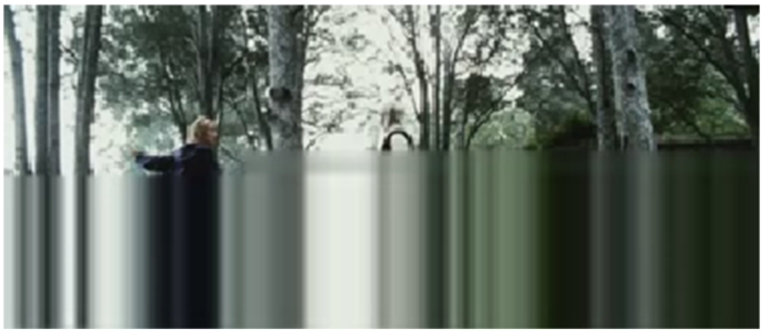

Figure 4 **Frame affected by vertical stripes.**

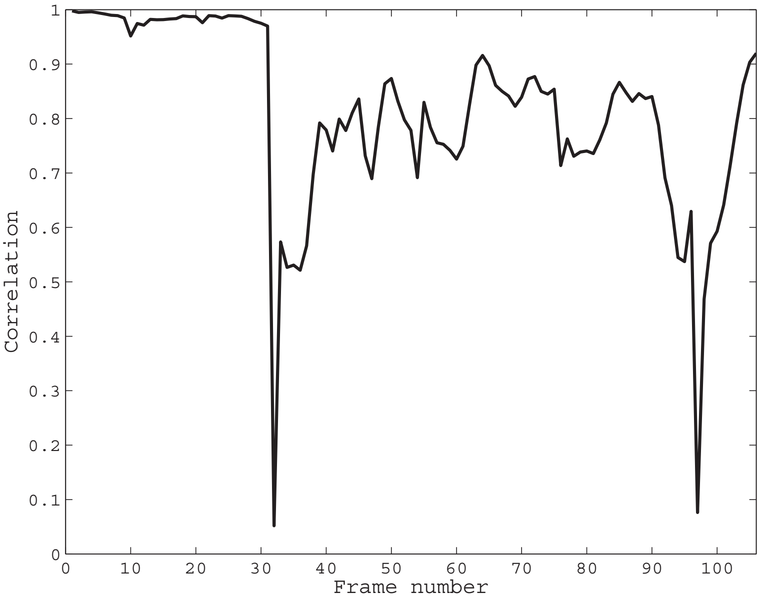

Figure 5 Normalized inter-frame correlation of the original sequence.

whose computation has been illustrated in the previous section. To account for the error propagation induced by predictive coding, a low-pass temporal filtering is applied to the degradation index maps. To this aim, let D_{k-1} denote the generic distortion map at time $(k-1)$; at time k, the distortion map D_k of frames belonging to dynamical shots is evaluated as follows:

$$\mathbf{D}_k^{\mathrm{CB}} = \mu\left[\Gamma_k^{\mathrm{CB}} + \varphi\left(\rho_k\right)\mathbf{D}_{k-1}^{\mathrm{CB}}\right]; \tag{13}$$

$$\mathbf{D}_k^{\mathrm{RLB}} = \mu\left[\Gamma_k^{\mathrm{RLB}} + \varphi\left(\rho_k\right)\mathbf{D}_{k-1}^{\mathrm{RLB}}\right] \tag{14}$$

where $\mu(x)$ is a non-linearity shown in Figure 6 and defined as follows:

$$\mu(x) = \begin{cases} 0 & x < \gamma \\ x & \gamma \leq x < 2 \\ 2 & x \geq 2 \end{cases} \tag{15}$$

This non-linearity shrinks small distortions and allows to account for saturation in case of consecutive degradations of the same block through an operation of hard limiting.

The number of frames to be low-pass filtered is determined by the inter-frame correlation and in the following it will be indicated as φ.

More in details:

- For a given block $b(i,j)$, φ is set to zero if the corresponding block in the previous frame is affected by repeated line distortion and the inter-block correlation is below a predefined threshold (i.e., $\rho_k^{B(i,j)} < \lambda_{\mathrm{RLB}}$) indicating that the block has been updated by I-frame coding.
- φ is set to zero when processing I-frames.
- φ is set to one for frames belonging to static shots.

In order to evaluate the overall distortion index, the map $\mathbf{D}_k^{\mathrm{CB}}$ of corrupted blocks is decomposed into two groups: the first one, denoted in the following with $\mathbf{D}_k^{\mathrm{CCB}}$, contains the entries of $\mathbf{D}_k^{\mathrm{CB}}$ associated to clustered corrupted blocks, while the second one, denoted in the following with $\mathbf{D}_k^{\mathrm{ICB}}$, contains the contributions corresponding to the remaining, isolated, blocks. A block $b(i,j)$ for which $D_k^{\mathrm{CB}}\left[i,j\right] > 0$ is considered member of a cluster if at least for one of its eight surrounding neighbors, $b(p,q)$, the condition $D_k^{\mathrm{CB}^{(p,q)}} > 0$ holds. Let

- $\mathbf{N}_k^{\mathrm{CCB}} = \left\|\mathbf{D}_k^{\mathrm{CCB}}\right\|_{L_1}$ be the the L_1 norm of clustered corrupted block map
- $\mathbf{N}_k^{\mathrm{ICB}} = \eta^{\mathrm{ICB}}\left(\left\|\mathbf{D}_k^{\mathrm{ICB}}\right\|_{L_1}\right)$ be the L_1 norm of isolated corrupted blocks where

$$\eta^{\mathrm{ICB}}(x) = \begin{cases} 0 & x \leq \lambda_{\mathrm{ICB}} \\ x & \text{otherwise} \end{cases} \tag{16}$$

- $\mathbf{N}_k^{\mathrm{RL}} = \left\|\mathbf{D}_k^{\mathrm{RLB}}\right\|_{L_1}$ be the column vector of the number of repeated lines for each image color component
- ρ^{LOSS} the packet loss rate

Then, denoting with $\boldsymbol{\xi}$ is the column vector

$$\boldsymbol{\xi} = \left[\, 1 \;\; \overline{N}^{\mathrm{CCB}} \;\; \overline{N}^{\mathrm{ICB}} \;\; \overline{N}^{\mathrm{RL}} \;\; \rho^{\mathrm{LOSS}} \,\right]^T \tag{17}$$

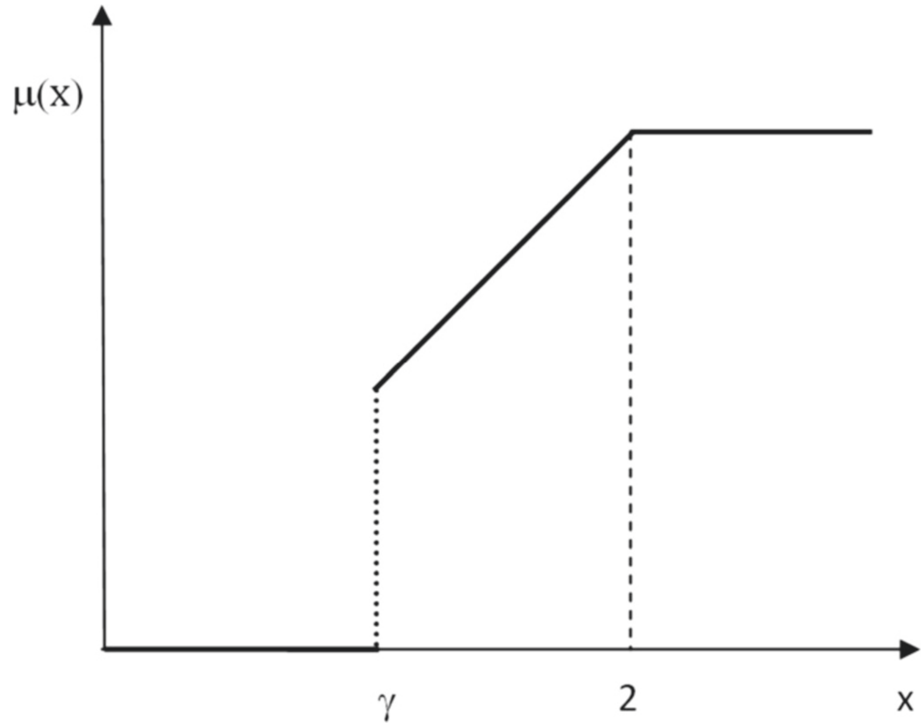

Figure 6 **Non linearity $\mu(x)$ used for distortion map evaluation.**

where

$$\overline{N}^{\mathrm{CCB}} = \frac{1}{L}\sum_{k=1}^{L} N_k^{\mathrm{CCB}}, \tag{18}$$

$$\overline{N}^{\mathrm{ICB}} = \frac{1}{L}\sum_{k=1}^{L} N_k^{\mathrm{ICB}}, \tag{19}$$

$$\overline{N}^{\mathrm{RL}} = \frac{1}{L}\sum_{k=1}^{L} N_k^{\mathrm{RL}}, \tag{20}$$

are the average values of the corresponding L_1 norms and L is the length of the video sequence, the NR metric $\mathrm{VQM}_{\mathrm{NR}}^{(\mathrm{Y})}$ based on the luminance component can be computed as follows:

$$\mathrm{VQM}_{\mathrm{NR}}^{(\mathrm{Y})} = \alpha \left[\boldsymbol{\xi}^T Q^{(Y)} \boldsymbol{\xi}\right]^{1/2} + \beta. \tag{21}$$

The weighting matrix $Q^{(Y)}$ (Table 1) and the regression coefficients α and β can be estimated by fitting subjective experiments, as illustrated in the next section.

We remark that, since $\boldsymbol{\xi}_1 = 1$, the quadratic form includes both linear and quadratic terms.

The above relationships can be directly extended to color images by building a degradation map for each color component. Therefore, assuming that each frame is represented by the luminance Y and color difference components C_b, C_r, the proposed video quality metric specifies as follows

$$\mathrm{VQM}_{\mathrm{NR}}^{(Y,C_b,C_r)} = \alpha_c \left[\boldsymbol{\zeta}^T Q^{(Y,C_b,C_r)} \boldsymbol{\zeta}\right]^{1/2} + \beta_c. \tag{22}$$

where $\boldsymbol{\zeta}$ is the column vector:

$$\boldsymbol{\zeta} = \begin{bmatrix} 1 & \overline{N}_Y^{\mathrm{CCB}} & \overline{N}_Y^{\mathrm{ICB}} & \overline{N}_Y^{\mathrm{RL}} & \overline{N}_{C_b}^{\mathrm{CCB}} & \overline{N}_{C_b}^{\mathrm{ICB}} \\ & \overline{N}_{C_b}^{\mathrm{RL}} & \overline{N}_{C_r}^{\mathrm{CCB}} & \overline{N}_{C_r}^{\mathrm{ICB}} & \overline{N}_{C_r}^{\mathrm{RL}} & \rho^{\mathrm{LOSS}} \end{bmatrix}^T \tag{23}$$

having demoted with

- $\overline{N}_Y^{\mathrm{CCB}}, \overline{N}_{C_b}^{\mathrm{CCB}}, \overline{N}_{C_r}^{\mathrm{CCB}}$ the average numbers of clustered corrupted blocks for the three color components,
- $\overline{N}_Y^{\mathrm{ICB}}, \overline{N}_{C_b}^{\mathrm{ICB}}, \overline{N}_{C_r}^{\mathrm{ICB}}$ the average numbers of isolated corrupted blocks for the three color components,
- $\overline{N}_Y^{\mathrm{RL}}, \overline{N}_{C_b}^{\mathrm{RL}}, \overline{N}_{C_r}^{\mathrm{RL}}$ the average numbers of repeated lines for the three color components.

Table 1 Weighting matrix $Q^{(Y)}$

$Q^{(Y)}_{i,j}$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$j = 1$	-0.0328	-0.0740	-0.0074	0.1138	0.0383
$j = 2$	-0.0740	0.0072	0.0007	0.0392	-0.0006
$j = 3$	-0.0074	0.0007	0.0006	0.0137	-0.0000
$j = 4$	0.1138	0.0392	0.0137	-0.1077	-0.0099
$j = 5$	0.0383	-0.0006	-0.0000	-0.0099	-0.0012

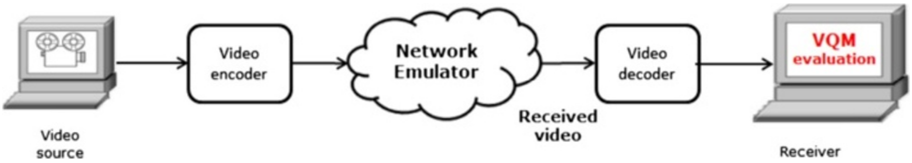

Figure 7 **Block diagram of the experimental setup.**

It is important to notice that the reduction of the impact of isolated corrupted blocks allows to mitigate the effects produced by misclassifications and to account for the lower sensitivity to artifacts in small areas compared to those in wider areas.

5 Experimental results

To identify the parameters specifying a NR metric and to verify its effectiveness, experiments involving human subjects should be performed. As already stated, this procedure is expensive, time-consuming, and often impossible to be performed. The alternative is to compare, under the same testing conditions, the gathered results with those provided by reliable full reference metrics. In the performed test, the NTIA video quality metric ($\mathrm{VQM}_{\mathrm{NTIA}}$) whose software implementation is publicly available and freely downloadable at the URL http://www.its.bldrdoc.gov/vqm, has been adopted.

5.1 Experimental setup

The experimental setup is shown in Figure 7, and it is composed by a streaming source, a network segment, and a receiver. The video server consists of a personal computer equipped with the open source VideoLAN server [57] and the FFmpeg tool [58]. The original video is encapsulated in a MPEG2 TS, packetized in the RTP/UDP/IP protocol stack and transmitted on a 100/1000Base-T Ethernet interface. The network segment has been accounted for by means of an open source network emulator: NETwork EMulator (NETEM) [59]. The emulator has been used for introducing packet losses in the incoming stream, in accordance to the statistics of real networks based on the best-effort paradigm.

Each considered media stream has been processed in order to simulate a set of increasing packet loss rates (PLRs). The selected PLRs are: 0.1%, 0.5%, 0.7%, 0.9%, 1.1%, 1.3%, 1.5%, 2.0%, 3.0%, 5.0%, 10%, 15%, and 20%. At the receiver side, the VLC client receives the media packets and uses concealment techniques for reconstructing the original video.

To evaluate the increase in $\mathrm{VQM}_{\mathrm{NR}}$ performance achievable when full color information is employed with respect to the use of the luminance alone, the parameter identification and the performance assessment have been performed for both gray scale and color videos. Two sets of sequences have been used in our tests. The first one (*test set 1*) is composed by eight video sequences and it has been used for calibrating the $\mathrm{VQM}_{\mathrm{NR}}^{(Y)}$ metric parameters. The sequences are of different content and characterized by still scenes and slow or fast motion rates. The second one (*test set 2*) has been used for evaluating the effectiveness of the proposed metric. The sequences have been extracted from the online database 'The Consumer Digital Video Library' [60]. All analyzed videos have a VGA resolution (640 × 480 pixels, progressive) and a frame rate equal to 30 fps and they are composed by 360 frames. The test set 2 video dataset characteristics are reported in Table 2 while sample frames from the videos are shown in Figure 8.

Table 2 Video dataset characteristics

Video name	Chroma sampling	File size (Mb)	Run time (s)
NTIA Bells: man directing handbell choir	4:2:2	223.5	0:12
NTIA Cargas: car pulls into a gas station	4:2:2	223.5	0:12
NTIA Cartalk: boy complains about sister	4:2:2	223.5	0:12
NTIA Catjoke: a man tells a joke	4:2:2	221.2	0:12
NTIA Diner: young man talks in a diner	4:2:0	223.5	0:12
NTIA Drmset: young man plays on a drum set	4:2:2	223.5	0:12
NTIA Fish: goldfish pond with fades	4:2:2	331.8	0:18
NTIA Guitar: man plays guitar using finger picking style	4:2:2	223.5	0:12
NTIA Magic: girl performs magic trick	4:2:2	223.5	0:12
NTIA Music: a man playing banjo	4:2:2	223.5	0:12
NTIA Rfdev: RF device	4:2:2	221.2	0:12
NTIA Schart: school art	4:2:2	221.2	0:12
NTIA Wboard: man draws a chart on a whiteboard	4:2:2	1.663,4	1:30

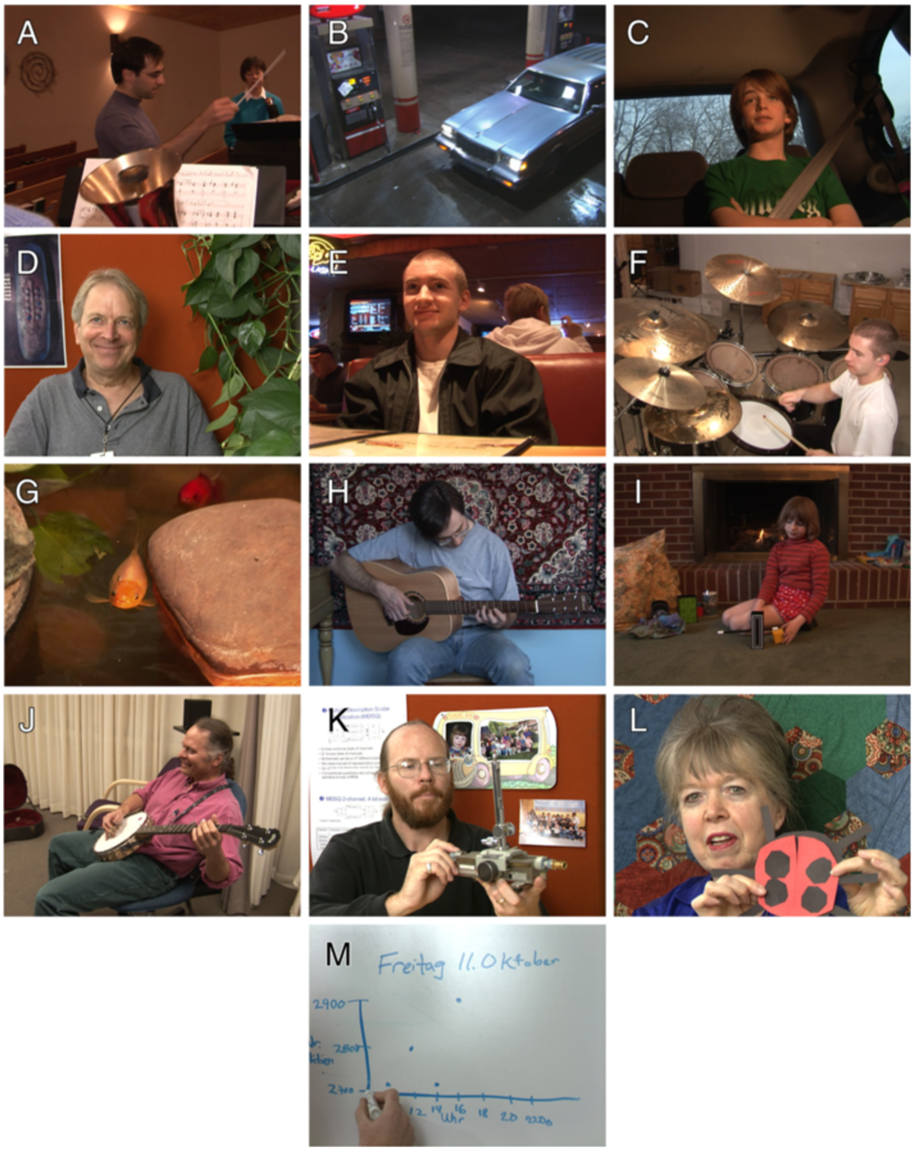

Figure 8 Sample frames extracted from the videos in the dataset used in the experimental tests. (A) Bells. **(B)** Cargas. **(C)** Cartalk. **(D)** Catjoke. **(E)** Diner. **(F)** Drmset. **(G)** Fish. **(H)** Guitar. **(I)** Magic. **(J)** Music. **(K)** Rfdev. **(L)** Schart. **(M)** Wboard.

5.2 Gray scale video tests

The luminance component of a the test set 1 training set has been used for calibrating the $\mathrm{VQM}_{\mathrm{NR}}^{(Y)}$ metric parameters. In this phase, the goal has been to mimic the $\mathrm{VQM}_{\mathrm{NTIA}}$ score computed on the training sequences as much as possible. Based on the achieved results, the thresholds θ_l and θ_h in Equation 4 have been set to 0.3 and 0.9, respectively. The parameters λ_H and λ_V, defined in Equation 7 and 8 have been set to 5 and 1, respectively. From the performed test, it can be noticed that λ_V, although small, is not null to account for small variations induced by partial decoding of a tile affected by errors. The length of the interval employed for the adaptation of η_S (as in Equation 12) is considered constant and $M_h = 7$ has been employed in the reported results. The parameters λ_I, P, and Q, as defined in Equation 10 have been set to 0.25, 2, and 5, respectively.

The capability of the proposed metric to mimic the behavior of the $\mathrm{VQM}_{\mathrm{NTIA}}$ for the training set is illustrated in Figures 9, 10, and 11. In more details, in Figure 9, the results concerning the *Taxi* sequence affected by three packet loss rates (1.9% top row, 3% middle row, and 6.7% bottom row) are reported. As can be noticed from the plots, the proposed metric scores are coherent with the $\mathrm{VQM}_{\mathrm{NTIA}}$ ones, especially for the PLR = 3% and PLR = 6.7% cases.

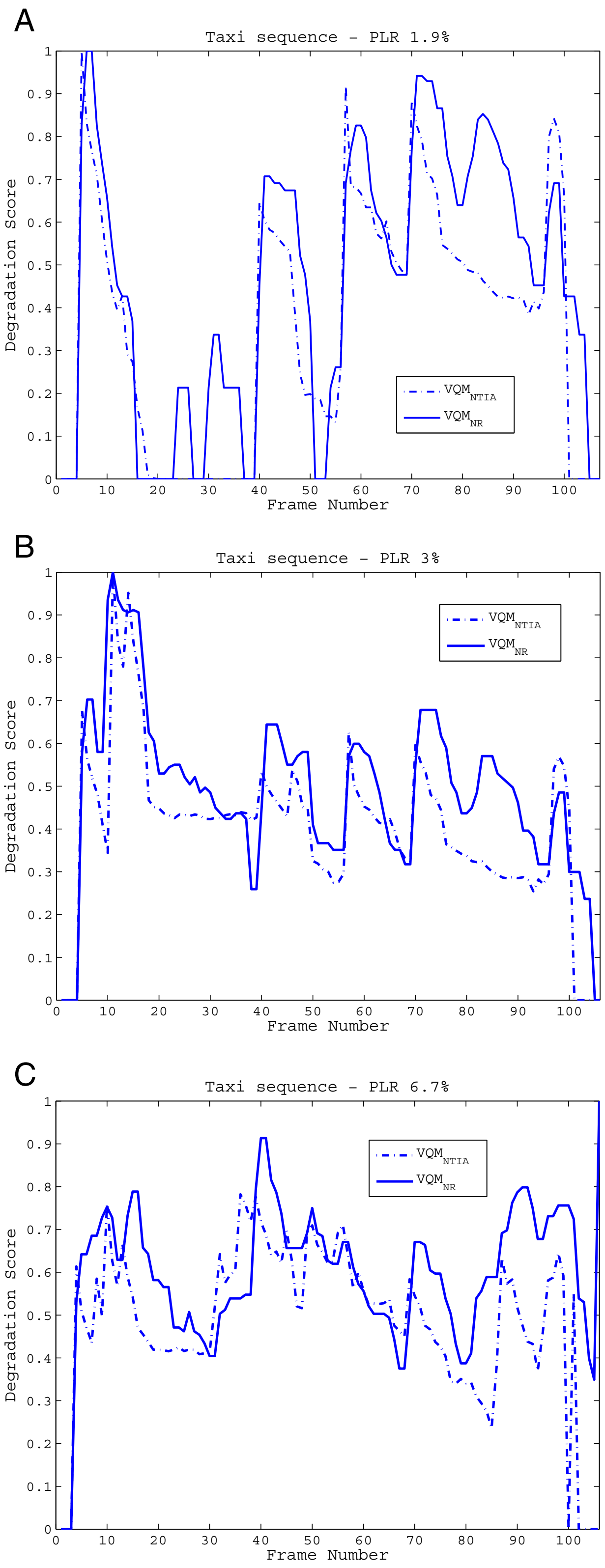

Figure 9 VQM$^{(Y)}_{\text{NR}}$ scores versus VQM ones for the Taxi sequence. **(A)** PLR = 1.9%. **(B)** PLR = 3%. **(C)** PLR = 7%.

It is worth noticing that in Figure 9A, around frame 85, the VQM$_{\text{NR}}$ presents a peak not corresponding to a similar quality variation detected by the VQM$_{\text{NTIA}}$. This behavior highlights the differences between the two metrics. As can be easily noticed by a visual inspection of the considered frames in the Taxi sequence and in its degraded version in Figure 10, there are errors resulting in block artifacts affecting both the main object and the road curbs. In this case, the proposed metric is able to cope with the masking effect of textures and with the perceived impact of silhouette definition and text readability.

The same behavior can be noticed for the sequence Field as reported in Figure 11. For almost the whole sequence, the two indexes show the same behavior. There is a slight tendency in overestimating the video artifacts by the VQM$^{(Y)}_{\text{NR}}$ index. Only for a few frames, the quality assessments provided by the two metrics are opposite: the value is over or below the quality threshold.

In the sequence Horse ride, the overlapping between the two curves is not homogeneous, as shown in Figure 12. Moreover, if the average behavior is compared, among the 25th and the 38th frame, the VQM$^{(Y)}_{\text{NR}}$ indicator shows high degradation while VQM$_{\text{NTIA}}$ only shows a slight degradation. The same different degradation rate can also be noticed in the last part of the sequence.

In order to evaluate the performances of the gray scale VQM$^{(Y)}_{\text{NR}}$ with respect to the quality estimation provided by the full color VQM$_{\text{NTIA}}$ metric, the test set 2 has been employed.

For comparing the performance achieved with the proposed gray scale and full color no-reference metrics, a Monte Carlo simulation of the transmission of the set of full color videos over an IP channel affected by packet losses for several packet loss rates has been performed. Then, only the luminance component of the decoded videos has been employed for computing VQM$^{(Y)}_{\text{NR}}$ while both luminance and color differences have been employed for computing both VQM$^{(Y,c_b,C_r)}_{\text{NR}}$ and VQM$_{\text{NTIA}}$.

In Figure 13, the results obtained for each sequence with the VQM$^{(Y)}_{\text{NR}}$ have been plotted versus VQM$_{\text{NTIA}}$ ones. As can be noticed, there is good matching between the two metrics and the root mean square error (RMSE) value is 0.14. The regression value is 0.86.

5.3 Color video tests

To verify the gain achieved when chrominance is employed, the multivariate regression procedure has then been applied to the full reference metric and to the packet loss rate, the average numbers of clustered corrupted blocks, isolated corrupted blocks, and repeated lines extracted from Y, C_b, and C_r decoded components, thus, obtaining the regression coefficients α_c= 1.0419 and

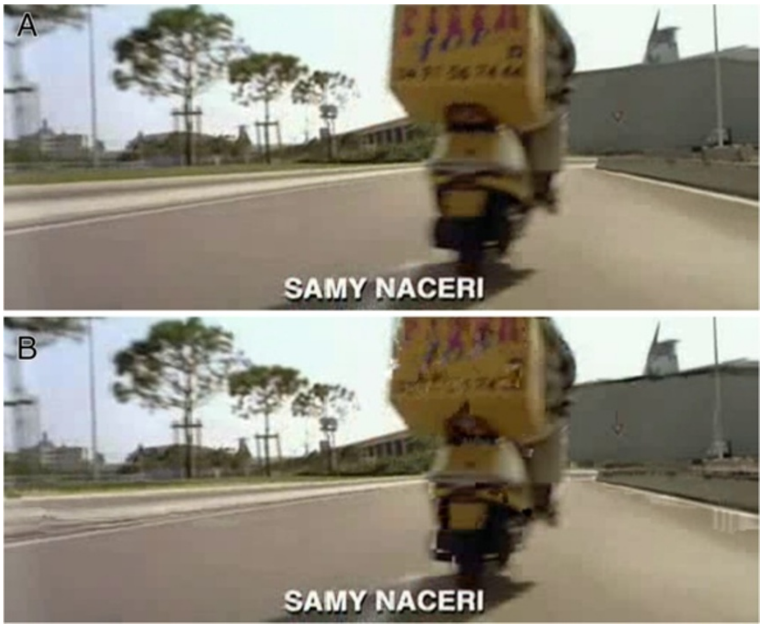

Figure 10 Degraded version of the *Taxi* sequence. Original **(A)** and impaired **(B)** version of the frame number 85 of the *Taxi* sequence.

$\beta_c = -0.0465$. The weighting matrix is reported in Table 3.

In Figure 14, the plot of $\mathrm{VQM}_{\mathrm{NR}}^{(Y,C_b,C_r)}$ versus the $\mathrm{VQM}_{\mathrm{NTIA}}$ is reported for the selected videos. The plot shows improved performances of the proposed metric in matching the full reference score. In fact, the use of color information increases the fitting performances resulting in a regression value of 0.91 and on a RMSE value equal to 0.11.

By analyzing the results, a few issues are open for future investigation. First of all, from the performed experiments for both metric parameters tuning and metric performance effectiveness assessment, a key factor for a successful comparison of NR and FR metrics is

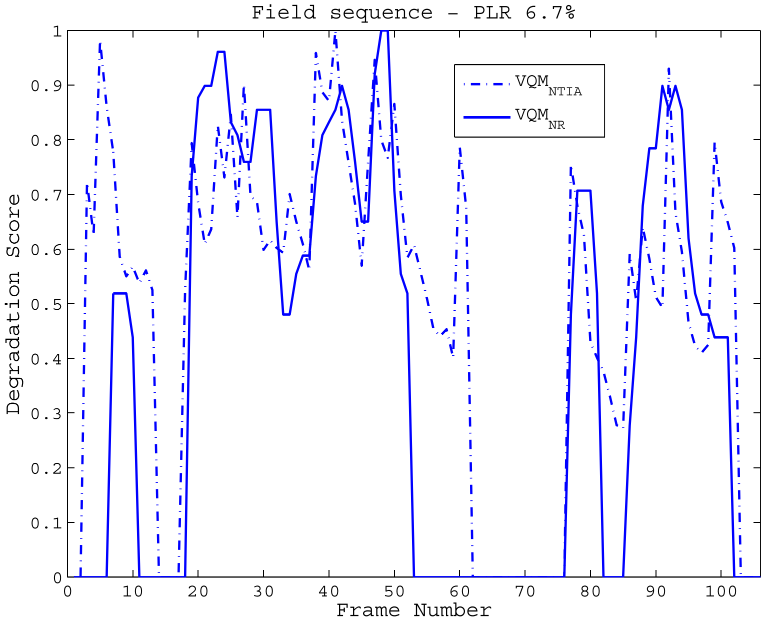

Figure 11 Field sequence: $\mathrm{VQM}_{\mathrm{NR}}^{(Y)}$ versus VQM one at PLR = 6.7%.

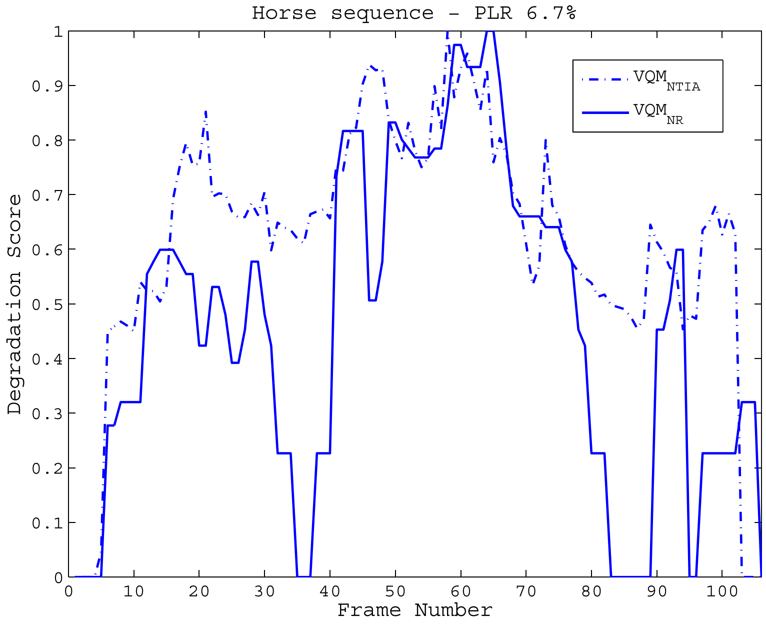

Figure 12 Horse ride sequence: $\mathrm{VQM}_{\mathrm{NR}}^{(Y)}$ versus VQM at PLR = 6.7%.

represented by the temporal realignment algorithm. In the presence of highly textured backgrounds, severe frame losses, and medium to high compression ratios, at least our implementation of the $\mathrm{VQM}_{\mathrm{NTIA}}$ algorithm does not provide reliable estimates of the variable delay between the original and decoded videos. This implies a potential bias in the estimated NR metrics induced by the wrong selection of the reference frame to be used for the comparison. Furthermore, we noticed that the adopted key-frame detection algorithm has an impact on overall distortion evaluation, since many elements, considered in the proposed metric, depend on the shot boundaries detection.

5.4 Subjective experiment

Finally, in order to further verify the effectiveness of the proposed metric, a subjective experiment has been performed.

Sixteen test subjects drawn from a pool of students of the University of Roma TRE have participated to the test. The students are thought to be relatively naive concerning video artifacts and the associated terminology. They were asked to wear any vision correcting devices (glasses or contact lenses) they normally wear to watch television. The subjects were asked to rate the quality of the videos in the test database (listed in Table 2) through a single stimulus quality evaluation method [61].

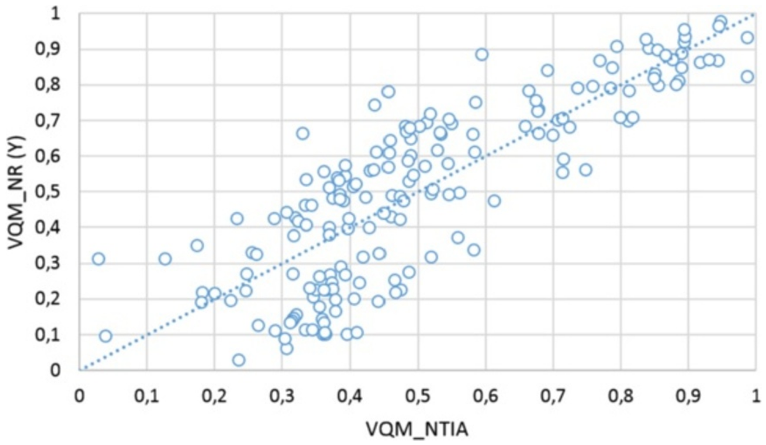

Figure 13 $\mathrm{VQM}_{\mathrm{NR}}^{(Y)}$ scores versus $\mathrm{VQM}_{\mathrm{NTIA}}$ ones for the gray scale sequences. Regression value = 0.86 and RMSE = 0.14.

Table 3 Weighting matrix Q

$Q_{i,j}$	i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8	i = 9	i = 10	i = 11
j = 1	0.0094	0.0728	0.0101	0.0099	-0.0202	-0.0003	0.0768	0.1929	-0.0065	0.0195	0.0266
j = 2	0.0728	-0.0085	-0.0011	-0.0520	-0.0633	0.0013	0.0071	0	-0.0245	0.2121	-0.0022
j = 3	0.0101	-0.0011	-0.0009	0.0039	-0.0119	0.0007	0.0041	0	-0.0043	0.0071	0.0001
j = 4	0.0099	-0.0520	0.0039	-0.0052	0.0855	-0.0016	-0.0524	0	-0.0254	0.0329	0.0008
j = 5	-0.0202	-0.0633	-0.0119	0.0855	-0.0242	-0.0456	0.1776	0	0.2104	-0.4598	0.0080
j = 6	-0.0003	0.0013	0.0007	-0.0016	-0.0456	-0.0005	-0.0030	-0.0094	0.0012	-0.0013	0.0004
j = 7	0.0768	0.0071	0.0041	-0.0524	0.1776	-0.0030	-0.1082	-0.1023	-0.0038	0.1061	-0.0007
j = 8	0.1929	0	0	0	0	-0.0094	-0.1023	-0.0120	0.0224	-0.3094	0.0109
j = 9	-0.0065	-0.0245	-0.0043	-0.0254	0.2104	0.0012	-0.0038	0.0224	0.0091	0.0220	-0.0006
j = 10	0.0195	0.2121	0.0071	0.0329	-0.4598	-0.0013	0.1061	-0.3094	0.0220	-0.1038	-0.0078
j = 11	0.0266	-0.0022	0.0001	0.0008	0.0080	0.0004	-0.0007	0.0109	-0.0006	-0.0078	-0.0011

A Panasonic Viera monitor (46") is used to display the test video sequences. The experiment is run with one subject at a time. Each subject was seated straight ahead in front of the monitor, located at or slightly below eye height for most subjects. The subjects are positioned at a distance of four screen heights (80 cm) from the video monitor in a controlled light environment. The experimental session consisted of four stages. In the first stage, the subject was verbally given instructions for performing the test. In the second stage, training sequences were 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 stage, the test subjects run 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 experiment is performed on the complete set of test sequences. After each video was displayed, the subject was asked to enter his/her judgment in a scale from 1 to 5, where 5 corresponds to best quality and 1 to worst quality.

In Figure 15, the comparison between the collected MOS and the two objective metrics are reported. The MOS has been normalized in the range 0 (best quality) to 1 (worst quality). The RMS is 0.10 for the $\mathrm{VQM}_{\mathrm{NR}}^{(Y,C_b,C_r)}$ metric and 0.16 for $\mathrm{VQM}_{\mathrm{NTIA}}$ metric, respectively. As can be noticed, the proposed metric is able to predict the subjective judgment.

6 Conclusions

In this paper, a no-reference metric for assessing the quality of MPEG-based video transmissions over IP-based networks is presented. The proposed approach is based on the analysis of the inter-frame correlation measured at the receiver side. Several tests have been performed for tuning and evaluating the performances of the proposed metric. The scores collected by this tool in evaluating

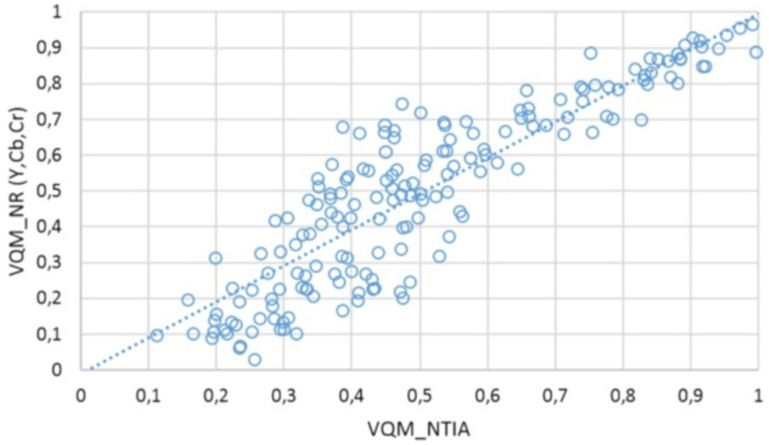

Figure 14 $\mathrm{VQM}_{\mathrm{NR}}^{(Y,C_b,C_r)}$ scores versus $\mathrm{VQM}_{\mathrm{NTIA}}$ ones for color videos. Regression value = 0.91 and RMSE = 0.11.

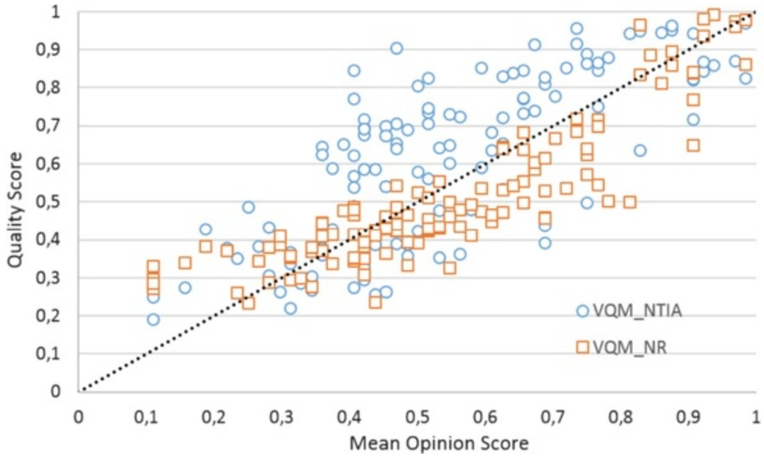

Figure 15 Normalized MOS versus VQM$_{\mathrm{NR}}^{(Y,C_b,C_r)}$ and VQM$_{\mathrm{NTIA}}$ metrics.

impaired videos have been compared with the ones gathered with the full reference VQM$_{\mathrm{NTIA}}$ metrics and with the MOS collected by means of a subjective experiment. The overall analysis demonstrates the effectiveness of the VQM$_{\mathrm{NR}}$. Current investigation is devoted to solve the problems arising when using evaluation methods that are not based on reference signals. In particular, for the temporal realignment algorithm that is needed for the FR metrics in order to correctly estimate the NR parameters, we plan to test a novel re-synchronization procedure. Recently, the NTIA group announced the release of a new version of VQM metrics especially tuned for variable packet loss rate. Even if the problem of realignment is still to be solved, the use of such a metric could probably be used for a more effective parameters tuning. As a general remark, the influence of the adopted key-frame detection algorithm should be investigated. In fact, if a fake key-frame is selected due to estimation errors, the quality metric immediately decreases. Another issue is related to the amount of motion characterizing the sequences. We noticed a difference in the scores when slow or almost null motion rate is present. The choice of the parameter λ_s should be based on the consideration, confirmed by many studies, that human attention is attracted by objects whose movement is relevant with respect to the other elements in the scene. Therefore, λ_s should probably be adapted to the relative motion of the surrounding areas. Finally, a key issue to be further investigated is the influence of the adopted error concealment technique implemented in the decoder. With the improving of error concealment masking techniques, the concealed video may present different error patterns from the ones we are experiencing at the moment. For example, we noticed that the latest 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 weight of such parameters may be different depending on the improvements achieved in the field of error concealment techniques.

Competing interests

The authors declare that they have no competing interests.

Received: 30 July 2013 Accepted: 24 January 2014
Published: 25 February 2014

References

1. F Battisti, M Carli, E Mammi, A Neri, A study on the impact of AL-FEC techniques on TV over IP quality of experience. EURASIP J. Adv. Signal Process. **2011**, 86 (2011). http://asp.eurasipjournals.com/content/2011/1/86.
2. A Perkis, Y Abdeljaoued, C Christopoulos, T Ebrahimi, JF Chicharo, Universal multimedia access from wired and wireless systems. Circuits, Syst., Signal Process. Special Issue Multimedia Commun. **20**(3), 387–402 (2001)
3. F Pereira, I Burnett, Universal multimedia experiences for tomorrow. IEEE Signal Process. Mag. **20**(2), 63–73 (2003)
4. Psytechnics Ltd, Psytechnics no-reference video quality assessment model, in *ITU-T SG9 Meeting, COM9-C190-E: Geneva, 5 May 2008* (Psytechnics Ltd., Ipswitch, 2008)
5. YH Kim, J Han, H Kim, J Shin, Novel no-reference video quality assessment metric with estimation of dynamic range distortion, in *Proceedings of the 12th International Conference on Advanced Communication Technology (ICACT): 7–10 Feb 2010; Phoenix Park*, vol. 2 (IEEE, Piscataway, 2010), pp. 1689–1692
6. S Chikkerur, V Sundaram, M Reisslein, LJ Karam, Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans. Broadcasting. **57**(2), 165–182 (2011)
7. M Vranješ, S Rimac-Drlje, K Grgić, Review of objective video quality metrics and performance comparison using different databases. Image Commun. **28**, 1–19 (2013)
8. S Winkler, Video quality measurement standards—current status and trends, in *Proceedings of the Seventh International Conference Information, Communications and Signal Processing: 8–10 Dec 2009; Macau* (IEEE, Piscataway, 2009), pp. 1–5
9. M Pinson, S Wolf, A new standardized method for objectively measuring video quality. IEEE Trans. Broadcasting. **50**(3), 312–322 (2004)
10. ITU-T, *Recommendation J.144, objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference* (ITU, Geneva, 2004)

11. A Basso, I Dalgic, FA Tobagi, CJ van den Branden Lambrecht, Feedback-control scheme for low-latency constant-quality MPEG-2 video encoding, in *SPIE:16 Sept 1996; Berlin*, vol. 2952 (SPIE, Berlin, 1996), pp. 460–471
12. CJ van den Branden Lambrecht, Color moving pictures quality metric, in *Proceedings of the International Conference on Image Processing (ICIP): 16–19 Sept 1996; Lausanne*, vol. 1 (IEEE, Piscataway, 1996), pp. 885–888
13. Z Wang, A Bovik, H Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004). citeseer.ist.psu.edu/article/wang04image.html
14. S Wolf, MH Pinson, SD Voran, AA Webster, Objective quality assessment of digitally transmitted video, in *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing: 9–10 May 1991; Victoria* (IEEE, Piscataway, 1991), pp. 477–482
15. AB Watson, QJ Hu, JFM Gowan, Digital video quality metric based on human vision. J. Electron. Imaging **10**, 20–29 (2001)
16. S Winkler, Issues in vision modeling for perceptual video quality assessment. Signal Process. **78**(2) (1999). citeseer.ist.psu.edu/winkler99issues.html
17. A Shnayderman, A Gusev, AM Eskicioglu, An SVD-based grayscale image quality measure for local and global assessment. IEEE Trans. Image Process. **15**(2), 422–429 (2006)
18. G Zhai, W Zhang, X Yang, S Yao, Y Xu, GES: a new image quality assessment metric based on energy features in Gabor transform domain, in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS): 21–24 may 2006; Island of Kos* (IEEE, Piscataway, 2006), p. 4
19. M Carnec, P Le Callet, D Barba, Full reference and reduced reference metrics for image quality assessment, in *Proceedings of the* $7th$ *International Symposium on Signal Processing and its Applications (ISSPA):1–4 Jul 2003; Paris*, vol. 1 (IEEE, Piscataway, 2003), pp. 477–480
20. TM Kusuma, HJ Zepernick, M Caldera, On the development of a reduced-reference perceptual image quality metric, in *Proceedings on Systems Communications: 14–17 Aug 2005* (IEEE, Piscataway, 2005), pp. 178–184
21. Z Wang, EP Simoncelli, Reduced-reference image quality assessment using a wavelet-domain natural image statistic model, in *Proceedings of SPIE Human Vision and Electronic Imaging X: San Jose*, vol. 5666 (SPIE, Berlin, 2005)
22. S Kanumuri, SG Subramanian, PC Cosman, AR Reibman, Predicting H.264 packet loss visibility using a generalized linear model, in *Proceedings of the IEEE International Conference on Image Processing (ICIP): 8–11 Oct 2006; Atlanta* (IEEE, Piscataway, 2006), pp. 2245–2248
23. S Kanumuri, PC Cosman, AR Reibman, VA Vaishampayan, Modeling packet-loss visibility in MPEG-2 video. IEEE Trans. Multimedia **8**(2), 341–355 (2006)
24. T Liu, X Feng, AR Reibman, Y Wang, Saliency inspired modeling of packet-loss visibility in decoded videos, in *Proceedings of the* $4th$ *International workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM): 14–16 Jan 2009; Scottsdale*. Online: www.vpqm.org
25. P Campisi, M Carli, G Giunta, A Neri, Blind quality assessment system for multimedia communications using tracing watermarking. Signal Process., IEEE Trans. **51**(4), 996–1002 (2003)
26. M Farias, M Carli, S Mitra, Objective video quality metric based on data hiding. IEEE Trans. Consum. Electron. **51**, 983–992 (2005)
27. M Carli, M Farias, E Drelie Gelasca, R Tedesco, A Neri, Quality assessment using data hiding on perceptually important areas, in *IEEE International Conference on Image Processing: 11–14 Sept 2005; Genoa*, vol. 3 (IEEE, Piscataway, 2005), pp. 1200–1203
28. F Battisti, M Carli, A Neri, Video error concealment based on data hiding in the 3D wavelet domain, in *Proceedings of the* $2nd$ *European Workshop on Visual Information Processing (EUVIP): 5–6 Jul 2010; Paris* (IEEE, Piscataway, 2010), pp. 134–139
29. A Ninassi, PL Callet, F Autrusseau, Pseudo no reference image quality metric using perceptual data hiding, in *Proceedings of the SPIE Human Vision and Electronic Imaging XI* (SPIE, Berlin, 2006)
30. A Phadikar, P Maity, C Delpha, Data hiding for quality access control and error concealment in digital images, in *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo (ICME): 11–15 Jul 2011; Barcelona* (IEEE, Piscataway, 2011), pp. 1–6
31. HR Wu, M Yuen, A generalized block edge impairment metric for video coding. Signal Process. Lett. **4**(11), 317–320 (1997)
32. P Marziliano, F Dufaux, S Winkler, T Ebrahimi, A no-reference perceptual blur metric, in *Proceedings of the IEEE International Conference on Image Processing*, vol. 3 (IEEE, Piscataway, 2002)
33. M Carli, D Guida, A Neri, No-reference jerkiness evaluation method for multimedia communications. Procs. SPIE Image Qual. Syst. Perform. III. **6059**, 350–359 (2006)
34. DS Turaga, Y Chen, J Caviedes, No reference PSNR estimation for compressed pictures. Proc. Elsevier Signal Process. Image Commun. **19**, 173–184 (2004)
35. W Ci, H Dong, Z Wu, Y Tan, Example-based objective quality estimation for compressed images. IEEE Multimedia. **99** (2009)
36. A Leontaris, PC Cosman, AR Reibman, Quality evaluation of motion-compensated edge artifacts in compressed video. IEEE Trans. Image Process. **16**(4), 943–956 (2007)
37. A Leontaris, PC Cosman, Compression efficiency and delay trade-offs for hierarchical B-Pictures and pulsed-quality frames. IEEE Trans. Image Process. **16**(7), 1726–1740 (2007)
38. J Gustafsson, G Heikkila, M Pettersson, Measuring multimedia quality in mobile networks with an objective parametric model, in *Proceedings of the* $15th$ *IEEE International Conference on Image Processing (ICIP): 12–15 Oct 2008; San Diego* (IEEE, Piscataway, 2008), pp. 405–408
39. M Naccari, M Tagliasacchi, F Pereira, S Tubaro, No-reference modeling of the channel induced distortion at the decoder for H.264/AVC video coding, in *Proceedings of the* $15th$ *IEEE International Conference on Image Processing (ICIP): 12–15 Oct 2008; San Diego* (IEEE, Piscataway, 2008), pp. 2324–2327
40. Y Liu, Y Zhang, M Sun, W Li, Full-reference quality diagnosis for video summary, in *Proceedings of the IEEE International Conference on Multimedia and Expo: 23 Jun–26 Apr 2008; Hannover* (IEEE, Piscataway, 2008), pp. 1489–1492
41. J Han, YH Kim, J Jeong, J Shin, Video quality estimation for packet loss based on no-reference method, in *Proceedings of the* $12th$ *International Conference on Advanced Communication Technology: 7–10 Feb 2010; Phoenix Park* (IEEE, Piscataway, 2010), pp. 418–421
42. SO Lee, DG Sim, Hybrid bitstream-based video quality assessment method for scalable video coding. Opt. Eng. **51**(6) (2012)
43. N Staelens, N Vercammen, Y Dhondt, B Vermeulen, P Lambert, R Van de Walle, P Demeester, ViQID: a no-reference bit stream-based visual quality impairment detector, in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)* (IEEE, Piscataway, 2010), pp. 206–211
44. N Staelens, D Deschrijver, E Vladislavleva, B Vermeulen, Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression, in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt*, vol. 23 (IEEE, Piscataway, 2013), pp. 1322–1333
45. N Staelens, GV Wallendael, K Crombecq, N Vercammen, *No-reference bitstream-based visual quality impairment detection for high definition H.264/AVC encoded video sequences*, vol. 58 (IEEE, Piscataway, 2012), pp. 187–199
46. C Keimel, J Habigt, M Klimpke, K Diepold, Design of no-reference video quality metrics with multiway partial least squares regression, in *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX): 7–9 Sep 2011; Mechelen* (IEEE, Piscataway, 2011), pp. 1322–1333
47. C Keimel, J Habigt, K Diepold, Hybrid no-reference video quality metric based on multiway PLSR, in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO): 27–31 Aug 2012; Bucharest* (IEEE, Piscataway, 2012), pp. 1244–1248
48. S Shi, K Nahrstedt, R Campbell, Distortion over latency: novel metric for measuring interactive performance in remote rendering systems, in *2011 IEEE International Conference on Multimedia and Expo (ICME): 11–15 July 2011; Barcelona* (IEEE, Piscataway, 2011), pp. 1–6
49. E Bosc, F Battisti, M Carli, P Le Callet, A wavelet-based image quality metric for the assessment of 3D synthesized views, in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, vol. 8648 (IEEE, Piscataway, 2013)
50. L Azzari, F Battisti, A Gotchev, M Carli, K Egiazarian, A modified non-local mean inpainting technique for occlusion filling in depth-image-based rendering, in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, vol. 7863 (IEEE, Piscataway, 2011)
51. AA Webster, CT Jones, MH Pinson, SD Voran, S Wolf, An objective video quality assessment system based on human perception, in *Proceedings of SPIE Human Vision, Visual Processing, and Digital Display IV*, vol. 1913 (SPIE, Berlin, 1993), pp. 15–26

52. P Bretillon, J Baina, M Jourlin, G Goudezeune, Method for image quality monitoring on digital television networks, in *Proceedings of the SPIE Multimedia Systems and Applications II*, vol. 3845 (SPIE, Berlin, 1999), pp. 298–306
53. Z Wang, A Bovik, B Evans, Blind measurement of blocking artifacts in images. Proc. IEEE Int. Conf. Image Process. **3**, 981–984 (2000)
54. J Caviedes, J Jung, No-reference metric for a video quality control loop, in *Proceedings of the 5th World Multiconference on Systemics, Cybernetics, and Informatics* (IIIS, Orlando, 2001), pp. 290–295
55. F Zhang, W Lin, Z Chen, KN Ngan, Additive log-logistic model for networked video quality assessment. Image Process., IEEE Trans. **22**(4), 1536–1547 (2013)
56. M Montenovo, A Perot, M Carli, P Cicchetti, A Neri, Objective quality evaluation of video services, in *Procs. of the 1st International Workshop on Video Processing and Quality Metrics for Consumer Electronic (VPQM)*, (2006). www.vpqm.org
57. VideoLan team, VideoLAN - VLC media player. http://www.videolan.org/. Accessed 3 Feb 2014
58. FFmpeg team, FFmpeg. http://www.ffmpeg.org/. Accessed 3 Feb 2014
59. NetEm team, Network Emulation with NetEm (2005). http://www.linuxfoundation.org/. Accessed 3 Feb 2014
60. CDVL Team, The consumer digital video library (2011). http://www.cdvl.org. Accessed 3 Feb 2014
61. ITU-T, *Recommendation BT.500-11, methodology for the subjective assessment of the quality of television pictures* (ITU, Geneva, 2002). www.itu.org

doi:10.1186/1687-5281-2014-13
Cite this article as: Battisti *et al.*: No reference quality assessment for MPEG video delivery over IP. *EURASIP Journal on Image and Video Processing* 2014 **2014**:13.