14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

NO REFERENCE QUALITY ASSESSMENT OF INTERNET MULTIMEDIA SERVICES

Alessandro Neri[1], Marco Carli[1], Marco Montenovo[2], and Francesco Comi[1]

[1]Department of Applied Electronics, ROMA TRE University, Roma, Italy

[2] HP C&I Italy, Roma, Italy

phone: + (39) 06 55177017, fax: + (39) 06 55177026, email: neri@uniroma3.it
web: http://comlab.uniroma3.it,

ABSTRACT

In this paper an objective No Reference metric for assessing the quality degradations 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. UMTS based video-telephony, multimedia messaging, video on demand over Internet, satellite, terrestrial, and mobile digital video broadcasting, experienced during the last year, increase the relevance of real time video quality assessment in the design and provisioning of new multimedia services. In this new scenario, the broadcast network is often out of control of content providers. Nevertheless, evaluation of effectiveness of the offered services and customer satisfaction, supporting, for instance, service plan options selection and price policy, do require continuous monitoring of the key performance indices. As a matter of fact, during the last decade research on video quality has been mainly focused of the development of *objective* video quality metrics mimicking the average *subjective* behaviour, based on the knowledge of both the original and the received video stream. The main drawback of these techniques, named *full reference quality metrics*, is that they cannot be used in real time applications, because the original video is not available at the receiver side.

On the other hand, since the perceived quality of a digital video is content dependent, it can not be directly inferred from the knowledge of the channel reliability and temporal integrity alone. To partially overcome this problem, reduced reference (RR) quality metrics as well as no reference metrics have been devised. RR methods are based on the extraction from the original video of features that summarize the video content and that can be transmitted to the receiver with a negligible bandwidth increment. Then, the performance assessment is driven by this 'side' information [6].

The only solution that can be used in real time application without modifying bitstream formats and communication protocols, is the "No-Reference" (NR) metric based on the received videostream only. Although less accurate than FR metric, NR metric can be obtained without extra information.

In this contribution we propose a no-reference method to assess the degradations 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 playout temporal constraints (e.g. maximum temporal delay) can appear at the decoder input as losses of isolated as well as clustered packets. 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.

Here we assume that we can just observe the output of the rendering algorithm, and therefore we have 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.

Here we attempt to recover this information from the analysis of the statistical properties of the rendered video.

The rest of the paper is organized as follow. In Section 2 the structure of the proposed metric is described. In Section 3 some results of the performed experiment are presented; finally, in Section 4 the conclusions are drawn.

2. THE NR PROCEDURE

Channel errors and end-to-end jittering delays can produce quite different effects ranging from the loss of isolated blocks to the loss of one or more consecutive frames. In the latter case, decoders can adopt several strategies including: (trivial) filling of the image field with a predefined colour, freezing of the last decoded frame, and predicting/interpolating the lost

frame(s) on the basis of the playout buffer content. Even more options are available to the designers when small clusters or isolated blocks are lost, for they can exploit both temporal and spatial correlation.

Since loss of one or more consecutive frames is equivalent to a dynamical variation of the frame rate, the quality loss can be effectively assessed by means of the tools proposed by the authors for both RR, [7], and NR jerkiness estimation [6]. Thus, in this contribution we mainly focus our attention on the loss of clustered and isolated blocks. At this aim, since in the first case the concealed frames are characterised by an high temporal correlation, we first segment the rendered sequence into static and dynamical shots. Then we test the static shots in order to detect whether one or more consecutive frames have been lost. Finally dynamical shots are tested to verify the presence of isolated and clustered corrupted blocks. Both static versus dynamical shot segmentation and construction of temporal distortion maps are based on the analysis of the interframe correlation. Thus, given a sequence $\{\mathbf{F}_k, k=1,...,L\}$ of L frames, each consisting of M_rxM_c square blocks $\mathbf{B}_k^{(i,j)}$ of size NxN, we denote with $\bar{F}_k$ and $\bar{B}_k^{(i,j)}$ the frame and the block average values and with $\Delta\bar{\mathbf{F}}_k = \mathbf{F}_k - \bar{F}_k$ and $\Delta\mathbf{B}_k^{(i,j)} = \mathbf{B}_k^{(i,j)} - \bar{B}_k^{(i,j)}$ their deviations. Then, the normalized interframe correlation coefficient at time t is:

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

where <,> denotes the inner product and $\|\bullet\|_{L^2}$ the L^2 norm. Similarly, the interblock correlation is given by:

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

Segmentation of the videosequence into static and dynamical shots is performed by comparing the interframe correlation ρ_k with a threshold λ_S. As illustrated by Fig.1, when frames are lost and the receiver holds the last decoded image until a new frame is received, the interframe correlation reaches values near 1. Nevertheless high correlation does not necessarily denotes concealment of total or partial frame loss, although in this case the inteframe correlation tends to become spiky.

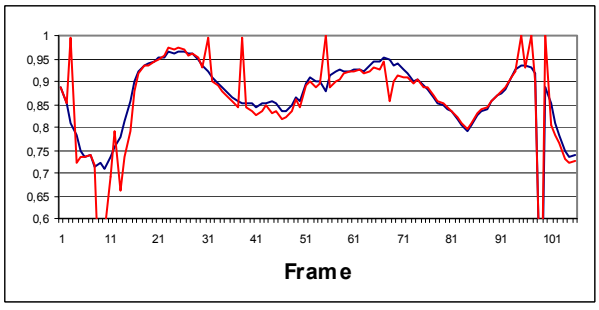

Fig. 1. Interframe normalized correlation for the original "Taxi" sequence (blue) and the sequence affected by a 6% packet loss (red)

The performed experiments evidenced that setting λ_S=0.99 allows to detect the major part of repeated frames still with an acceptable level of false alarms. Let us recall that jerkiness assessment is then applied to static shots. Thus segmentation is mainly aimed at the reduction of the computational complexity. On the other hand the first of a group of repeated frames can be still affected by artefacts. Thus apart from jerkiness, the construction of the distortion map of the first frame of a static shot follows the same procedure employed for frames of dynamical shots. That map is then inherited by the whole set of frames of a static shot.

Temporal analysis

Evaluation of the distortion associated to isolated and clustered blocks affected by errors is performed by applying first a coarse temporal analysis extracting, for each frame, those blocks potentially affected by artefacts produced by lost packets and a finer spatial analysis refining the degradation map. In essence, as first step, we classify each block of each frame either as affected by medium content variations, or as affected by an unusually large temporal variation, or, finally, as practically unchanged. The corresponding tentative distortion map is computed by comparing the interframe correlation of each bock with a set of thresholds:

$$\Gamma_k^{CB}[i,j] = \begin{cases} 0 & \theta_l \le \rho_k^{B(i,j)} \le \theta_h \\ 1 & \theta_l < \rho_k^{B(i,j)} \\ 2 & \rho_k^{B(i,j)} > \theta_h \end{cases} \qquad (3)$$

In practice θ_l=0.3 and θ_h=0.9 have been employed in the numerical experiments. let us observe that the highest distortion index is assigned to blocks practically copied from the previous frame, while zero distortion is assigned at this stage to blocks with medium content variation.

Spatial analysis

Blocks classified as potentially affected by packet loss artefacts undergo a spatial analysis consisting, essentially, of a static regions detection, a vertical and horizontal edge consistency check and repeated lines test.

a) Static regions detection

For each block with $\Gamma_k^{CB}[i,j] = 2$ we test if at least n out of the 8 neighbours have been also classified as practically unchanged. In case of positive result, the block is classified as belonging to a static region and its potential distortion is reset to zero.

b) Edge consistency check

The edge consistency check controls the presence of edge discontinuities at the block boundaries. Let us denote with E_l and E_r the L^1 norms of the vertical edges respectively falling on the left and on the right boundary of the block, and with A_c, A_l and A_r the average L^1 norms of the vertical edges falling inside the current block and on the two adjacent blocks respectively on the left and on the right side.

A block for which $\Gamma_k^{CB}[i,j] \neq 0$ is classified as affected by visible distortion if

$$\left| E_l - \frac{(A_c + A_l)}{2} \right| > \theta \quad OR \quad \left| E_r - \frac{(A_c + A_r)}{2} \right| > \theta \,. \qquad (4)$$

Similar procedure is applied along the horizontal direction.

c) Repeated lines test

As illustrated in Fig.2, when the packet loss affects an intra-frame encoded image, it can happen that only a portion of the frame is properly decoded, while the remaining part is filled with the content of the last row correctly decoded. This produces a band of vertical stripes ending on the bottom line.

Let us denote with $\mathbf{f}_k[m]$ the m-th row of the k-th frame. To detect repeated lines the following procedure is proposed.

- If the L^1 norm of the horizontal gradient component on the bottom line exceeds a threshold λ_H, i.e.,

$$\left\| \Delta \mathbf{f}_k[N_r] \right\|_{L^1} > \lambda_H \qquad (5)$$

starting from the bottom line, and moving upward we verify if consecutive lines are identical by comparing the L^1 norm of their difference wit a threshold λ_v, i.e.,

$$\left\| \mathbf{f}_k[m] - \mathbf{f}_k[m+1] \right\|_{L^1} < \lambda_V \qquad (6)$$

until the test fails.

Fig.2. frame affected by vertical stripes.

Based on the performed experiments, λ_H e λ_V have been set to 5 and 1 respectively. We observe that λ_V, although small, is not null to account for small variations induce by partial decoding of a tile affected by errors. Once the repeated lines test has been performed, a map $\Gamma_k^{RL}[i,j]$, with value equal to 1 if a block belongs to a region filled with vertical stripes and 0 otherwise, is generated.

Reference frame detection

The previous procedure allows to assess the presence of blocks affected by distortions caused by loss of packets belonging to the current frame. Nevertheless, due to error propagation these distortions persist until an intraframe encoded image, denoted in the following as I-frame, is received. As illustrated in Fig.3 where the normalized inter-frame correlation of the "Kill Bill" sequence is depicted, an I-frame is usually characterized by a low correlation with previous frame and an high correlation with the next frame. Thus, the k-th frame is classified as an I-Frame if

$$\rho_{k-1} - \rho_k > 2\eta_P \quad AND \quad \rho_{k+1} - \rho_k > 2\eta_S \qquad (7)$$

and no more than P out of Q previous frames or no more

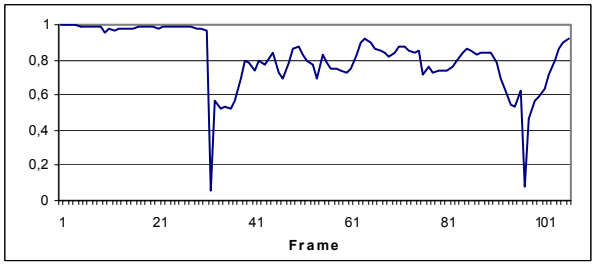

Fig.3. Normalized interframe correlation of the original "Kill Bill" sequence.

than P out of Q next frames present an overall distortion Γ_k greater than a threshold λ_I (values of P=2, Q=5, and λ_I =0.25 have been employed in the performed experiments).

As outlined in the scheme of Fig.4, 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 coefficient over the intervals [k-M_l, k] and [k, k+M_h], i.e.:

$$\eta_P = \frac{1}{M_l} \sum_{h=k-M_l+1}^{k} \left| \rho_h - \rho_{h-1} \right| . \qquad (8)$$

and

$$\eta_S = \frac{1}{M_h} \sum_{n=k+1}^{k+M_h} \left| \rho_n - \rho_{n-1} \right| . \qquad (9)$$

M_l is set in such a way that the time interval employed for the adaptation of η_P starts at the frame next the last detected I-frame. On the other hand, when processing the k-th frame, no information about location of next I-frames is available and the extent of the interval employed for the adaptation of η_S is kept constant (a value of M_h=7 has been employed in the reported results).

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. DISTORTION MAP EVALUATION

The evaluation of the NR video quality metric is based on the degradation index maps $\Gamma_k^{RL}[i,j]$ and $\Gamma_k^{CB}[i,j]$ whose evaluation has been illustrated in previous section.

To account for error propagation induced by predictive coding, a low pass temporal filtering is applied to the degradation index maps. Specifically, let us denote with $\mathbf{D}_{k-1}$ the generic distortion map at time (k-1), then the distortion map $\mathbf{D}_k$ at time k of frames belonging to dynamical shots is evaluated as follows

$$\mathbf{D}_k = \mu\left[\mathbf{\Gamma}_k + \varphi(\rho_k) \mathbf{D}_{k-1} \right], \qquad (10)$$

where μ is the non-linearity:

$$\mu(x) = \begin{cases} 0 & x < \gamma \\ x & \gamma \le x < 2 \\ 2 & x \ge 2 \end{cases} \qquad (11)$$

that accounts for both shrinking of small distortions and hard limiting to account for saturation in case of consecutive degradations of the same block.

The gain φ controlling the temporal extent of the memory associated to the low pass filter is varied in accordance to the interframe correlation as summarized in Table .1

Table 1. Low pass filter gain.

ρ_k	$\rho_k < 0.9$	$0{,}9 \le \rho_k < 0{,}98$	$0.98 \le \rho_k \le 1$
$\varphi(\rho_k)$	0.1	0.3	0.8

In addition, for a given block, let say [i, j], the gain φ is set to zero if the corresponding block of previous frame contained repeated lines and the interblock correlation drops below a predefined threshold (i.e. $\rho_k^{B(i,j)} < \lambda_{RL}$) indicating that the block has been updated by intraframe coding.
Similarly, the gain φ is set to zero when processing I-frames.
Finally, φ is set to one for frames belonging to static shots.
In order to evaluate the overall distortion index, the map $\mathbf{D}_k^{CB}$ of corrupted blocks is further split into two contributions: the first one, denoted in the following with $\mathbf{D}_k^{CCB}$ contains the entries of $\mathbf{D}_k^{CB}$ associated to clustered corrupted blocks, while the second one, denoted in the following with $\mathbf{D}_k^{ICB}$ contains the contributions referred to the remaining, isolated, blocks. A block (i,j) for which $D_k^{CB}[i,j] > 0$ is considered member of a cluster if at least for one of its 8 neighbours, let say (p,q) the following condition $D_k^{CB}[p,q] > 0$ holds.
The overall distortion index is finally computed as follows.

$$D_k^{Tot} = \frac{1}{N_{Blocks}}\left[a_{CCB} \left\| \mathbf{D}_k^{CCB} \right\|_{L_1} + \eta^{ICB}\left(\left\| \mathbf{D}_k^{ICB} \right\|_{L_1} \right) + a_{RL} \left\| \mathbf{D}_k^{RL} \right\|_{L_1} \right] \tag{12}$$

where

$$\eta^{ICB}(x) = \begin{cases} 0 & x \le \lambda_{ICB} \\ x & otherwise \end{cases}. \tag{13}$$

We note that the shrinking of the contribution of isolated corrupted blocks allows to mitigate the effects produced by misclassifications and to account for the lower sensitivity to artefacts confined to small areas compared to those interesting wider areas. The weights a_{CCB} and a_{RL} can be computed by fitting full reference video quality metrics with Eq. (12). Following this procedure, in the performed tests we obtained a_{CCB}=1 and a_{RL}=1/9.
Finally the no reference Video Quality Metric $nrMOS[k]$ is computed as

$$nrVQM[k] = \underset{k-M \le h \le k+M}{Median} \left\{ 5 - \sqrt{15 \frac{D_h^{Tot}}{c_0 + c_1 \overline{\rho^{10-90}}}} \right\}, \tag{14}$$

where $\overline{\rho^{10-90}}$ is the average correlation coefficient computed on the central part of the histogram (between 10-th and 90-th percentile) and c_1=0.56136 and c_2= 0. 78513 are normalizing factors obtained by fitting MOS obtained by subjective experiments as well as full reference Video Quality Metrics.

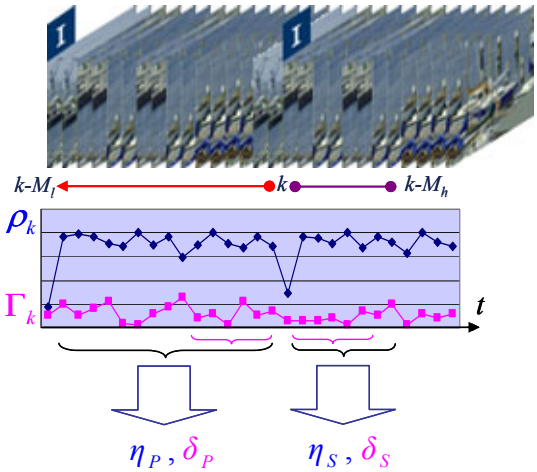

Fig.4. I-frame detection scheme.

4. NUMERICAL RESULTS

To evaluate the performances of the proposed method, a comparison has been performed between the proposed quality indicator and the full reference index VQMg(MSU Video Quality Measurement Tool) developed by "The Graphics & Media laboratory" - "Lomonosov Moscow State University" and available at the URL http://www.compression.ru/video/ . Several test sequences, with different content, have been used. a few of them are illustrated in Fig.5.
In figure 6 the results concerning the '*Taxi*' sequence affected by a packet loss in the range 1.9% (top), 3% (middle), and 6.7%(bottom) are reported. A possible evaluation of the goodness of the proposed method is the amount of overlapping between the proposed and the VQMg metric. It can be noticed that the proposed metric is overlapping quite well the VQMg one, especially in the 3% and in the 6.7% case. While the shape of the two curves is more or less the same, a small displacement between them is present. This source of mismatching is under analysis to understand if a particular content features or particular motion rates present in the video can cause such a behaviour.

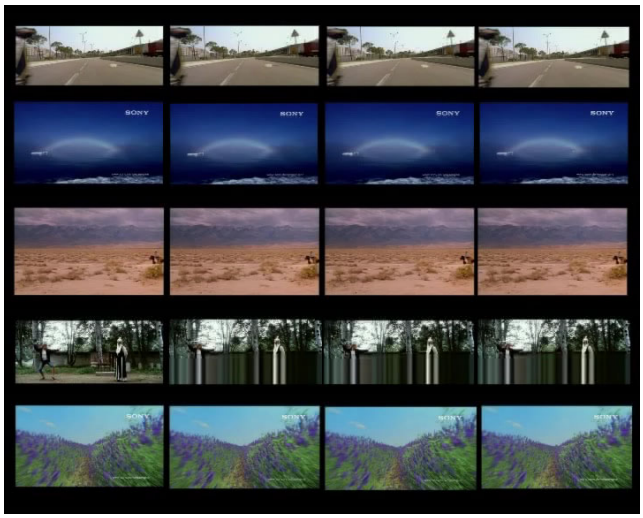
Fig. 5. Test sequences.

In the following we further discuss the results corresponding to the highest level of degradation. For the '*Taxi'* sequence (see Figure 6.c) the matching is quite good for almost the complete sequence. Only in the last part of it, the NR metric is over estimating the degradation rate with respect to the VQMg.

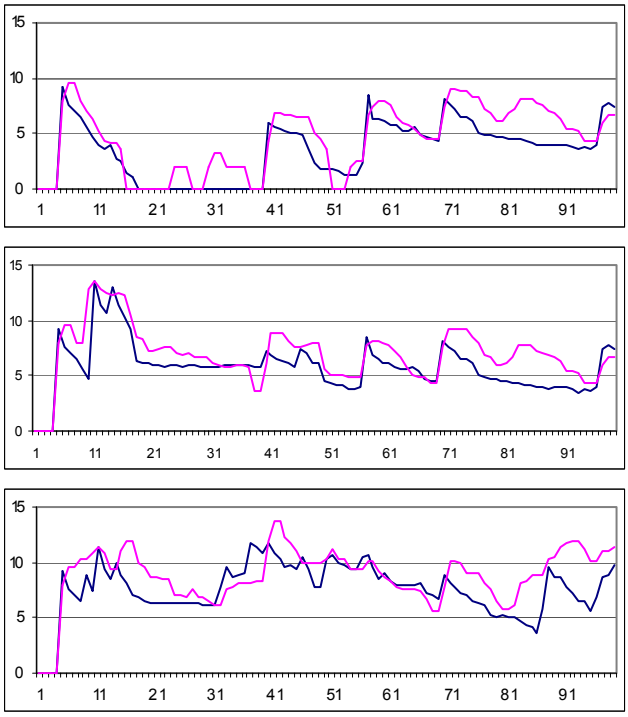

Fig. 6. *Taxi* sequence: nrVQM (magenta) versus VQMg (blue) at a) Packet loss = 1.9% (top), b) 3% (middle), c) 6.7%(bottom).

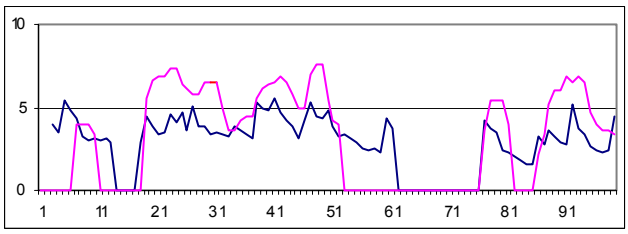

Fig. 7. *Field* sequence: nrVQM (magenta) versus VQMg (blue) at Packet loss = 6.4%.

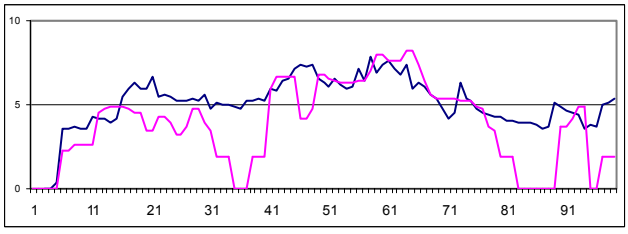

Fig. 8. "*Horse ride*" sequence: nrVQM (magenta) versus VQMg (blue) at Packet loss = 6.4%.

The same behavior can be notice for the sequence '*Field*' as reported in Figure 7. Almost for the whole sequence, the two indexes present the same behavior. There is a slight tendency in overestimating the video artifacts by the NR index. Only for a few frames the indication given 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. Also if the average behavior can be compared, among the 25th and the 38th frame the nrVOM indicator shows high degradation while VQMg shows only a slight degradation. The same different degradation rate can also be noticed in the last part of the sequence.

5. CONCLUDING REMARKS

In this paper a No Reference metric for assessing the quality video transmission 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 to evaluate the proposed metric. The overall analysis demonstrate the effectiveness of the nrVOM. By analyzing the results, it can be noticed that higher difference between the NR metric and the classical VQM is when long sequences with slow motion are presented to the metric. In this situation, the estimated degradation level caused by the previous frames is underestimated. Another problem is the automatic selection of the key frames. Due to estimation errors, if a fake key frame is extracted, the quality metric immediately decreases.

Also if the sequences are characterized by slow or almost null motion, a difference between the two metrics is present. In this case, the detected quality is really matching the quality behavior, while the metrics values may be different.

REFERENCES

[1] Z. Wang, A.C. Bovik, B.L. Evans, "Blind measurement of blocking artifacts in images*", IEEE Int. Conf. on Image Processing*, vol. 3, Sept. 2000, pp. 981-984

[2] Zhou Wang, Hamid R. Sheikh, Alan C. Bovik, "No-Reference perceptual quality assessment of JPEG compressed images", *IEEE Intl Conf. on Image Processing*, Sept. 2002, pp. 477-480.

[3] Xin Li, "Blind image quality assessment", IEEE Int. *Conf. on Image Processing*, Sept. 2002, pp. 449-452

[4] K.T. Tan, M. Ghambari, D.E. Pearson, "An objective measurement tool for MPEG video quality", *Signal Processing* 70 (1998),pp. 279-294.

[5] K.T. Tan, M. Ghambari, "A combinational automated MPEG video quality assessment model ", *Image Processing and its Applications*, Conf. Pub. no. 465 IEEE 1999, pp. 188-192.

[6] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, A. Neri "Objective quality evaluation of video services", in *Proc. VPQM 2006*, Scottsdale, Arizona, United States, January 2006), no page numbers (on CD-ROM).

[7] M. Carli, D. Guida, A. Neri, "No-reference jerkiness evaluation method for multimedia communications", in *Proc. SPIE Vol. 6059, Im. Qual. and System Performance III*; L.C. Cui, Y. Miyake; Eds, Jan. 2006, p. 350-359.

[8] M. Farias and S. Mitra, "No-Reference Video Quality Metric Based on Artifact Measurements," *ICIP 2005*, Genova, Italy, Sept. 2005.