Considering temporal variations of spatial visual distortions in video quality assessment

*Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba

***Abstract*—The temporal distortions such as flickering, jerkiness and mosquito noise play a fundamental part in video quality assessment. A temporal distortion is commonly defined as the temporal evolution, or fluctuation, of the spatial distortion on a particular area which corresponds to the image of a specific object in the scene. Perception along time of a spatial distortion can be largely modified by its temporal changes, such as increase or decrease of the distortions, or as periodic changes of the distortions. In this work, we have chosen to design a perceptual full reference video quality assessment metric by focusing on the temporal evolutions of the spatial distortions. As the perception of the temporal distortions is closely link to the visual attention mechanisms, we have chosen to first evaluate the temporal distortion at the eye fixation level. In this short-term temporal pooling, the video sequence is divided into spatio-temporal segments in which the spatio-temporal distortions are evaluated resulting in spatio-temporal distortion maps. Afterwards, the global quality score of the whole video sequence is obtained by the long-term temporal pooling in which the spatio-temporal maps are spatially and temporally pooled. Consistent improvement over existing video quality assessments methods is observed. Our validation has been realized with a dataset build from video sequences of various contents.**

***Index Terms*—Video quality assessment, temporal distortions, temporal pooling, spatio-temporal tube, visual fixation.**

I. INTRODUCTION

The purpose of an objective image or video quality evaluation is to automatically assess the quality of images or video in agreement with human quality judgments. Over the past few decades, image and video quality assessment has been extensively studied and many different objective criteria have been built. Video quality metric may be classified into Full Reference metrics (FR), Reduced Reference metrics (RR), and No Reference (NR). This paper is dedicated to the design of a FR video quality metric, for which the original video and the distorted video are both required. One obvious way to implement video quality metrics is to apply a still image quality assessment metric on a frame-by-frame basis. The quality of each frame is evaluated independently, and the global quality of the video sequence can be obtained by a simple time average, or with a Minkowski summation of per-frame quality. However, a more sophisticated approach would model the temporal aspects of the Human Visual System (HVS) in the design of a quality metric. A number of methods have been proposed to extend the HVS features towards the temporal dimension and motion [1]–[5].

In the scope of the error sensitivity-based approaches, Van den Branden Lambrecht *et al.* [2], [4] has extended the HVS modelling into the time dimension by modelling the temporal dimension of the Contrast Sensitivity Function (CSF), and by generating two visual streams tuned to different temporal aspects of the stimulus from the output of each spatial channel. The two streams model the transient and the sustained temporal mechanisms of the HVS respectively, which play an important role in other metrics such as in [1], or in [5] where only sustained temporal mechanism is taken into account. But, in these metrics, the temporal variations of the errors are not considered.

The approach of Wang *et al.* [6]–[8] was different. Rather than assessing the error in term of visibility Wang *et al.* used structural distortion [6] as an estimate of perceived visual distortion. This approach had been extended to the temporal dimension by using motion information in a more [7] or less [8] sophisticated way. In [8], Wang *et al.* proposed an heuristic weighting model which take into account the fact that the accuracy of the visual perception is significantly reduced when the speed of the motion is large. In [7], the errors are weighted by the perceptual uncertainty based on the motion information, which is computed from a model of human visual speed perception [9]. However, these metrics do not take into account the temporal variations of the errors.

Another approach is the one from the National Telecommunications and Information Administration (NTIA), which has developed a Video Quality Model (VQM) [10] adopted by the ANSI as a U.S. national standard [11], and as international ITU Recommendations [12], [13]. NTIA's research has focused on developing technology independent parameters that model how people perceive video quality. These parameters have been combined using linear models. The *General Model* contains seven independent parameters. Four parameters are based on features extracted from spatial gradients of the Y luminance component. Two parameters are based on features extracted from the vector formed by the two (C_B, C_R)

*A. Ninassi is both with Thomson Corporate Research, 1 Avenue Belle Fontaine, 35511 Cesson-Sevigne, France; and with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), Nantes, France (phone. +33(0)299273830; fax: +33(0)299273015; e-mail: alexandre.ninassi@thomson.net).

Olivier Le Meur is with Thomson Corporate Research, 1 Avenue Belle Fontaine, 35511 Cesson-Sevigne, France (phone. +33(0)299273654; fax: +33(0)299273015; e-mail: olivier.le-meur@thomson.net).

Patrick Le Callet is with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) UMR 6597 CNRS, Ecole Polytechnique de l'Universite de Nantes, rue Christian Pauc, La Chantrerie, 44306 Nantes, France (phone. +33(0)240683047; fax: +33(0)240683232; e-mail: patrick.lecallet@univ-nantes.fr).

Dominique Barba is with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) UMR 6597 CNRS, Ecole Polytechnique de l'Universite de Nantes, rue Christian Pauc, La Chantrerie, 44306 Nantes, France (phone. +33(0)240683022; fax: +33(0)240683232; e-mail: dominique.barba@univ-nantes.fr).

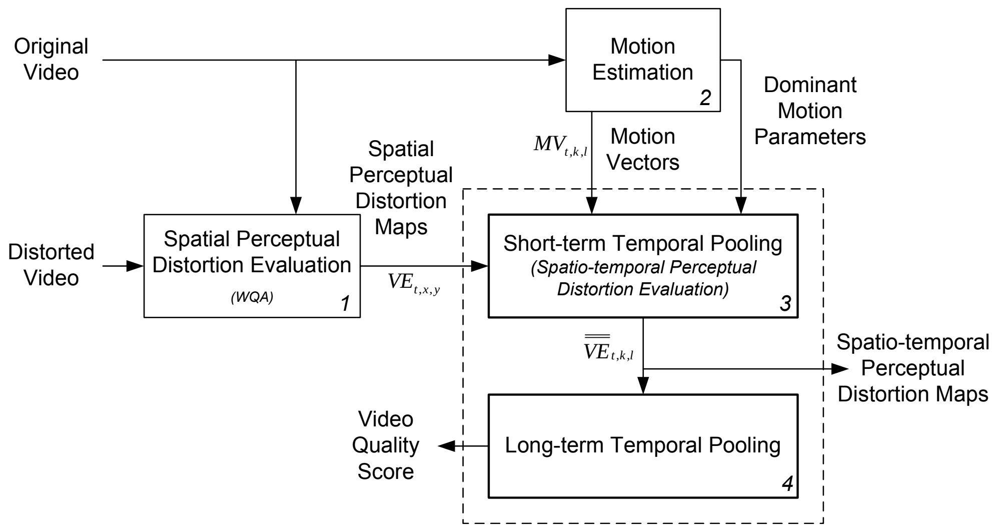

Fig. 1. Block diagram of the proposed video quality assessment system.

chrominance components. One parameter is based on the product of features that measure contrast and motion, both of which are extracted from the Y luminance component. This last parameter deals with the fact that perception of spatial impairments can be influenced by the amount of motion, but once again the temporal variations of spatial impairments are not considered.

The effects of the introduction of the temporal dimension in a quality assessment context can be addressed in a different way. A major consequence of the temporal dimension is the introduction of temporal effects in the distortions such as flickering, jerkiness and mosquito noise. Broadly speaking, a temporal distortion can be defined as the temporal evolution, or fluctuation, of the spatial distortion on a particular area which corresponds to the image of a specific object in the scene. Perception along time of a spatial distortion can be largely modified (enhanced or lessened) by its temporal changes. The time frequency and the speed of the spatial distortion variations, for instance, can significantly influence the human perception. The temporal variations of the distortions has been study in the scope of continuous quality evaluation [14], [15], where objective quality metrics try to mimic the temporally varying subjective quality of video sequences as recorded by subjective continuous evaluation such as Single Stimulus Continuous Quality Evaluation (SSCQE). In [15], the existence of both a short-term and a long-term mechanisms in the temporal pooling of the distortions is introduced. The short-term mechanisms is a smoothing step of per-frame quality scores, and the long-term mechanisms is addressed by a recursive process on the smoothed per-frame quality scores. This process includes perceptual saturation and asymmetrical behavior.

In this work, we have chosen to address the effects of the introduction of the temporal dimension by focusing on the temporal evolutions of the spatial distortions. Then, the question is how does a human observer perceive a temporal distortion?

The perception of the temporal distortions is closely link to the visual attention mechanisms. HVS is intrinsically a limited system. The visual inspection of the visual field is performed through many visual attention mechanisms. The eye movements can be mainly decompose into three types [16]: saccades, fixations and smooth pursuits. Saccades are very rapid eye movements allowing human to explore the visual field. Fixation is a residual movement of the eye when the gaze is fixed on a particular area of the visual field. Pursuit movement is the ability of the eyes to smoothly track the image of a moving object. Saccades allow human to mobilize the visual sensory resources (i.e. all parts of the HVS dedicated to processing of the visual signal coming from the central part of the retina: the fovea) on the different parts of a scene. Between two saccade periods a fixation (or smooth pursuit) occurs. When a human observer assesses a video sequence, different spatio-temporal segments of the video sequence are successively assessed. These segments are spatially limited by the area of the sequence projected on both the fovea and the perifovea. Even if the perifovea plays a role in the perception of the temporal distortion, we have chosen to simplify the problem by using a foveal model. Motion information is essential to perform the temporal distortions evaluation of a moving object, because the eye movement is very likely a pursuit in this situation. In that case, the evaluation of the temporal distortions must be done according to the apparent movement of this object. Furthermore, these segments are temporally limited by the fixation duration, or by the smooth pursuit duration. The perception of a temporal distortion likely happens during a fixation, or during a smooth pursuit. The fixation duration being shorter than the smooth pursuit duration, the temporal distortions must be evaluated first at the eye fixation level. This short-term evaluation constitutes the first stage of our approach. This stage then is completed by a long-term evaluation in which the global quality of the whole sequence is evaluated from the quality perceived over each fixation.

In this paper, a objective video quality assessment method is proposed. The spatio-temporal distortions are evaluated

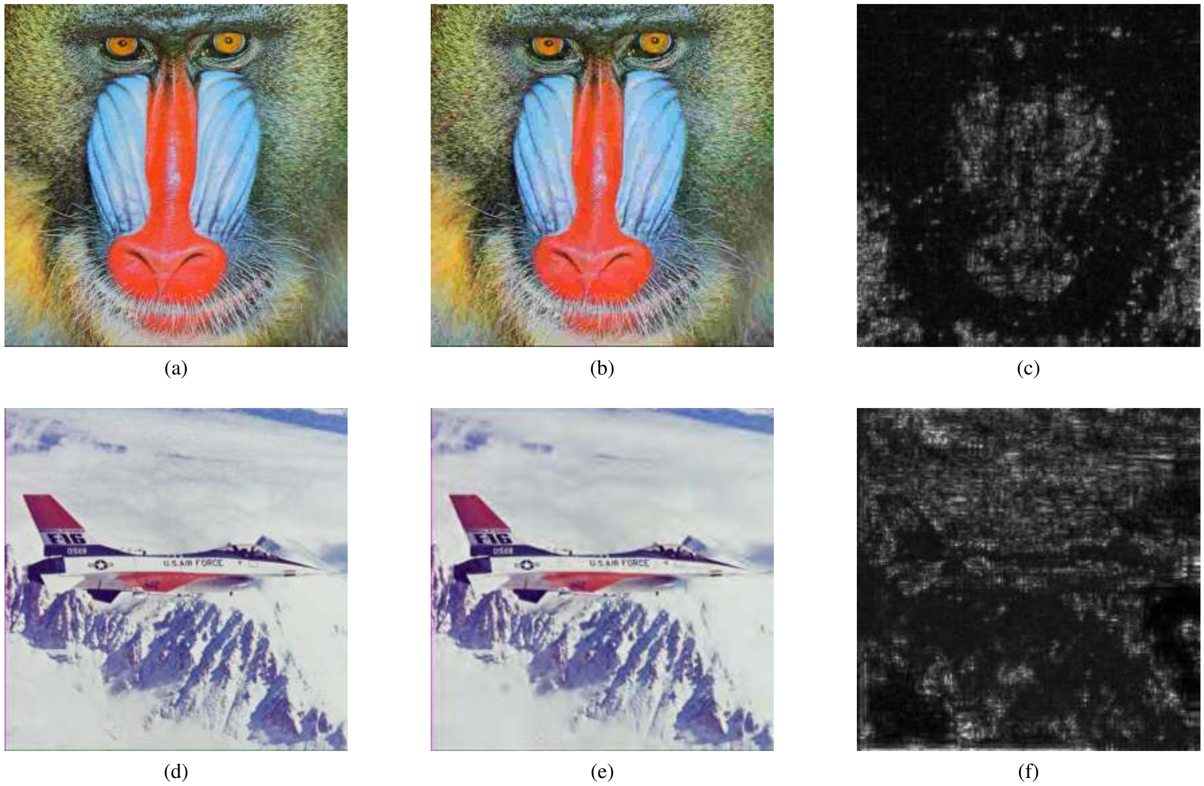

Fig. 2. Examples of WQA perceptual distortion maps: (a) and (d) are original Mandrill and Plane respectively; (b) is JPEG compressed Mandrill image; (c) is WQA perceptual distortion map of JPEG compressed Mandrill image; (e) is JPEG2000 compressed Plane image; (f) is WQA perceptual distortion map of JPEG2000 compressed Plane image. In (c) and (f), brightness indicates the magnitude of the perceptual distortion (black means no perceptual distortion).

through a temporal analysis of spatial perceptual distortion maps. The spatial perceptual distortion maps are computed for each frame with a wavelet based quality assessment (WQA) metric developed in a previous study [17]. This paper is decomposed as follows. In section II, the new video quality assessment metric (VQA) is presented. In order to investigate its efficiency, the VQA metric is compared with subjective ratings and two state-of-the-art metrics (VSSIM [8], VQM [10]) in section III. Finally conclusions are drawn.

II. Video quality assessment method

In the proposed video quality assessment system, the temporal evolution of the spatial distortions is locally evaluated, at short-term, through the mechanisms of the visual attention. The mechanisms of the visual attention indicate that the HVS integrates most of the visual information at the scale of the fixations [16]. So, the spatio-temporal distortions are locally observed and measured for each possible fixation. It does not make sense to evaluate the distortion variations on a period longer than the fixation duration, because it does not happen in the reality. The duration of 400 ms is chosen in accordance to the average duration of the visual fixation. This is the most simple and straightforward solution. A better solution, but much more complex, would be to adjust this value depending on the local spatial and temporal properties. A rather simple content, such as flat areas, probably requires less attentional resources than a more complex area [18]. Moreover, a smooth pursuit movement can be longer than a fixation duration. The complexity as well as the validation of such a solution still remains an issue.

Since the variations of the spatial distortions are evaluated locally according to where humans gaze, a special attention must be paid to the moving objects. In the case of a moving object, the quality of its rendering cannot be evaluated if it is not well stabilized on the fovea, which means that eye movement is a pursuit. Consequently, the evaluation of the temporal distortions must take into account the motion information, and the *locality of evaluation* must be motion compensated. These spatio-temporal segments of the sequence, evaluated by human observer during fixations, can be roughly linked to spatio-temporal distortion tubes (cf. section II-B1). These structures contain the spatial distortion variations for each possible fixation.

The description of the proposed method is divided into three subsections. The general architecture of the proposed metric is presented in section II-A. Section II-B is devoted to the evaluation of the spatio-temporal distortions at the eye fixation level. Finally, the evaluation of the temporal distortion on the whole video sequence is described in section II-C.

A. General architecture

The proposed video quality assessment system is composed of four steps as shown in Fig. 1. In the first step, numbered 1 in Fig. 1, for each frame t of the video sequence, a spatial perceptual distortion map $VE_{t,x,y}$ is computed. Each site (x, y) of this map encodes the degree of distortion that is perceived at the same site (x, y) between the original and the distorted frame. In this first step, there is no temporal consideration. In this work, the spatial perceptual distortion maps are obtained through the WQA metric developed in our previous work [17]. The WQA metric is a still image quality metric, which is based on a multi-channel model of HVS. HVS model of the low-level perception used in this metric includes subband decomposition, spatial frequency sensitivity, contrast masking and semi-local masking. The subband decomposition is based on a spatial frequency dependent wavelet transform. The spatial frequency sensitivity of the HVS is simulated by a wavelet CSF derived from Daly's CSF [19]. Masking effects include both contrast masking and semi-local masking. Semi-local masking allows to consider the modification of the visibility threshold due to the semi-local complexity of an image. The objective quality scores computed with this metric are well correlated with subjective scores [17]. The WQA distortion maps of a JPEG compressed image, and of a JPEG2000 compressed image, are shown in Fig. 2.

The second step, numbered 2 in Fig. 1, performs the motion estimation, in which the local motion between two frames are estimated, as well as the dominant motion. This step is achieved with the use of a classical Hierarchical Motion Estimator (HME). The local motion is a block-based motion estimation (block 8×8). The motion estimated is expected to be as close as possible to the real apparent movement. Local motion and dominant motion is used to construct the spatio-temporal structure (spatio-temporal tube) in which the spatio-temporal distortions are evaluated. The local motion is used to track a moving object in the past, and the dominant motion is used to determine the temporal horizon on which the object can be tracked (appearance or disappearance of the object). Local motion $\overrightarrow{V}_{local}$ at each site (x, y) of an image (or the motion vector) is produced by a hierarchical block matching. It is computed through a series of levels (different resolution), each providing input for the next. Dominant motion corresponds the motion of the camera. To estimate the global transformation that two successive images undergo, the dominant motion or the global transformation is estimated from the previous estimated local motion. The displacement $\overrightarrow{V}_{\Theta}(x, y)$, at site (x, y) related to a motion model parametrized by Θ is given by a 2D affine motion model:

$$\overrightarrow{V}_{\Theta}(s) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix}, \qquad (1)$$

where $\Theta = [a_1, a_2, a_3, a_4, a_5, a_6]$ represents the 2D affine parameters of the model. The affine parameters are computed with a popular robust technique based on the M-estimators [20].

Temporal evaluation of the quality is performed through steps 3 and 4. Step 3 realizes the short-term evaluation of the temporal distortions, in which the spatio-temporal perceptual distortion maps $\overline{\overline{VE}}_{t,k,l}$ are computed from the spatial distortion maps and the motion information. For each frame of the video sequence, a temporal perceptual distortion map is computed. Each site (k, l) of this map encodes the degree of distortion that is perceived between the block (k, l) of the original frame and the block (k, l) of the distorted frame including temporal considerations (temporal distortions, etc.). The time scale of this evaluation is the one of the human eye fixation [21] (around 400ms). This step is accurately described in section II-B. Step 4 performs the long-term evaluation of the temporal distortions, in which the quality score for the whole video sequence is computed from the temporal perceptual distortion maps. Section II-C will describe this last part.

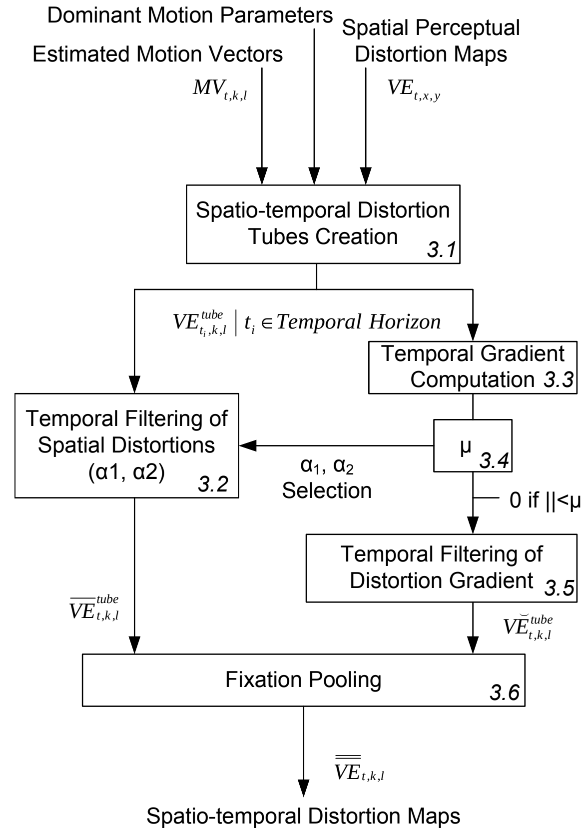

Fig. 3. Block diagram of the Spatio-temporal Perceptual Distortion Evaluation.

B. Spatio-temporal distortion evaluation at the eye fixation level

Spatio-temporal distortion evaluation is a complex problem. The purpose of this step is to perform the short-term evaluation of the temporal distortions at the eye fixation level. The video sequence must be divided into spatio-temporal segments corresponding to each possible fixation. It means that a fixation can start at every time t, and every site (x, y) of the sequence. At the eye fixation level, the temporal distortions evaluation depends both on the mean distortion level, and on the temporal variations of distortions. The temporal variations of distortions have to be smoothed to obtain the mean distortion level that is perceptible during fixation. The insignificant temporal variations of distortions have to be discard, and only the most perceptually important temporal variations of distortions have to be taken into account. Fig. 3 gives the main components involved in this evaluation. The first component (3.1) is dedicated to the creation of the spatio-temporal structures required to analyze the variation of the distortion during a fixation, i.e.

the spatio-temporal tubes. The process is then separated into two parallel branches. The purpose of the first branch is to evaluate a mean distortion level during the visual fixation. The aim of the second branch is to evaluate the distortion variations occurring during a fixation, and at which humans are the most sensitive. Next, these two branches are merged resulting in the spatio-temporal distortion maps.

1) Spatio-temporal tube creation: In step 3.1, the Spatio-temporal Distortion Tubes are created. The aim of this step is to divided the video sequence into the spatial-temporal segments corresponding to each possible fixation. A spatio-temporal distortion tube is computed for each block of a frame t. A spatio-temporal distortion tube is a spatio-temporal structure containing the past of a block in terms of spatial distortion (cf. Fig. 4). It means that this structure contains the different distortion values of this block for each past frame over a specific temporal horizon. The motion vectors $MV_{t,k,l}$ are used to find the different position of a block (k, l) in the past frames of its temporal horizon. The positions of the past blocks are then motion compensated in order to have its real trajectory. The temporal horizon is limited to 400ms, and can be shortered if the object inside the block appears or disappears in the image of the scene. To detect appearance or disappearance of an object, each block is classified by comparing its motion with the parametric representation of the dominant motion: each block is either inlier or outlier to the dominant motion. A modification of the classification (inlier/outlier) of a block between two consecutive frames means appearance or disappearance of the object it belongs to, and so indicates the limit of the temporal horizon of this block.

2) Temporal filtering of the spatial distortion in the tube: Step 3.2 realizes the Temporal Filtering of Spatial Distortions. The goal of this step is to obtain a mean distortion level over the fixation duration. The large temporal variations of distortions are the most annoying for observers, and their contribution should be more important than limited temporal variations of distortions. The spatial distortions are then temporally filtered in each tube of a frame t. The temporal filter is a recursive filter. The characteristics of the filter are modified according to the importance of the temporal variations of distortions. The contribution of the large temporal variations of the distortions is increased compare to the contribution of the limited temporal variations of distortions. Time constant of this filter changes depending on the value of the corresponding distortion gradient value (cf. step 3.4). Time constant $\alpha_1 = 200ms$ is used if the absolute value of the distortion gradient value is greater than μ, otherwise $\alpha_2 = 400ms$ is used. The output of this step is the map $\overline{VE}^{tube}_{t,k,l}$ where each block (k, l) is the result of the temporal filtering of the spatial distortions in each tube finishing at frame t.

3) Temporal distortion evaluation in the tube: The purpose of step 3.3 is to evaluate the temporal variation of distortions. The temporal gradients of the spatial distortions in the tubes are computed, in order to evaluate the most perceptually important temporal variations of distortions during fixations. In a tube, the distortion gradient $\nabla VE^{tube}_{t_i,k,l}$ at time t_i is computed as follows:

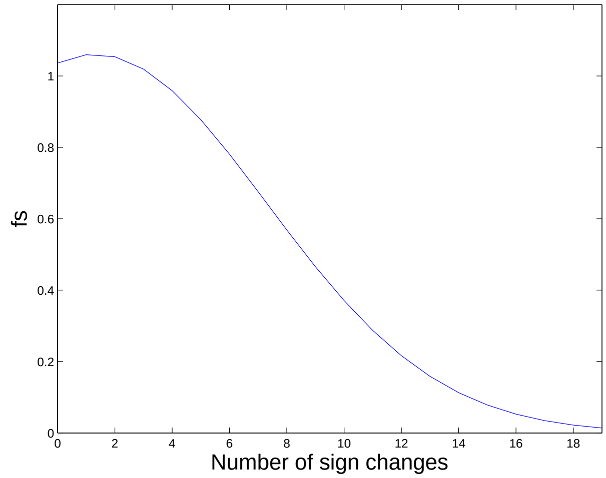

Fig. 5. Plot of the fs response. The function reaches his maximum around one sign change of the distortion gradients per fixation.

$$\nabla VE^{tube}_{t_i,k,l} = \frac{\delta VE^{tube}_{t_i,k,l}}{\delta t} \Bigg| \begin{array}{l} \delta t = t_i - t_{i-1} \\ t_i \in TemporalHorizon \end{array} , \quad (2)$$

where $VE^{tube}_{t_i,k,l}$ is the distortion value at instant t_i.

The limited temporal variations of distortions which are probably not annoying must not be taken into account. The aim of step 3.4 is to delete them. In this step, a thresholding operation is performed on the absolute value of the gradient values. The purpose is to reduce the weight of the limited temporal variations of distortions (below μ) compare to large temporal variations of distortions (above μ). If the absolute value of the gradient is lower than μ the gradient value becomes 0. This thresholding operation is also used to manage the temporal filtering of the step 3.2, as described in the previous section.

The characteristics of temporal distortions, such as time frequency and amplitude of the variations, impact the perception. The purpose of step 3.5 is to evaluate the perceptual impact of temporal distortions according to the characteristics of the temporal variations of distortions. In this step, the temporal filtering of distortion gradient is realized, in which the distortion gradients are temporally filtered in each tube of a frame t. This temporal filtering operation is achieved by counting the number of sign changes of the distortion gradients $nS^{tube}_{t,k,l}$ along the tube duration. The maximal distortion gradient $Max\nabla VE^{tube}_{t,k,l}$ is computed, and used as maximal response of the filter. The temporal filtering result is obtained by:

$$\breve{VE}^{tube}_{t,k,l} = Max\nabla VE^{tube}_{t,k,l} \cdot fs(nS^{tube}_{t,k,l}) , \quad (3)$$

where fs is the response of the filter depending on the number of sign changes:

$$fs(n) = \frac{g_s}{\sigma_s\sqrt{2\pi}} \cdot e^{-\frac{(n-\mu_s)^2}{2\sigma_s^2}} , \quad (4)$$

The response of the function $fs(n)$ is given Fig. 5. Function $fs(n)$ gives more importance to temporal distortion at medium frequencies than at low or high frequencies. The HVS is the

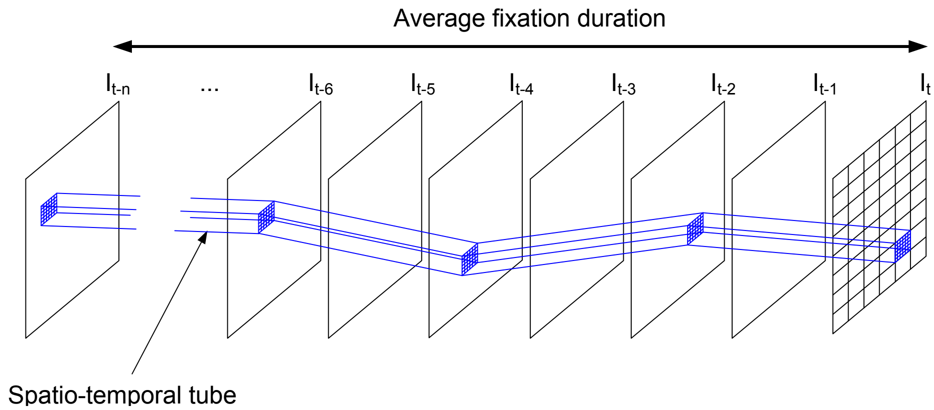

Fig. 4. Spatio-temporal tube illustration. The past trajectory of a block of the frame I_t is reconstituted by using the past motion vectors of this block.

most sensitive to temporal variations around $2cy/s$, which correspond to about one sign change by fixation duration. The output of this step is the map $\breve{VE}^{tube}_{t,k,l}$ where each block (k,l) is the result of the temporal filtering of the distortion gradient in each tube finishing at frame t.

The results coming from the two branches are then mixed together in step 3.6. This step performs the Fixation Pooling, in which the map $\overline{VE}_{t,k,l}$ and the map $\breve{VE}_{t,k,l}$ are merged in order to obtain the final spatio-temporal distortion map $\overline{\overline{VE}}_{t,k,l}$. If there is no temporal variations of distortions in the video sequence the final map $\overline{\overline{VE}}_{t,k,l}$ is equal to the $\overline{VE}_{t,k,l}$ map. But when temporal variations of distortions occurred, the $\overline{VE}_{t,k,l}$ map are consolidated by the temporal variation evaluation of the map $\breve{VE}_{t,k,l}$. This map is computed according to the following relation:

$$\overline{\overline{VE}}_{t,k,l} = \overline{VE}_{t,k,l} \cdot (1 + \beta \cdot \breve{VE}_{t,k,l}), \quad (5)$$

where value of parameter β is empirically deduced from experiments on synthetic sequences. These experiments aimed at obtaining relevant spatio-temporal distortion maps from synthetic sequences with synthetic distortions. It was achieved by setting the value β at 3.

Until now, the impact of the temporal distortions has been evaluated at the fixation level, resulting in the final spatio-temporal distortion maps $\overline{\overline{VE}}_{t,k,l}$. However, a human observer scores a video sequence using the impairments he perceived during the whole sequence. This is the issue addressed by the next section.

C. Temporal distortion evaluation on the whole video sequence

The long-term temporal pooling is the final stage that allows to construct the global objective quality score of a video sequence. The global objective quality score depends both on the mean distortion level over the whole sequence, and on the temporal variations of distortions over the whole sequence. The temporal variations of the distortions along a video sequence play an important part in the global score, and a mean distortion level on the whole sequence is not sufficient to evaluate the quality of the video. The evaluation process of

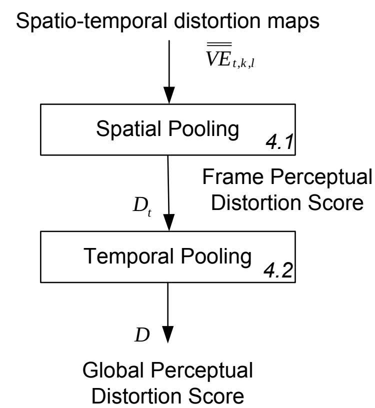

Fig. 6. Block diagram of the long-term temporal pooling.

a human observer could be sum up by the following sentence *quick to criticize and slow to forgive*. So, the overall temporal distortions evaluation of the whole video sequence is divided in two steps as shown in Fig. 6.

1) Spatial pooling: The purpose of the step 4.1 is to obtain a perceptual distortion score for each frame. A per-frame perceptual distortion score D_t is computed from the spatio-temporal distortion map of each frame through a classical Minkowski summation:

$$D_t = \left(\frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(\overline{\overline{VE}}_{t,k,l} \right)^{\beta_s} \right)^{\frac{1}{\beta_s}}, \quad (6)$$

where K and L are the height and the width of the spatio-temporal distortion maps respectively (i.e. the vertical and the horizontal number of blocks in the original frame), and β_s is the Minkowski exponent ($\beta_s = 2$).

2) Temporal pooling: The global objective perceptual distortion score, called D, depends both on the average of distortion level over the whole sequence, and on the temporal variations of distortions over the whole sequence. The perceptual distortion is increase by the temporal variations of distortions over the whole sequence. The proposed temporal pooling contains two main elements: perceptual saturation and asymmetric behavior. There are limitations in viewer's ability to observe any further changes in the frame quality after it exceeds certain thresholds, either toward better or worse quality [14]. This is what we called the perceptual saturation. The asymmetrical behavior is the fact that humans are better able to remember unpleasant experiences than pleasant moments, and also experience with great intensity of feelings from disliked situations compared to favorable situations [14].

The global perceptual distortion score D of a video is computed from every per-frame perceptual distortion scores D_t, as the sum ($D = \bar{D}+\Delta_D$) of the time average of distortion $\bar{D}$, and a term representing the variation of distortions along the sequence Δ_D. But to limit the influence of too high distortion variations, D is computed with a saturation effect as follows:

$$D = \begin{cases} \bar{D} + \Delta_D & \text{for } \Delta_D < \lambda_1 \cdot \bar{D} \\ \bar{D} + \lambda_1 \cdot \bar{D} & \text{for } \Delta_D \geq \lambda_1 \cdot \bar{D} \end{cases} . \quad (7)$$

The global distortion score D increases linearly with the temporal variation until a saturation threshold value proportional to $\bar{D}$. The term Δ_D favours the most important variations of distortions, and is computed as follows:

$$\Delta_D = \lambda_2 \cdot avg_{n\ \%}(abs(\nabla' D_t))\,, \quad (8)$$

where $\nabla' D_t$ is the temporal gradient of the per-frame distortion values D_t after the asymmetrical transformation of the gradient values, $abs(X)$ is the absolute value of X, and $avg_{n\ \%}(X)$ is the average of X values above the *n*th percentile of X. The asymmetrical transformation of the gradient values is computed as follows:

$$\nabla' D_t = \begin{cases} \lambda_3 \cdot \nabla D_t & \text{for } \nabla D_t < 0 \\ \nabla D_t & \text{for } \nabla D_t \geq 0 \end{cases} \Bigg| \ \lambda_3 \leq 1, \quad (9)$$

where value of λ_3 controls the asymmetrical behavior. If $\lambda_3 < 1$, more weight is given to the distortion increases than to distortion decreases.

Finally, the global quality score VQA is computed from perceptual distortion score D by using a psychometric function, as recommended by the Video Quality Expert Group (VQEG) [22]:

$$\text{VQA} = \frac{b1'}{1 + e^{-b2' \cdot (D - b3')}}\,, \quad (10)$$

where $b1'$, $b2'$ and $b3'$ are the three parameters of the psychometric function. These psychometric function is also used to compared VQA, with state-of-the-art metrics (cf. section III-C).

III. Experimentation

A. Video database

1) Participants: Thirty six compensated participants are asked to assign each sequence with a quality score, indicating the extent to which the artifacts were more or less annoying. Prior to the test, subjects were screened for visual acuity by using a Monoyer optometric table. Besides, test for normal color vision were performed using Ishihara's tables. So, all observers had normal or corrected to normal visual acuity (Monoyer test), and normal color perception (Ichihara test). All were inexperienced observers (not familiar with video processing) and naive to the experiment.

2) Method: The standardized method DSIS (Double Stimulus Impairment Scale) is used to determine the Mean Opinion Score (MOS). In DSIS, each observer views an unimpaired reference video sequence followed by its impaired version, each lasting 8s. Experiments were conducted in normalized viewing conditions [23]. The scale used to score the distortion level is composed of 5 distortion grades:

- imperceptible (MOS=5);
- not annoying (MOS=4);
- slightly annoying (MOS=3);
- annoying (MOS=2);
- very annoying (MOS=1).

3) Stimuli: The video database is build from ten unimpaired video sequences of various contents as illustrated in Fig. 7. The spatial resolution of video sequence is 720x480 with a frequency of 50Hz in a progressive scan mode. Each clip lasts 8s. They were displayed at a viewing distance of four times the height of the picture (66 cm). These video sequences have been degraded by using a H.264/AVC compression scheme at five different bitrates, resulting in fifty impaired video sequences. The five different bitrates were chosen in order to generate degradations all over the distortion scale (from imperceptible to very annoying).

The impairments produced by the encoding are evidently neither spatially nor temporally uniform, and therefore depend on each video content. Fig. 8a illustrates the temporal variations of the quality through the scores given by the WQA metric (cf. Section II). This example indicates that the quality of the sequences varies from frame to frame, which is a clue on the presence of temporal distortions.

B. Video quality metrics tested

Several quality assessment metrics have been compared with subjective scores (MOS):

- The proposed video quality metric VQA (achromatic version),
- The usual PSNR (achromatic version). The PSNR global score is the temporal average of the per-frame PSNR.
- VSSIM developed by Wang *et al.* [8]. We used all the parameters described in [8], except for the normalization factor K_M of the frame motion level which has been adapted to our frame rate.
- VQM developed by NTIA [10]. Among the different models of VQM, we have chosen to use the *General Model* which is considered to be the most accurate. The *General Model* is known as metric H in the Video Quality Experts Group (VQEG) Phase II Full Reference Television (FR-TV) tests [24].

In order to evaluate the different steps of the VQA metric two alternative video perceptual distortion scores (VQA_1, VQA_2) are computed in addition to the global quality score.

The first intermediate video perceptual distortion score is a purely spatial quality score called VQA_1; It is computed from the spatial distortion maps of the still image metric WQA [17] as follows:

$$VQA_1 = \frac{1}{T}\sum_{t=1}^{T} d_t\,, \quad (11)$$

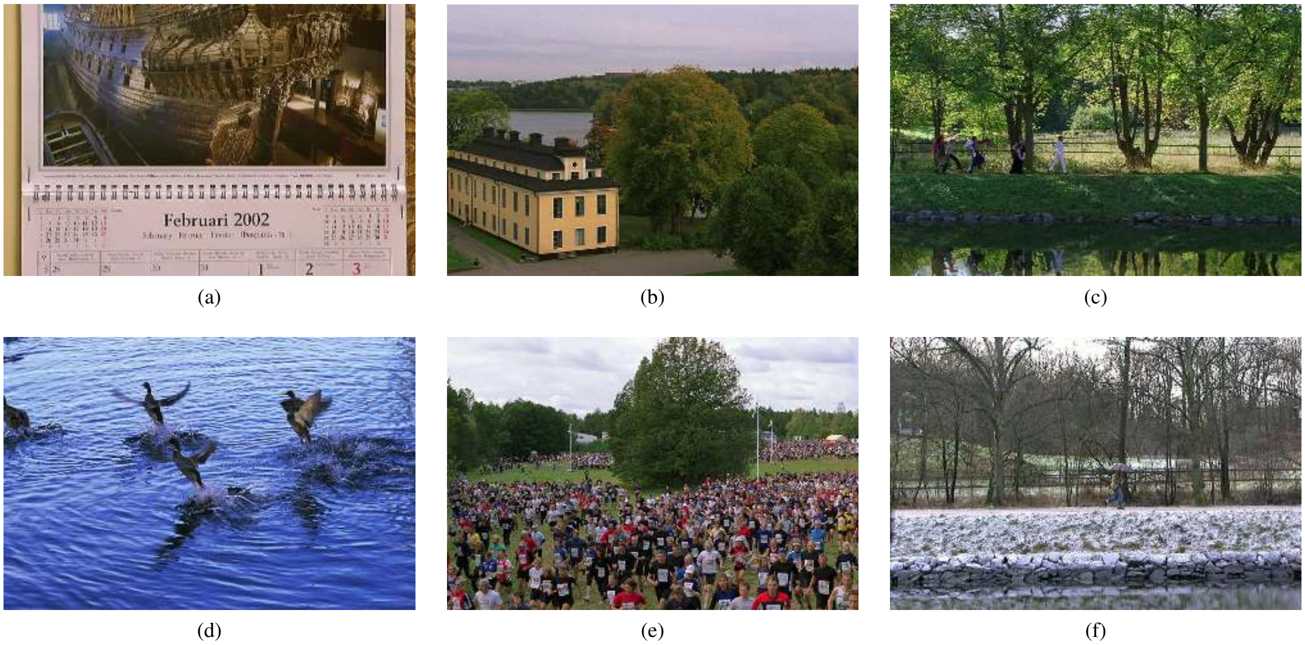

Fig. 7. Examples of video sequences from the database. (a) MobCal, (b) InToTree, (c) ParkJoy, (d) DucksTakeOff, (e) CrowdRun, and (f) ParkRun.

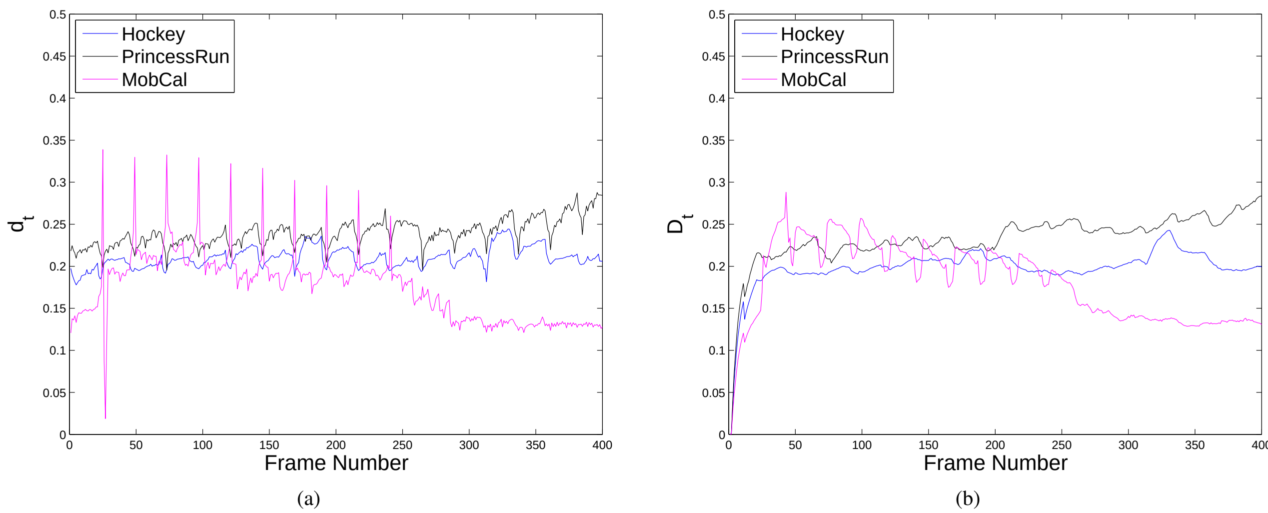

Fig. 8. Temporal evolution of the per-frame distortion score d_t (a), and the per-frame distortion score D_t (b) for the three impaired sequences of the database: Hockey (MOS=1.4), PrincessRun (MOS=2.6) and MobCal (MOS=1.3). The horizontal scale is the frame number, and the vertical scale is a distortion scale, which goes from 0 (best quality) to 0.5 (worst quality).

where T is the total number of frames and d_t is a frame score computed as follows:

$$d_t = \left(\frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(VE_{t,k,l} \right)^{\beta_s} \right)^{\frac{1}{\beta_s}}, \tag{12}$$

where $VE_{t,k,l}$ are the spatial distortion maps computed with WQA [17], K and L are the height and the width of the spatial distortion maps, respectively, and β_s is the Minkowski exponent.

In the second intermediate quality score called VQA_2 the fixation temporal pooling is disabled, which means that perceptual distortion score is computed from the long-term temporal pooling (Eq. 7) where D_t is replaced by d_t. D_t is the spatio-temporal per-frame distortion score (with the fixation temporal pooling), whereas d_t is the purely spatial per-frame distortion score (without the fixation temporal pooling).

Comparison between VQA_2 and VQA allows to evaluate the improvement due to spatio-temporal distortion evaluation at the eye fixation level (or short-term temporal pooling). On the other hand, comparison between VQA_1 and VQA allows to evaluate the improvement due to temporal pooling.

C. Results

As said previously, prior to evaluate the objective image quality measures, a psychometric function (Eq. 10) is used to transform the different objective quality score in predicted

MOS (MOSp), as recommended by VQEG [22]. The objective quality metrics are evaluated using three performance indicators recommended by VQEG [22]. The three performance indicators are the linear correlation coefficient (CC), the Spearman rank order correlation coefficient (SROCC) and the root-mean-square-error (RMSE). The CC between the MOS and MOSp scores provides an evaluation of the *prediction accuracy*. The SROCC between the MOS and MOSp is considered as a measure of *prediction monotonicity*.

TABLE I
PERFORMANCE COMPARISON OF QUALITY METRICS ON THE ENTIRE DATASET IN TERMS CC, SROCC AND RMSE.

Metrics (MosP)	CC	SROCC	RMSE
PSNR	0.516	0.523	0.982
VQM	0.854	0.898	0.597
VSSIM	0.738	0.758	0.773
VQA	0.892	0.903	0.519
VQA_1	0.831	0.872	0.638
VQA_2	0.834	0.863	0.633

Results, presented in Table I, are reported for the different metrics (VSSIM, VQM and VQA) and for the two intermediate quality score (VQA_1 and VQA_2) of VQA. PSNR results are provided for information and to allow readers to make their own opinions on the image dataset. Fig. 9 shows the scatter plots of the MOS/MOSp comparisons on the whole database given by PSNR, VSSIM, VQM, VQA, and by the two intermediate video quality score (VQA_1 and VQA_2) of VQA. The PNSR does not lead to a good prediction of quality as CC is only 0.516. This result gives a clue of how the quality of the video sequences of the database is difficult to evaluate.

The proposed method provides good results compared with the other approaches. It is important to mention that the parameters of the proposed method (VQA) were selected empirically, without any optimization process for the video database (λ_1=1, λ_2=10, λ_3=0.25, and n=95). Fig. 9 shows that the prediction performances of the metrics depend of the video content, and the video content does not disturb the different metrics in the same way. For example, VQM overestimates the quality of sequence Ducks, whereas VQA does not overestimate it. VQA overestimates the quality of sequences PrincessRun and Dance, and underestimates the quality of sequence Hockey. A possible explanation lies in the fact that the spatial distortions are also overestimated, and underestimated respectively. Fig. 8 shows that the per-frame distortion scores (d_t and D_t) of sequence Hockey are lower that the per-frame distortion scores of sequence PrincessRun, whereas the MOS of sequence Hockey are lower than the MOS of sequence PrincessRun. In these sequences, the temporal variations of the distortions could not explain the prediction errors of the quality. It shows that, in the proposed metric, the evaluation of temporal distortions is dependent of a good evaluation of the spatial distortion in the first step of the metric.

Comparison between the results from VQA_1, VQA_2 and VQA shows the positive contribution of the different steps of the proposed metric. The prediction improvement of the quality from the purely spatial quality score (VQA_1) to the spatio-temporal quality score (VQA) is significant. For example, ΔCC between these two configurations is $+0.061$. As expected, it shows that temporal distortions play an important part in video quality assessment. The prediction improvement of quality between VQA_2 and VQA shows the importance of the spatio-temporal distortion evaluation at the eye fixation level (short-term temporal pooling). This step seems fundamental prior to the long-term temporal pooling. One possible explanation is the smoothing effect of the short-term temporal distortion variations due to the fixation temporal pooling. This effect enables a better analysis of the long-term temporal distortion variations, by eliminating parasite temporal distortion variations. This smoothing effect is illustrated Fig. 8, by comparing the temporal variation of the per-frame distortion scores d_t (Fig. 8(a)) and D_t (Fig. 8(b)). The fixation temporal pooling does not only improve the prediction performance of the metric, but it also improves the relevancy of distortions maps.

TABLE II
PERFORMANCE COMPARISON OF VQA FOR DIFFERENT VALUES OF THE PARAMETERS λ_3 AND n, IN TERMS CC, SROCC AND RMSE. THE PARAMETERS λ_1 AND λ_2 ARE CHOSEN TO OPTIMIZE PREDICTION PERFORMANCES. RESULTS ON THE ENTIRE DATASET.

λ_3	*n*th percentile	CC	SROCC	RMSE
0	0	0.85	0.874	0.605
0	80	0.879	0.892	0.547
0	85	0.885	0.893	0.535
0	90	0.892	0.901	0.518
0	95	0.895	0.912	0.512
0.25	0	0.851	0.874	0.601
0.25	80	0.88	0.892	0.545
0.25	85	0.885	0.893	0.533
0.25	90	0.892	0.901	0.518
0.25	95	0.895	0.912	0.511
0.5	0	0.853	0.875	0.599
0.5	80	0.877	0.89	0.551
0.5	85	0.883	0.895	0.539
0.5	90	0.89	0.901	0.522
0.5	95	0.894	0.912	0.513
0.75	0	0.854	0.878	0.597
0.75	80	0.872	0.89	0.561
0.75	85	0.876	0.893	0.552
0.75	90	0.883	0.896	0.538
0.75	95	0.892	0.91	0.519
1	0	0.854	0.877	0.596
1	80	0.867	0.883	0.571
1	85	0.87	0.886	0.565
1	90	0.875	0.89	0.554
1	95	0.887	0.908	0.53

Results, presented in Table II, are reported for VQA and for different values of the parameters λ_3 and n. In this experiment, values of parameters λ_1 and λ_2 are selected to optimize prediction performances. The parameter λ_3 modifies the asymmetrical behavior of the long-term temporal pooling. The prediction modification of quality as function of λ_3 shows that long-term temporal pooling with symmetrical behavior (λ_3=1) leads to lower results than long-term temporal pooling with asymmetrical behavior. It is interesting to note that, to reach the best prediction performances, asymmetrical behavior must give, at least, twice more weight to the distortion increases than to distortion decreases. Besides, the choice of the empirical value of λ_3 (λ_3=0.25), seems to be a good option.

The parameter n modifies the weight given to maximal temporal gradients of per-frame distortion values. The worst results are obtained when all temporal gradients of per-frame distortion values are considered (n=0). The prediction modification of the quality as function of n shows that long-term temporal pooling takes advantage of using maximal temporal gradients of per-frame distortion values. Even if the best prediction performances are obtained with n=95, the results are robust to high value of n. It is interesting to note that n=95 means that the most important distortion variations occurring 5 percent of the time are the most significants in term of prediction performance. It strengthens the fact that distortion variations with high dynamic range must be considered.

TABLE III
PERFORMANCE COMPARISON OF VQA_2 FOR DIFFERENT VALUES OF THE PARAMETERS λ_3 AND n, IN TERMS CC, SROCC AND RMSE. THE PARAMETERS λ_1 AND λ_2 ARE CHOSEN TO OPTIMIZE PREDICTION PERFORMANCES. RESULTS ON THE ENTIRE DATASET.

λ_3	nth percentile	CC	SROCC	RMSE
0	0	0.831	0.872	0.638
0	80	0.831	0.872	0.638
0	85	0.831	0.872	0.638
0	90	0.831	0.872	0.638
0	95	0.832	0.869	0.636
0.25	0	0.831	0.872	0.638
0.25	80	0.831	0.872	0.638
0.25	85	0.831	0.868	0.638
0.25	90	0.832	0.867	0.636
0.25	95	0.834	0.863	0.633
0.5	0	0.831	0.872	0.638
0.5	80	0.831	0.868	0.638
0.5	85	0.832	0.866	0.636
0.5	90	0.833	0.87	0.634
0.5	95	0.839	0.866	0.624
0.75	0	0.831	0.872	0.638
0.75	80	0.832	0.868	0.636
0.75	85	0.833	0.867	0.635
0.75	90	0.834	0.869	0.633
0.75	95	0.846	0.869	0.611
1	0	0.831	0.872	0.638
1	80	0.832	0.867	0.636
1	85	0.833	0.87	0.634
1	90	0.835	0.869	0.632
1	95	0.85	0.865	0.605

Results are also reported for VQA_2 (without the fixation temporal pooling), presented in Table III, and for different values of parameters λ_3 and n. In this experiment, values of parameters λ_1 and λ_2 are selected to optimize prediction performance. The results show that long-term temporal pooling failed to improve the prediction performance when the fixation pooling is disabled. This observation is still valid whatever are the values of the parameters λ_1, λ_2, λ_3, and n. Consequently, the fundamental nature of fixation pooling step is enhanced by these results.

IV. CONCLUSION

This paper described a full reference video quality assessment metric. This metric focuses on the temporal variations of the spatial distortions. The temporal variations of the spatial distortions are evaluated both at the eye fixation level, and on the whole video sequence. The former, and the latter are assimilated to a short-term temporal pooling, and a long-term temporal pooling respectively.

Consistent improvement over existing video quality assessments methods is observed. CC between VQA and subjective scores is 0.892, and the prediction improvements in term of CC are $+73\%$, $+21\%$ and $+4\%$ compare to PSNR, VSSIM and VQM, respectively. Results also show the positive contribution of the different steps of the proposed metric. In particularly it shows that the short-term temporal pooling is fundamental prior to the long-term temporal pooling, as its use significantly improves the prediction performances of VQA. An interesting point of the proposed method is that the spatial distortion maps could be considered as an *input*. In this work, we used a still image quality metric WQA developed in a previous work to compute the spatial perceptual distortion map, but we can imagine to replace it by any still image quality metric which compute a spatial perceptual distortion map. The performance comparison of the proposed method, using different models to obtain the spatial perceptual distortion maps, could be an interesting investigation.

Further work includes further research to find a more sophisticated way to realize the long-term temporal pooling. In the proposed metric, we think that relevant information are lost in the spatial pooling step, and a more sophisticated long-term temporal should suppress this step.

REFERENCES

[1] S. Winkler, "A perceputal distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175–184, 1999.
[2] C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 2291–2294, 1996.
[3] A. B. Watson, J. Hu, and J. F. I. McGowan, "Dvq: A digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, pp. 20–29, 2001.
[4] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video coding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 9, pp. 766–782, 1999.
[5] M. Masry, S. S. Hemami, and S. Yegnaswamy, "A scalable wavelet-based video distortion metric and applications," *IEEE transactions on circuits and systems for video technology*, vol. 16, no. 3, pp. 260–273, 2006.
[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, 2004.
[7] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America A*, vol. 24, pp. B61–B69, 2007.

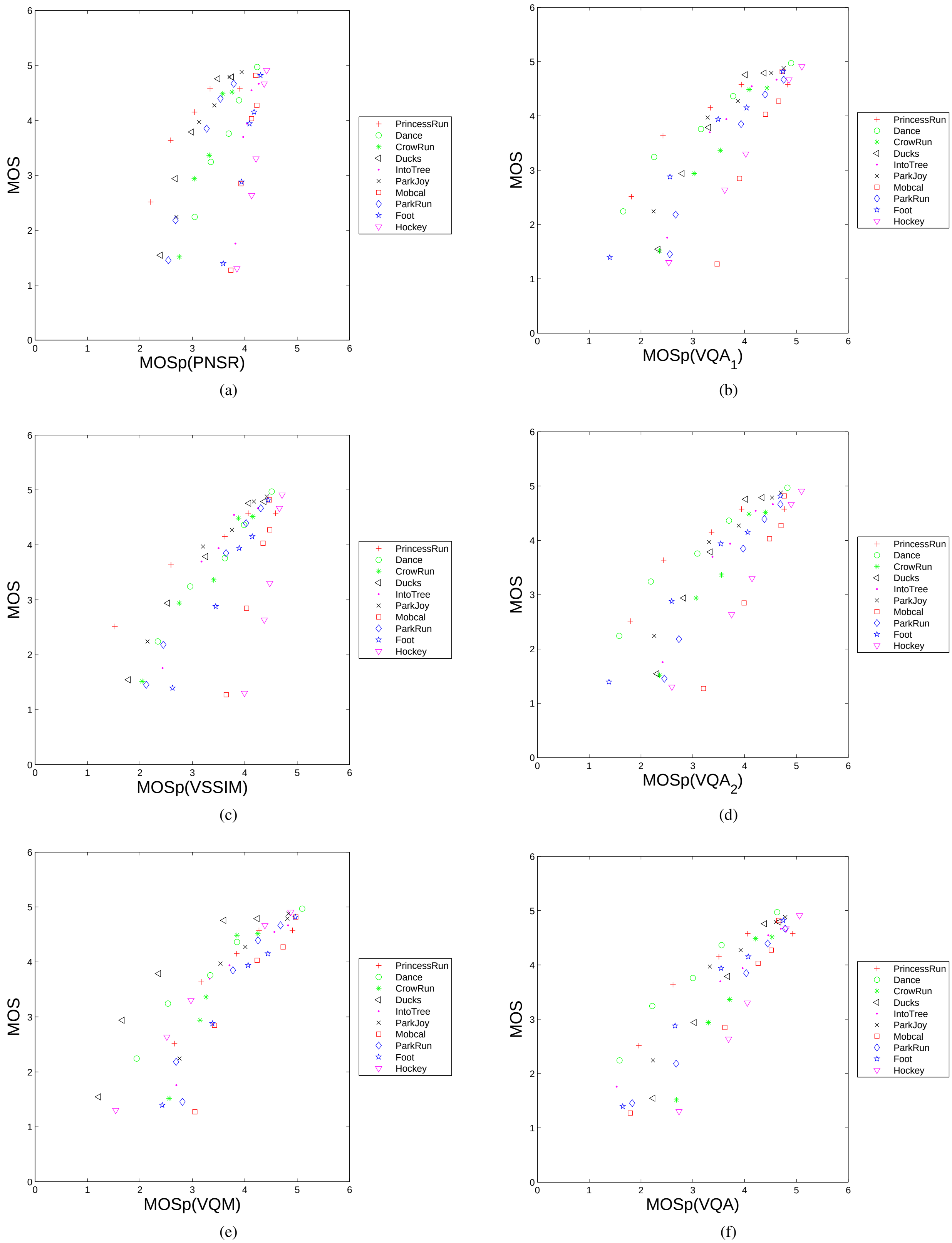

Fig. 9. Scatter plot comparison of different video quality assessment metrics on our video database. Vertical and horizontal axes are for subjective (MOS) and objective measurement (MOSp), respectively. Each sample point represents one test video sequence. The same marker type is used for each impaired video obtained from the same original video: (a) PSNR, (c) VSSIM, (e) VQM, (b) VQA$_1$, (d) VQA$_2$, and (f) VQA.

[8] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication, special issue on objective video quality metrics*, vol. 19, pp. 121–132, 2004.

[9] A. A. Stocker and E. P. Simoncelli, “Noise characteristics and prior expectations in human visual speed perception,” *Nature Neuroscience*, vol. 9, pp. 578–585, 2006.

[10] M. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *Broadcasting, IEEE Transactions on*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[11] ANSI T1.801.03 2003, “American national standard for telecommunications - digital transport of one-way video signals - parameters for objective performance assessment,” *American National Standard Institute*.

[12] IUT-T Recommendation J.144, “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” *Recommendations of the ITU, Telecommunication Standardization Sector*.

[13] IUT-R Recommendation BT.1683, “Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference,” *Recommendations of the ITU, Radiocommunication Sector*.

[14] K. T. Tan, M. Ghanbari, and D. E. Pearson, “An objective measurement tool for mpeg video quality,” *Signal Process.*, vol. 70, no. 3, pp. 279–294, 1998.

[15] M. Masry and S. Hemami, “A metric for continuous quality evaluation of compressed video with severe distortions,” vol. 19, no. 2, pp. 133–146, February 2004.

[16] J. E. Hoffman, “Visual attention and eye movements,” In H. Pashler, Ed. Hove, UK: Psychology Press, 1998, pp. 119–154.

[17] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, “On the performance of human visual system based image quality assessment metric using wavelet domain,” *Proc. SPIE Human Vision and Electronic Imaging XIII*, 2008.

[18] J. M. Wolfe, “Visual search,” In H. Pashler, Ed. East Sussex, UK: Psychology Press, 1998, pp. 13–74.

[19] S. Daly, “The visible differences predictor : an algorithm for the assessment of image fidelity,” *Proc. SPIE*, vol. 1666, pp. 2–15, 1992.

[20] J. Odobez and P. Bouthemy, “Robust multiresolution estimation of parametric motion models,” *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, 1995.

[21] O. Le Meur, P. Le Callet, and D. Barba, “Predicting visual fixations on video based on low-level visual features,” *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[22] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” March 2000, http://www.vqeg.org/.

[23] ITU-R Recommendation BT.500-10, “Methodology for the subjective assessment of the quality of television pictures,” *Recommendations of the ITU, Radiocommunication Sector*.

[24] Video Quality Experts Group (VQEG), “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii,” 2003, http://www.vqeg.org/.