Cover Page

Title of the paper

Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric.

Authors' affiliation and address and e-mail address

O. Le Meur, University of Rennes 1, olemeur@irisa.fr
A. Ninassi (was before at Technicolor R&D)
P. Le Callet, IRCCyN-IVC (UMR CNRS 6597), Polytech'Nantes, patrick.lecallet@univ-nantes.fr
D. Barba, IRCCyN-IVC (UMR CNRS 6597), Polytech'Nantes, Dominique.Barba@univ-nantes.fr

Journal & Publisher information

Elsevier Signal Processing: Image Communication

http://www.elsevier.com/wps/find/journaldescription.cws_home/505651/description#description

Bibtex entry

@article{LeMeur_2010_ImgComm,
Author={O. Le Meur, A. Ninassi, P. Le Callet and D. Barba},
Journal={Elsevier Signal Processing: Image Communication},
Title={ Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric.},
Year={2010}}

DOI = doi:10.1016/j.image.2010.05.006

Overt visual attention for free-viewing and quality assessment tasks
Impact of the regions of interest on a video quality metric.

O. Le Meur[a], A. Ninassi[b,c], P. Le Callet[c], D. Barba[c]

[a] *University of Rennes 1*
IRISA-TEMICS
35042 RENNES FRANCE
[b] *Technicolor R&D*
1 avenue Belle Fontaine
35576 CESSON SEVIGNE FRANCE
[c] *IRCCyN*
La Chantrerie
44306 NANTES FRANCE

Abstract

The aim of this study is to understand how people watch a video sequence during free-viewing and quality assessment tasks. To this end, two eye tracking experiments were carried out. The video dataset is composed of ten original video sequences and fifty impaired video sequences (5 levels of impairments obtained by a H.264 video compression). A first experiment consisted in recording eye movements in a free-viewing task. The ten original video sequences were used. The second experiment concerned an eye tracking experiment in a context of a subjective quality assessment. Eye movements were recorded while observers judged on the quality of the fifty impaired video sequences. The comparison between gaze allocations indicates the quality task has a moderate impact on the visual attention deployment. This impact increases with the presentation number of impaired video sequences. The locations of regions of interest remain highly similar after several presentations of the same video sequence, suggesting that eye movements are still driven by the low level visual features after several viewings. In addition, the level of distortion does not significantly alter the oculomotor behavior. Finally, we modified the pooling of an objective full-reference video quality metric by adjusting the weight applied on the distortions. This adjustment depends on the visual importance (the visual importance is deduced from the eye tracking experiment realized on the impaired video sequences). We observe that a saliency-based distortion pooling does not significantly improve the performances of the video quality metric.

Key words: Visual attention, video quality assessment, video quality metric, free-viewing and quality tasks.

1. Introduction

A great deal of interest and research has been devoted to the design and development of visual quality metrics, leading to the definition of three types of quality metric: no-reference, reduced-reference and full-reference video quality metric. A full-reference video quality metric requires

to have the original and the impaired video sequences. This could be a strong limitation in practice. To overcome this limitation, a reduced-reference quality metric can be used. It requires to dispose only of a more or less reduced description of the reference video. This description is compared to a similar description coming from the impaired video in order to compute a quality score. This can be more useful than a full-reference video quality metric for some applications, but still impossible to use for many others. The last solution is to use a no-reference quality metric for which only the impaired video sequence is available. Some of image and video metrics (IQM (Image Quality Metric) or VQM (Video Quality Metric)) rest on the use of human visual system properties. Hierarchical perceptual decomposition, contrast sensitivity functions, visual masking, etc are the common building blocks of a perceptual metric. These operations simulate different levels of human perception and are now well mastered.

However, assessing the quality of an image or video sequence is a complex process, involving the visual perception but also the visual attention closely linked to our prior knowledge. It is wrong to think that all areas of the picture or video sequence are accurately inspected during a quality assessment task. People preferentially and unconsciously focus on regions of interest. For these types of regions, our sensitivity to distortions might be significantly increased to the detriment of the other. Even though we are aware of this, very few IQM or VQM approaches take this property into account. To go one step further on this topic, we need to understand how people perceive the quality of a video and how they adapt their visual strategies to judge the quality of an image or video sequence. The reality today is that we still don't know precisely how people judge the quality of a video sequence. For continuous quality evaluation, we know that humans are quick to criticize and slow to forgive. This experimental property can be used to improve the pooling stage of a video quality metric [1, 2, 3].However, there is almost no study that has examined the visual strategy of an observer during the quality scoring of a video sequence. It is intuitively obvious that the areas of the video sequence do not have the same visual importance and the same capacity to attract our visual attention. The hypothesis is that an impairment appearing on a region of interest is probably more annoying than an impairment on a non visually interesting area. Is this intuition relevant and does the use of the visual importance of an area bring a significant improvement? Previous studies dealing with the quality assessment of still color pictures [4, 5] showed that the relationship between visual importance and the quality assessment is not as simple as one would expect.

The aim of this paper is to examine the visual attention deployment during both quality assessment and free-viewing tasks. More specifically, the goal is to determine whether a significant difference exists between these two experimental contexts. The context of this study concerns the full-reference video quality metric. This paper is organized as follows. Section II presents the eye-tracking and the quality assessment experiments. Section III examines both the impact of the quality assessment tasks and of the visual distortion on the visual attention deployment. We will try to answer the following question: is there any significant difference between eye movements which take place during a free-viewing and during a quality assessment task? Section IV presents a video quality metric in which the pooling of the spatio-temporal distortion gives more importance to degradations appearing on visually interesting areas. We conclude the paper in Section V.

2. Eye-tracking and quality assessment experiments

2.1. Stimuli

Ten unimpaired video sequences are used and are then degraded through a H.264/AVC video encoder. Five levels of degradation have been used. The five levels of degradation cover roughly the range of visual quality (the set of bitrate is then depending on the video content). The impairments caused by the encoding process are neither spatially nor temporally stationary. Some areas are impaired, whereas the quality of others remains almost unchanged. The video dataset is composed of the ten original video sequences and fifty impaired video sequences. The spatial resolution of each video sequence is 720×480 with a frame rate of 50Hz in a progressive scan mode. Each video clip lasts 8s. Figure 1 presents key pictures for the ten original video sequences.

2.2. Subjects

Thirty six unpaid subjects participated to the experiments (male and female). They came from the University of Nantes. All had normal or corrected to normal vision and all had normal color vision. All were inexperienced observers (not expert in video processing) and naive to the experiment.

2.3. Eye-tracking protocol

Eye-tracking experiments were conducted using a dual-Purkinje eye tracker from Cambridge Research Corporation. The eye tracker is mounted on a rigid EyeLock headrest that incorporates an infrared camera, an infrared mirror and two infrared illumination sources. To obtain accurate data regarding the diameter of the subjects' pupil, a calibration procedure is required. The calibration process consists in presenting the subject with a number of targets on the same screen from a known distance. Once the calibration procedure is completed and a stimulus has been loaded and displayed, the system is able to track the subject's eye movement. To maintain the data accuracy all along the test duration, the calibration procedure is repeated regularly during the test. The camera records a sequence of close-up images of the eye. This sequence is processed in real-time in order to extract the spatial locations of the position of the eye. Both Purkinje reflections are used to calculate the eye's location estimation. The guaranteed sampling frequency is 50Hz and the accuracy of the measurement is 0.5 degree of visual angle.

2.4. Experiments

Two types of experiments have been conducted. The first experiment is performed in a free-viewing task, meaning that observers are free to explore the field of the displayed video sequence. This situation where no explicit task is given to the observer is often used in order to lessen the contribution coming from the top-down or cognitive attention mechanisms. The goal is to give more importance to the low-level visual features. However, top-down influences cannot be ruled out and they can significantly contribute to the attentional allocation. The ten original video sequences are used in this first experiment. The viewing duration was of 8 seconds. Prior to the onset of each sequence, two flickering black discs sequentially appeared at two different positions during one second each. Then a gray picture was displayed for two seconds. There was no fixation marker prior the onset of the clip. The goal is to avoid any influence on fixation behavior coming from a particular area of the screen. The video clips are displayed at a viewing distance of four times the height of the displayed video (66 cm). Video sequences are positioned

in a random fashion around the center of the screen. The rationale of that relies on the willingness to be less sensitive to the bias center. Generally, when observers watch videos on monitors, they tend to look more frequently at the center of the screen than at its periphery. This central tendency has been noticed in different studies [6].

The second eye movement recording test is achieved during a video quality assessment campaign. It is expected that this high-level visual task will significantly impact the visual scanning of the scene. After the viewing of a video sequence, observers had to give a quality score. A specific graphical user interface was designed for that purpose. More explanations are given in section 2.6.

2.5. Human priority maps

From the collected fixation data, a human priority map [7] is computed for each observer and for each video sequence. It encodes the degree of interest of each spatial location of the video sequence. To compute this kind of map, the raw eye tracking data is first parsed in order to separate data into fixation and saccade periods. Each sample coming from the eye tracking apparatus is treated according to the following algorithm:

1. Calculate point-to-point velocity for each sample;
2. Label each sample below a given velocity threshold (25 degree/second) as belonging to a potential visual fixation period, otherwise as to a saccade period;
3. Merge consecutive potential visual fixation samples into a fixation group, removing saccade samples. The length of these groups, or in other words the fixation duration must be higher than 100 ms. Under this threshold, the samples belonging to either a saccade or a short fixation, are discarded;
4. Compute the spatial coordinates of each visual fixation (as the gravity center of the coordinates of the samples in the considered group).

More precisely, the parsing of the raw eye tracking data determines the fixation sequence, called $SM^{(k)}$ (for observer k), given by:

$$SM^{(k)}(x, y, t) = \sum_{i=1}^{M_k} \sum_{d=start_i}^{end_i} \Delta(x - x_i, y - y_i, t - t_d) \tag{1}$$

where M_k is the number of visual fixations; $start_i$ and end_i represent the start and the end of the visual fixation i, respectively. Δ is the Kronecker symbol.
Sequences $SM^{(k)}$ are grouped together, leading to an average fixation sequence SM. SM could be interpreted as a map indicating where an average observer (or standard observer) would look at:

$$SM(x, y, t) = \frac{1}{N} \sum_{k=1}^{N} SM^{(k)}(x, y, t) \tag{2}$$

where N is the number of observers. This sequence is eventually smoothed with a 2D Gaussian filter, leading to the human priority map. The rational of the Gaussian filtering is two-fold. Observers do not gaze at a specific point of the visual field but rather at an area having a size close to the size of the fovea (in visual angle).

The second reason is related to the eye tracking apparatus (accuracy of the apparatus is about 0.5 degree of visual angle). To simulate this, the standard deviation of the Gaussian filter is set

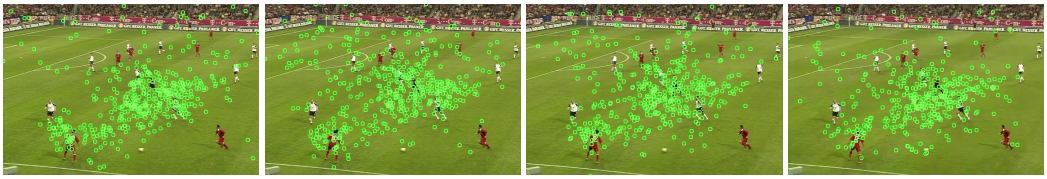
(a) Princess Run (b) Dance (c) Crowd Run

(d) Ducks (e) Intotree (f) ParkJoy

(g) Mobcal (h) ParkRun (i) Foot

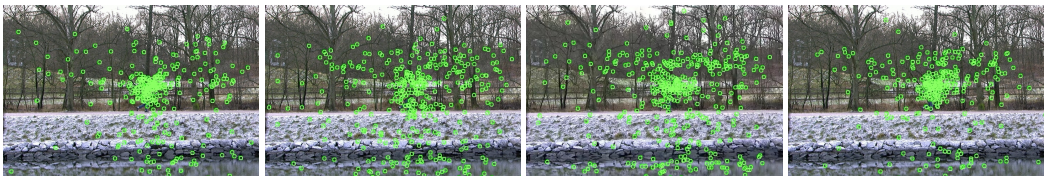
(j) Hockey

Figure 1: Representative pictures of the 10 original video sequences.



(a) Foot (b) $MOS = 4.81 \pm 0.38$ (c) $MOS = 3.93 \pm 0.95$ (d) $MOS = 1.39 \pm 0.64$

(a) Dance (b) $MOS = 4.97 \pm 0.17$ (c) $MOS = 3.75 \pm 0.85$ (d) $MOS = 2.24 \pm 0.74$



(a) ParkRun (b) $MOS = 4.66 \pm 0.58$ (c) $MOS = 3.84 \pm 0.74$ (d) $MOS = 1.45 \pm 0.49$

(a) CrowdRun (b) $MOS = 4.51 \pm 0.60$ (c) $MOS = 3.36 \pm 0.97$ (d) $MOS = 1.51 \pm 0.60$

Figure 2: All visual fixations obtained on the whole sequence have been accumulated on key picture of the corresponding video sequence. The green circles correspond to the fixation points (notice that the radius of the circle has been arbitrary chosen). First row: Foot. Second row: Dance. Third row: ParkRun. Fourth row: CrowdRun. From left-hand side to right: original video sequence with visual fixations stemming from the free-viewing experiment (a). The fixation points superimposed on the three last pictures ((b), (c) and (d)) of a row are fixation points obtained during the quality task assessment with three different levels of distortion ($q1$ (b), $q3$ (c) and $q5$ (d)).

to 0.75 degree of visual angle.
Figure 2 shows all the fixation points recorded during the eye tracking experiments superimposed on a the representative picture. The spatial distribution of the fixation points is given for four original video sequences and for three levels of degradations ($q1$ (lowest compression), $q3$ and $q5$ (highest compression)). Notice that the spatial distribution of the fixation points seems to be qualitatively similar whatever the level of distortion may be. It suggests that the level of distortion would not significantly influence the deployment of the visual attention. These fixation points were obtained from the quality assessment task results.

2.6. Quality assessment protocol

During the video quality assessment, the eye movements of the subjects were recorded. The fact that a particular task is assigned may likely alter the oculomotor behavior (compared to an experiment of free-viewing task).
The standardized method DSIS (*Double Stimulus Impairment Scale*) was used to access the quality of the video sequence. In DSIS, each observer views an unimpaired reference video sequence followed by one of its impaired version. Observer then rates the visual quality of the impaired video sequence using a 5-score scale. We use an impairment scale:

1. very annoying: observer is very annoyed by the impairments;
2. annoying: observer is annoyed by the impairments;
3. slightly annoying: observer is slightly annoyed by the impairments;
4. not annoying: observer is not annoyed by the visible impairments;
5. imperceptible: impairments are imperceptible to the observer.

The notation is performed by the observer by focusing his visual attention on a particular target related to the quality assessment he wanted to score. The observer just gazes at the scoring screen area (see figure 3) corresponding to his choice. The chosen area becomes red, and the observer then validates or rectifies his choice by directing his gaze to one of the two screen areas called *confirm* or *reset*, respectively. Figure 3 illustrates the principle of the proposed scoring method. The rationale of this method is simple. Through the use of this visual scoring interface, the head of the observer remains fixed over the experiment. Therefore, it is not required to re-calibrate the apparatus after each quality assessment scoring.
The mean opinion score (MOS) as well as its standard deviation (STD) are given in Table 1 for each sequence and for each level of distortion.

3. Impact of the task and of the distortion on eye movements

Eye movements of a panel of observers were recorded in two different contexts, namely during a free-viewing task (FT) and during a quality task (QT). Still again, the protocol DSIS was used to assess the quality of the impaired video sequences. A reference video sequence is always displayed before the test video sequence.
Therefore, for the latter context, two human priority maps were available: one computed from the collected data on the reference video sequence, the second one from eye movements recorded when observers watched the impaired video sequence.
The differences in the visual strategy are examined, taking into account the type of the task and also the strength of distortion. The first analysis concerns the fixation durations. In a second step of analysis, the degree of similarity between human priority maps is computed.

Table 1: MOS obtained during the quality campaign. The mean quality score is given per video sequence and per level of degradation.

	$MOS \pm STD$ $q1$: lowest compression – $q5$: highest compression				
Sequence name	q1	q2	q3	q4	q5
Princess Run	4.57 ± 0.65	4.57 ± 0.65	4.15 ± 0.74	3.63 ± 0.80	2.51 ± 0.95
Dance	4.97 ± 0.17	4.36 ± 0.64	3.75 ± 0.85	3.24 ± 0.81	2.24 ± 0.74
Crowd Run	4.51 ± 0.60	4.48 ± 0.60	3.36 ± 0.97	2.93 ± 0.91	1.51 ± 0.60
Ducks	4.78 ± 0.40	4.75 ± 0.55	3.78 ± 0.80	2.93 ± 0.85	1.54 ± 0.55
Intotree	4.66 ± 0.47	4.54 ± 0.55	3.93 ± 0.64	3.69 ± 0.57	1.75 ± 0.81
ParkJoy	4.87 ± 0.32	4.78 ± 0.47	4.27 ± 0.82	3.96 ± 0.83	2.24 ± 0.85
Mobcal	4.27 ± 0.78	4.81 ± 0.45	4.03 ± 0.71	2.84 ± 0.78	1.27 ± 0.44
ParkRun	4.66 ± 0.58	4.39 ± 0.64	3.84 ± 0.74	2.18 ± 0.83	1.45 ± 0.49
Foot	4.81 ± 0.38	4.15 ± 0.65	3.93 ± 0.95	2.87 ± 0.80	1.39 ± 0.64
Hockey	4.90 ± 0.28	4.66 ± 0.58	3.30 ± 0.83	2.63 ± 0.81	1.30 ± 0.45

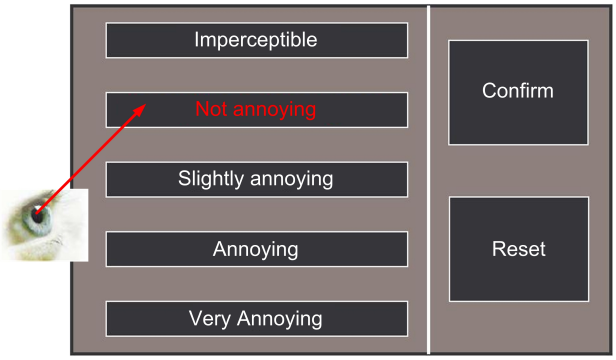


Figure 3: Scoring visual interface. The observer's eye is used to select the quality score (impairment scale).

3.1. Visual fixation duration

The first analysis aims at investigating the impact of the task on the duration of the visual fixations. These durations are computed for three conditions: on the reference video sequence in free-viewing task, on the reference and on the impaired video sequences in quality task.
Results shown in figure 4 indicate that the average duration of the visual fixations remains almost the same for the three conditions. The average duration spread between 400 and 550 *ms*. In video, the quality task does not alter in a significant manner the oculomotor behavior. It is not consistent with a previous finding [8] obtained on still color pictures. In this study, there was a significant difference between the duration of the visual fixations from a free-viewing task and from a quality task. This difference between the conclusions of these two studies is probably due

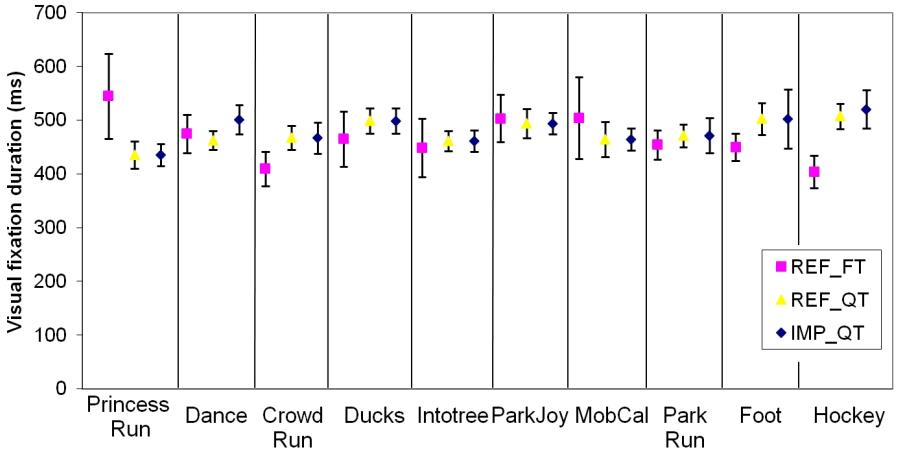


Figure 4: Comparison of the duration of the visual fixations for three conditions: during a free-viewing task, during a quality task on the video reference and on the impaired video sequence, respectively. The average value is given as well as the confidence interval.

to the fact that we do not watch a video clip as we watch a still color picture. For a still color picture, after few seconds of viewing, it is well accepted that top-down mechanisms can override the bottom-up mechanism and can influence the gaze allocation [9]. Another explanation can rely on the fact that observers might try to memorize some parts of the picture in order to compare them to the impaired one. Such visual strategy would provoke an increase of the visual fixation durations. Considering the viewing of video sequence, it is reasonable (but not demonstrated yet) to suppose that visual attention is less dependent on the top-down contributions. Indeed, the temporal dimension of a video sequence might give much more importance to the low-level visual features than for a still color picture, just as observers have to follow the action in the scene. It is well known that the motion contrast closely linked to the movement in the scene is one of the strongest attractors of our gaze [10, 11]. Therefore, it is probably more difficult for an observer to gaze at a particular area in a dynamic context in order to memorize it than on a still picture.
What is certain is that different visual strategies due to different visual tasks can produce major

differences in the visual attention deployment. The seminal work of A. Yarbus [12] is the perfect example.

3.2. Similarity of human priority maps

The degree of similarity between the human priority maps is computed in two conditions, as illustrated in figure 5. The two comparisons are described below:

- Comparison $REF(QT)vsREF(FT)$: the comparison involves the human priority maps deduced from the reference video sequences in the both experimental conditions. The goal here is to analyze the impact of a quality assessment task on the oculomotor behavior when observers look at reference video sequences;

- Comparison $IMP(QT)vsREF(FT)$: the comparison is performed between the human priority maps obtained in free-viewing task and obtained in quality task, respectively. The difference with the first comparison relies on the fact that the impaired video sequence is used instead of the reference. The goal is to examine the extent to which the impairments and the task could alter the deployment of visual attention. One may argue that the comparison $IMP(QT)vsREF(QT)$ (comparison between human priority maps obtained in quality-task) is more interesting than $IMP(QT)vsREF(FT)$ since only one parameter would be changed. However, it is important to remind the context of this study. The question we want to answer is: can we use a purely bottom-up computation model of visual attention in order to enhance the prediction of quality scores? The parameter FT, standing for Free-viewing Task, is then fundamental in the context of this study. In addition, in previous work we found that the visual attention deployment was not influenced by the level of impairments [13]. Therefore, as the visual attention is invariant to impairments (the aforementioned study was conducted with impairments coherent with a typical TV broadcast), we made the assumption that comparisons $IMP(QT)vsREF(FT)$ and $IMP(QT)vsIMP(FT)$ would lead to the same result.

3.2.1. ROC Analysis

The degree of similarity between the human priority maps has been computed through a ROC (Receiver Operating Characteristic) analysis [14]. It consists in estimating the true positive rate (TPR) and the false positive rate (FPR), by labeling each pixel of the human priority maps as *fixated* or *not fixated*. A first map, considered as the reference, is first labeled in two classes (1 for *fixated* and 0 for *non fixated* areas) by using a unique decisional system (DS). The second map, which is the predicted map, is also labeled into two classes. Contrary to the reference map, different thresholds are applied to predict the labels. The problem can be expressed as follow:

$$SM^{Ref}(x,y,t) = \begin{cases} 1 & \text{if } DS(SM(Ref_{FT}(x,y,t))) = fixated \\ 0 & otherwise(non\ fixated) \end{cases} \quad (3)$$

$$SM^{Test}(x,y,t) = \begin{cases} 1 & \text{if } SM(Ref_{QT} \text{ or } Imp_{QT})(x,y,t)) \geq T_i^{Test} \\ 0 & otherwise(non\ fixated) \end{cases} \quad (4)$$

SM^{Ref} and SM^{Test} are the binary maps of the reference and the predicted human priority maps, respectively. The dynamic range of the human priority maps is coded on 8 bits. DS is the decisional system used to label the human priority maps (from the eye tracking experiment involving the reference video sequence in a free-viewing task). The decisional system is a simple

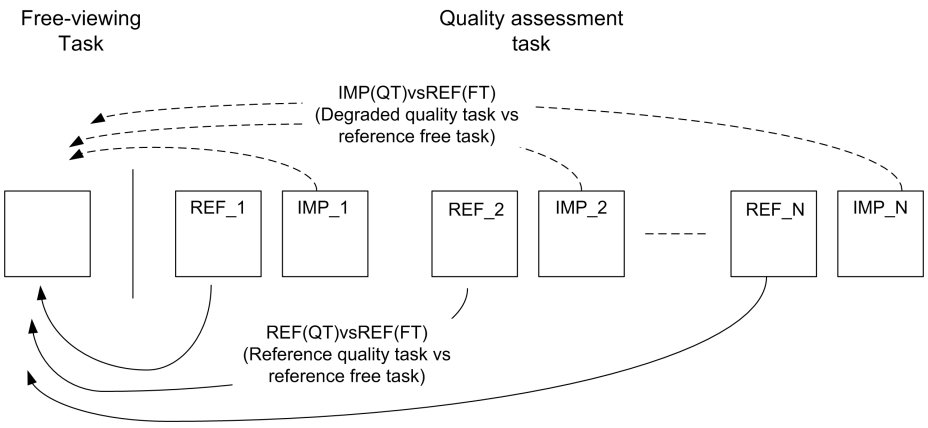


Figure 5: Two comparisons of priority maps are performed. (1) the human priority maps obtained in quality task (on the reference video sequence) and in free-viewing task (on the reference video sequence). (2) the human priority maps obtained in quality task (on the impaired video sequence) and in free-viewing task (on the reference video sequence).

threshold operation. The threshold equals to 14 (it indicates that an area will be set to one only if there are at least two observers out of 36 that focus on it at the same time $\frac{255}{36} \times 2 = 14$). $\left\{T_i^{Test}\right\}$ are the set of thresholds used to classify the predictions. As the dynamic range of the human priority maps is coded on 8 bits, we can have up to 256 thresholds going from 0 to 255.
From these classifications, a two-by-two confusion matrix is computed. Many features can be extracted from this matrix. Among them, the true positive rate (TPR) and the false positive rate (FPR) are given by:

$$\begin{aligned} TPR &\approx \frac{\text{Number of Positives correctly classified}}{\text{Total number of positives}} \\ FPR &\approx \frac{\text{Number of Negatives incorrectly classified}}{\text{Total number of negatives}} \end{aligned} \tag{5}$$

A pair of values (TPR,FPR) is obtained for each threshold T_i^{Test} used. A ROC graph depicting the tradeoff between true positive and false positive rates is plotted. The TPR rate is plotted on the Y axis whereas the FPR rate is plotted on the X axis. On this graph, the point (0,1) represents a perfect similarity. The closer the curve the top left-hand corner, the better the classification is. The diagonal line (if it is plotted on a linear-linear graph) indicates that the classification is a pure random process. One interesting indicator is the area under the curve, called AUC. This indicator is indeed very useful to compare the quality of the prediction. The AUC value lies between 0 and 1. An AUC value of 0.5 means that there is no similarity between the two sets of data. A value of 1 is obtained for a perfect classification.
To assess the degree of similarity between the human priority maps, AUC is computed for each picture. All the AUC values are then averaged over the video sequence duration.

3.2.2. Impact of the task on the unimpaired video

The first analysis concerns the quality-task impact on the oculomotor behavior when observers watch the reference video sequences (comparison called $REF(QT)vsREF(FT)$). The ROC graph is plotted in figure 6 for three video sequences: *Ducks*, *Foot* and *CrowdRun*.

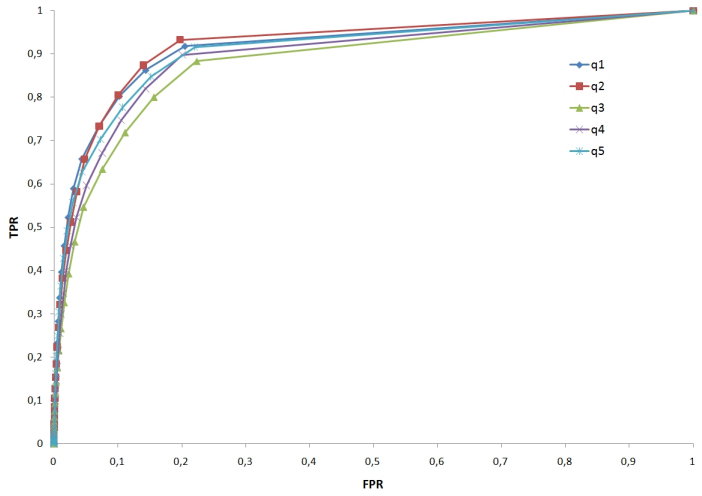
Figures 6 (a) and (b) suggest that there is no major difference between the human priority maps coming from a free-viewing task and a quality task. ROC graph for the last video sequence, called *CrowdRun*, has a global shape significantly different from the two others. We thought that it is not due to the task but rather to the video content itself. Indeed, the two first video sequences contain objects of interest that clearly come into view from their neighborhood (see the key pictures of these sequences in figure 1). Our visual attention is then naturally attracted by these regions, and therefore the variability inter-observers is small. This is not the case for the last sequence for which there is no region that pops out (see figure 2). The variability between the observers is important and could explain the difference in the shape of the ROC curve. Table 2 gives the AUC values for the ten reference video sequences. Three values are computed per sequence. The first is the average value computed over the whole sequence duration. The second corresponds to an averaging limited to the first second of viewing. Finally, the third is the average of the AUC values obtained between the first and the fourth second of viewing. The goal of these three indicators is to test whether there is a modification in the visual strategy at the beginning of the viewing. The general tendency of the results indicates that there is a good agreement between the priority maps. The fact that the tasks are not the same does not seem to impact the visual attention deployment. For instance, the smallest AUC value for a whole sequence equals 0.8. The lowest AUC value is obtained for the sequence *CrowdRun*, and that for the same reasons given previously. We can notice that the highest AUC values are obtained just after the stimulus onset (first second of viewing). It is coherent with previous studies [9] showing that the variability between observers is less important at the beginning of the viewing than after few seconds of viewing. Just after the stimulus onset, the role of the low-level visual features or in other words the bottom-up visual attention is more important than the top-down attentional effects. Between the first and the fourth second of viewing, the similarities between priority maps are a bit less important but still remain high. Overall, these results indicate that eye movements are mostly stimulus-driven not only under a free-viewing task but also under a quality assessment-task when we consider an unimpaired video sequence.

3.2.3. Impact of the task on the impaired video sequences

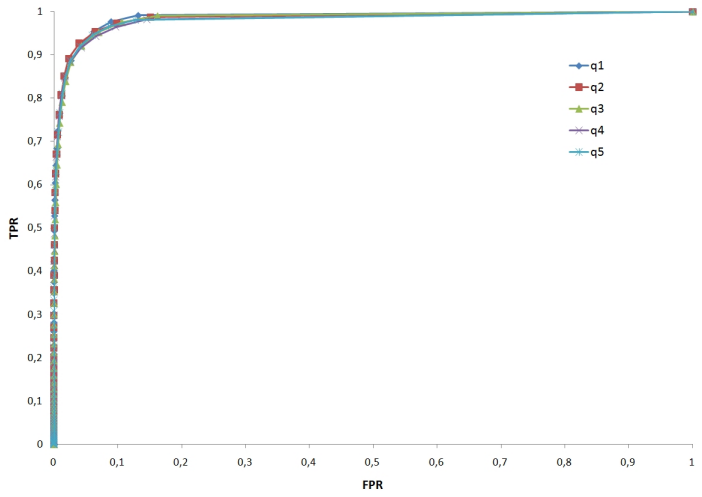
The impact of the quality task on the eye movements is examined in this section. This comparison is called $IMP(QT)vsREF(FT)$.

As illustrated in figure 7, the quality task significantly affects the similarity of the human priority maps for 6 sequences (*Princess Run*, *Ducks*, *Intotree*, *ParkRun*, *Foot* and *Hockey*). For these 6 sequences, the level of quality does not contribute to the modification of the similarity. For the four other sequences, there is at least one quality level for which the average value AUC is similar to the one obtained in a free-viewing task.

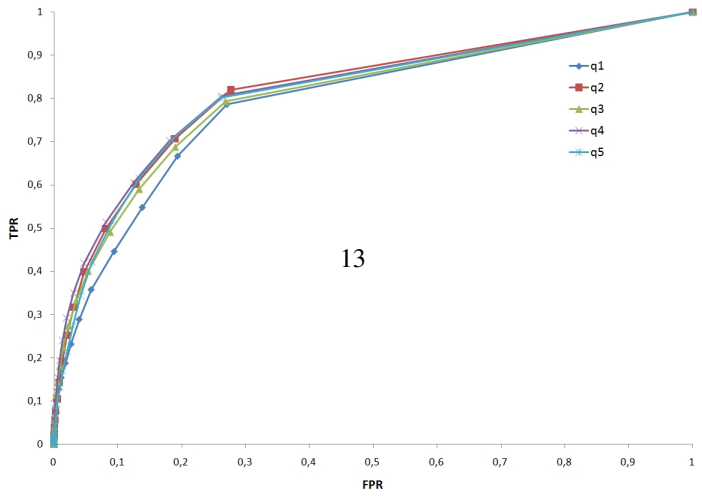
Figure 7 gives the average AUC value for each level of degradation. It is interesting to note from figure 7 that the degree of similarity between human priority maps is not systematically dependent on the level of quality (for the set of impairment level used in the experiment). It was reasonable to think that when the amount of degradation increases, the similarity decreases. Indeed, a poor video quality could significantly alter the visual deployment leading to a decrease of the degree of similarity between the priority maps. This type of configuration is only observed in the sequence *Dance*. In this case, the AUC values decrease with the level of quality: the visual



(a) Ducks



(b) Foot



(c) CrowdRun

Table 2: Comparison of the human priority maps during free-viewing and quality tasks (on the reference sequence presented just before the impaired video). The AUC value is computed on the whole duration of the sequence, on the first second duration of viewing and on the duration between the first and the fourth second of viewing.

	$AUC \pm STD$		
Sequence name	Whole duration	First second	1s to 4s of viewing
Princess Run	0.95 ± 0.08	0.95 ± 0.04	0.93 ± 0.11
Dance	0.86 ± 0.09	0.95 ± 0.06	0.84 ± 0.09
Crowd Run	0.80 ± 0.12	0.94 ± 0.05	0.75 ± 0.12
Ducks	0.90 ± 0.07	0.97 ± 0.03	0.88 ± 0.07
Intotree	0.91 ± 0.06	0.96 ± 0.03	0.87 ± 0.06
ParkJoy	0.96 ± 0.04	0.94 ± 0.07	0.96 ± 0.02
Mobcal	0.86 ± 0.11	0.91 ± 0.08	0.85 ± 0.10
ParkRun	0.96 ± 0.03	0.98 ± 0.02	0.96 ± 0.03
Foot	0.98 ± 0.02	0.96 ± 0.04	0.97 ± 0.02
Hockey	0.97 ± 0.02	0.95 ± 0.04	0.97 ± 0.02

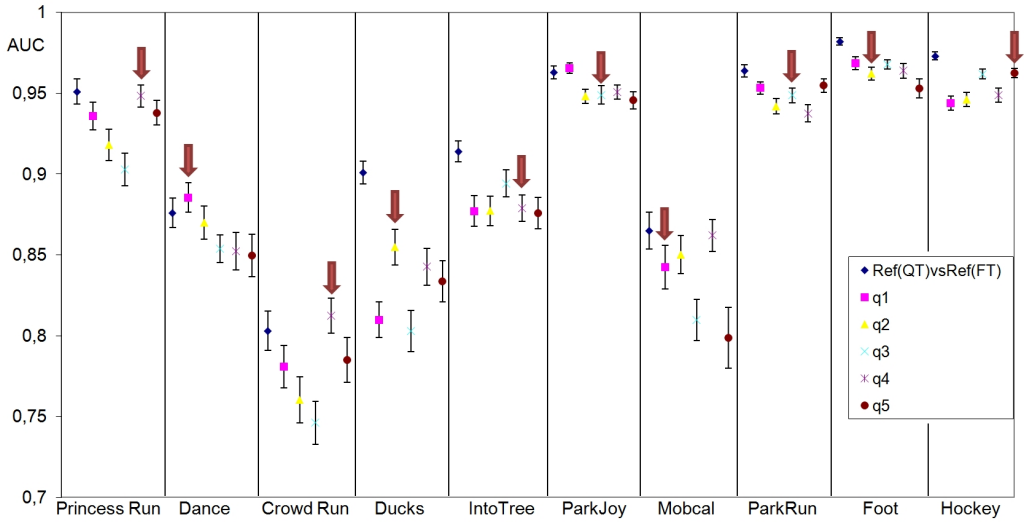


Figure 7: Comparison of the human priority maps during a free-viewing and quality tasks. The AUC values are computed over the whole sequences. The terms q1 to q5 represent the level of quality of the impaired video sequence (1 is for the best quality). On the graph, the previous result (see Table 2) has been added in order to compare the AUC values for both tasks. The upper and lower limits are the 95% confidence interval. The vertical arrow indicates the first impaired video sequence presented to the observers.

distortion could be the factor that influences the way observers watch the video. However, for the other sequences, this is not the case. For instance, for the sequence *Princess Run*, the two highest AUC values are obtained for the distortion levels $q4$ and $q5$. For these two levels of impairment, the MOSs equal to 3.63 and 2.5, respectively, indicating that the degradations are slightly annoying or really annoying.
To go one step further, the linear correlation coefficient[1] is computed between the MOSs and the average AUC values, given in Table 3. Results indicate the quality level of the sequences and the similarity between the priority maps issued from a free-viewing and a quality task are not correlated. Nevertheless, there are two particular cases: sequences *Dance* and *Foot*. For these two video sequences, the linear correlation coefficient is rather high. Notice that the correlation coefficient is also high for sequence *Hockey* but negative. It indicates that the more the sequence is impaired the more the similarity between priority maps increases.
These two first analyses provided contrasted results and it is not possible to give a clear and definitive trend. The degree of similarity varies with the level of distortion but the decrease of the similarity is not systematically due to an increase of distortion.
Another test was performed. The goal is to examine whether the order of viewing of the impaired video sequences has an influence on the similarity of the priority maps. For instance, we can intuitively suppose that there may be a learning effect becoming more and more important with the presentation number. Based on this hypothesis, the similarity of the priority maps should decrease with the rank of the presentation. The linear correlation coefficient is computed between the rank of the presentation and the average AUC values. Results are also given in Table 3 (second column). Most of the linear correlation coefficients are negative, in spite of the various content of the tested video sequences. It would indicate that the similarity between the scan paths decreases with the presentation number. It suggests there is a learning effect (in the condition of our experiment) but very limited. The level of expertise of the subjects does not dramatically increase after several viewings (reasonable number) of the same video sequence. The difference between the highest and the lowest AUC value is indeed given in Table 3 (third column). The maximum value does not exceed 0.1. In Table 3, the highest correlation is obtained by the sequence *Princess Run*. The presentation order was the following: $IMP_{q4}, IMP_{q5}, IMP_{q1}, IMP_{q2}, IMP_{q3}$. The first impaired video sequence viewed by subjects was the video sequence having a quality level $q4$, followed by the video sequence having a quality level $q5$, etc. The similarity decreases with this order. For this sequence, this result indicates that subjects have continuously modified theirs visual strategies, going relatively farthest from a free-viewing strategy. The vertical arrow in Figure 7 indicates the first impaired video sequence presented to the observers. For this sequence, the AUC value is very close to the AUC obtained in a free-viewing task (except for the sequence Ducks). It would indicate that the quality task does not significantly modify the visual scan path of the observers.

3.2.4. Brief summary

Two comparisons were performed in order to test the following assumption: does it make sense to use a purely bottom-up saliency map to improve the prediction of visual quality assessment ? The two following paragraphs summarize the conclusions of these comparisons:

- On the degree of similarity between the human priority maps obtained on the reference video sequences (unimpaired) in a quality assessment task (REF(QT)) and in a free-viewing

[1] We would like to draw the attention of the readers on the fact that only six points are used to compute the linear coefficients. As this coefficient is very sensitive to outliers, the interpretation is difficult.

Table 3: Linear correlation coefficients between the MOS and the average value AUC, between the rank of the presentation and the average value AUC. The rang of AUC is given in the last column.

Sequence name	$CC(MOS, AUC)$	$CC(RANK, AUC)$	$MAX(AUC) - MIN(AUC)$
Princess Run	−0.429	−0.878	0.045
Dance	0.896	−0.444	0.036
Crowd Run	−0.351	−0.481	0.066
Ducks	−0.137	−0.591	0.1
Intotree	0.172	−0.429	0.039
ParkJoy	0.573	−0.430	0.018
Mobcal	0.487	−0.450	0.067
ParkRun	0.000	−0.511	0.034
Foot	0.879	0.105	0.03
Hockey	−0.758	−0.263	0.03

task (REF(FT)), respectively: in our experimental conditions, a quality task does not significantly influence the attentional allocation. The gaze deployment is mostly driven by the low-level visual features;

- On the degree of similarity between the human priority maps obtained on impaired video sequences in a quality assessment task (IMP(QT)) and on the reference video sequences in a free-viewing task (REF(FT)), respectively: first, our results indicate that the eye movements are not significantly influenced by the level of impairment. Regarding the quality task, the results are more diversified. Results would indicate that the quality task does not influence the way we inspect the video. This is especially true for the first viewing. When the number of presentation increases, the similarity between scan paths decreases but this decrease is rather limited. It indicates that there is a memory or learning effect. Observers tend to adjust their visual behavior in function of the previous viewing.

In conclusion, the use of a purely bottom-up computational model of visual attention might be a good solution to improve the efficiency of a video quality metric, especially if the number of presentation of the same video sequence is small. Indeed, in a quality task assessment, the similarity between scan paths is strong but decreases with the number of viewing.

4. Objective quality metric using a saliency-based pooling

In this section, we present an adaptation of an objective full-reference video quality metric we previously designed. The modification consists in taking into account the visual importance of the video sequence areas. The objective is to test the validity of the previous conclusions by using a simple modification of the distortion pooling stage.

4.1. Wavelet-based quality metric

The video quality metric proposed in [15] is used here to test whether it is possible to improve the prediction of the quality scores by using a saliency-based pooling.
For the spatial dimension, this video quality metric is based on the use of a wavelet transform,

a contrast sensitivity function and a visual masking function. To take into account the temporal dimension, a short-term pooling, based on spatio-temporal tubes, is proposed. This is followed by a long-term temporal pooling. The spatio-temporal tubes are used to track the fluctuation of the spatial distortions over the duration of visual fixations (400 *ms* on average). The goal is to examine the time frequency and the magnitude of the spatial distortion fluctuations present in the tube in order to strengthen or to lessen the potential annoyance due to these distortions. A spatial pooling of the spatio-temporal distortion is performed in the final step.
In the next section, we propose to modify this spatial pooling part in order to give more importance to the visually interesting regions.

4.2. Modification of the spatial pooling

The pooling of spatio-temporal distortions is modified by guessing that certain areas of an image may be visually more important than others. The modified pooling function is given by:

$$D_t^S = \left(\frac{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(x,y,t) \cdot \left(\overline{\overline{VE}}(x,y,t) \right)^{\beta_s}}{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(x,y,t)} \right)^{\frac{1}{\beta_s}}, \tag{6}$$

Where D_t^S is the perceptual distortion value for the frame at time t weighted by the visual saliency. K and L are the height and the width of the picture, respectively. $w_i(x,y,t)$ is the weighting factor i applied at pixel (x,y) of the frame at time t. $\overline{\overline{VE}}(x,y,t)$ is the spatio-temporal map of the visual distortion at t. For more details readers could refer to [15]. Two β_s values were tested: 1 and 2. A higher β_s value will favor the strongest distortions in the map to the detriment of others. Seven different weighting functions w_i are given by:

$$\left| \begin{array}{l} w_0(x,y,t) = 1 \\ w_1(x,y,t) = SM_n(x,y,t) \\ w_2(x,y,t) = 1 + SM_n(x,y,t) \\ w_3(x,y,t) = SM(x,y,t) \\ w_4(x,y,t) = 1 + SM(x,y,t) \\ w_5(x,y,t) = SM_b(x,y,t) \\ w_6(x,y,t) = 1 + SM_b(x,y,t) \end{array} \right. \tag{7}$$

where $SM(x,y,t)$ is the unnormalized human saliency map, $SM_n(x,y)$ is the human saliency map normalized in the range $[0,1]$ and $SM_b(x,y,t)$ is a binarized human saliency map (a threshold equals to 14 is used for the reason explained in section 3.2.1). Recall that the saliency maps stem from the eye tracking experiment performed on the impaired video sequence in quality task. The final distortion value D, pooled over the sequence, is obtained by the formula (7) given in [15]:

$$D = \begin{cases} \overline{D} + \Delta_D & , \Delta_D < \lambda \overline{D} \\ \overline{D} + \lambda \times \overline{D} & \text{Otherwise} \end{cases} \tag{8}$$

where, $\overline{D}$ is time average distortion, Δ_D represents the variation of distortion along the sequence and λ is a weighting factor.
The weighting functions w_1, w_3 and w_5 give more importance to the salient areas than the others. Indeed, the offset value of 1 in the weighting functions w_2, w_4 and w_6 allows us to take into account distortions appearing also on the non salient areas. w_0 is the unmodified version of the video quality metric.

4.3. Results

A psychometric function is used to transform the perceptual distortion D into a global quality score $MOSp$, as recommended by the Video Quality Expert Group (VQEG) [16]. This is given by:

$$\mathrm{MOSp} = \frac{b_1}{1 + e^{-b_2 \cdot (D - b_3)}} \tag{9}$$

The three parameters of the function have been optimized.

The impact of each weighting function was evaluated using the linear correlation coefficient (CC), the Spearman rank ordered correlation coefficient (SROCC) and the Root Means Squared Error (RMSE) between the MOS and its prediction $MOSp$. These results are compared to a traditional approach where the visual importance of the areas is not taken into account. All the results are given in Table 4. Whatever the weighting functions used, there is no significant performance improvement. The best results are obtained with a constant weighting w_0, meaning that all the areas of the video sequences are considered as having the same visual importance. These results suggest that a simple saliency-based pooling function is not a good solution to improve the visual quality prediction.

The same results were observed for still color pictures [4]. These results were also confirmed by a recent study [5]. Indeed they did not find a statistically significant improvement of different quality metrics weighted by the visual importance (PSNR, SSIM [17], VIF [18] and VSNR [19]). A similar methodology was used in [20] leading to the same conclusion. In this study, eye tracking experiments were conducted and human visual fixations were then used to weight a distortion map.

Weighting			Metrics		
Saliency	w_i	β_s	CC	SROCC	RMSE
IMP(QT)	w_0	***1***	***0.889***	***0.904***	***0.526***
	w_1	1	0.875	0.903	0.554
	w_2	1	0.889	0.904	0.525
	w_3	1	0.875	0.903	0.554
	w_4	1	0.883	0.908	0.538
	w_5	1	0.876	0.904	0.553
	w_6	1	0.89	0.906	0.524
IMP(QT)	w_0	***2***	***0.892***	***0.9***	***0.519***
	w_1	2	0.878	0.904	0.548
	w_2	2	0.892	0.901	0.519
	w_3	2	0.878	0.904	0.548
	w_4	2	0.886	0.912	0.532
	w_5	2	0.88	0.905	0.546
	w_6	2	0.893	0.902	0.517

Table 4: Impact of the human saliency on the performances of a video quality metric. Different weighting functions are used.

5. Conclusion

In this study, the comparison between the visual strategies used in a free-viewing task and a quality assessment task are examined. The goal was to compare the eye movements recorded

during the two different tasks. More accurately, the question was to determine whether a high-level visual task, such as the video quality assessment, implies a dedicated oculomotor behavior. The comparison between eye movements collected during the two tasks on ten video sequences (comparison called $REF(QT)vsREF(FT)$) indicates that the degree of similarity between human priority maps is rather high. The quality task when the unimpaired video sequences are considered does not seem to modify in a significant manner the oculomotor behavior. Therefore, the low-level visual features play an important role but the extent to which they contribute to the quality judgment is still an open-issue.

The second comparison called $IMP(QT)vsREF(FT)$ involves impaired video sequences in the context of a quality task. Results indicates that the gaze allocation is not disturbed by the level of distortion (compared to a free-viewing task). The difference observed would be due to either the visual task or the rank of the presentation. Concerning the latter point, there is evidence in our data (see figure 7) for a memory or a learning effect. For most of the sequences, the similarity between human priority sequences is the highest for the first presentation, whatever the level of impairment. It would mean that observers tend to adapt their own visual strategy throughout the experiment. It should be noted however that the similarity still remains high even after 5 presentations.

The previous conclusion suggests that it would be possible to use a bottom-up computational model of visual attention in order to predict the quality scores. The relevance would be high for the first viewing and would decrease with the number of presentation. In addition, nothing allows us to assert that there is an obvious relationship between quality score and regions of interest obtained in a free-viewing task. The relationship between quality scores and region of interest is much more complicated than one would expect it to be. The hypothesis postulating that impairments contribute more to the elaboration of the quality score when it occurs on a region of interest is likely true, but it is only a particular case. The generalization of this hypothesis will be difficult to prove.

The video quality metric presented in [15] has been modified in order to take into account the visual importance of the areas of the impaired video sequence. Different weighting functions based on the human saliency maps have been proposed. Neither of them succeeds in improving the performance of the quality metric. That is consistent with the idea it is not as simple as we might have thought.

What is certain is that observers have to inspect some areas more or less accurately in order to assess the video quality. Among all the visual fixations, some of these contribute to the quality assessment whereas others have low or no impact. In the future, the relationship between the duration of the visual fixations and the amount of distortion could be investigated. The idea is that observers do not require to focus a long time on strong distortions whereas when the amount of the distortion of an area is small, observers need more time to inspect and to judge the quality. This new hypothesis could be of strong importance since the definition of the saliency would be drastically modified.

References

[1] V. Seferidis, M. Ghanbari, D. E. Pearson, Forgiveness effect in subjective assessment of packet video, in: Electronics Letters, Vol. 21, 1992, pp. 2013–2014.

[2] K. Tan, M. Ghanbari, D. Pearson, An objective measurement tool for mpeg video quality, Signal Processing 70 (3) (1998) 279–294.

[3] M. Masry, S. Hemami, A metric for continuous quality evaluation of compressed video with severe distortions, Signal Processing: Image Communication 19 (2) (2004) 133–146.

[4] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Does where you gaze on an image affect your perception of quality ? applying to image quality metric, in: IEEE International Conference on Image Processing, Vol. 2, 2007, pp. 169–172.
[5] E. C. Larson, C. T. Vu, D. M. Chandler, Can visual fixation patterns improve image fidelity assessment?, in: IEEE International Conference on Image Processing, Vol. 3, 2008, pp. 2572–2575.
[6] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A coherent approach to model bottom-up visual attention, IEEE PAMI 28 (5) (2006) 802–817.
[7] J. Fecteau, D. Munoz, Salience, relevance and firing: a priority map for target selection, Trends in Cognitive Sciences 10 (8) (2006) 617–631.
[8] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Task impact on the visual attention in subjective image quality assessment, in: EUSIPCO, 2006.
[9] B. Tatler, R. Baddeley, I. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vision Research 45 (2005) 643–659.
[10] O. Le Meur, P. Le Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Research 47 (19) (2007) 2483–2498.
[11] R. Carmi, L. Itti, Causal saliency effects during natural vision, in: Proc. ACM Eye Tracking Research and Applications, 2006, pp. 11–18.
[12] A. Yarbus, Eye movements and vision, New york: Plenum Press.
[13] O. Le Meur, A. Ninassi, P. Le Callet, D. Barba, Do video coding impairments disturb the visual attention deployment?, Signal Processing: Image Communication, Under second revision.
[14] T. Fawcett, An introduction to roc analysis, Pattern Recognition Letters 27 (2006) 861–874.
[15] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Considering temporal variations of spatial visual distortions in video quality assessment, IEEE journal of selected topics in signal processing 3 (2) (2009) 253–265.
[16] VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment, http://www.vqeg.org/ (2000).
[17] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. on Image Processing 13 (2004) 600–612.
[18] H. R. Sheikh, A. C. Bovik, Image information and visual quality, IEEE Trans. on Image Processing 15 (2) (2006) 430–444.
[19] D. M. Chandler, S. S. Hemami, Vsnr : A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. on Image Processing 16 (9) (2007) 2284–2298.
[20] C. T. Vu, E. C. Larson, D. M. Chandler, Visual fixation patterns when judging image quality : Effects of distortion type, amount, and subject experience, in: IEEE Southwest Symposium on Image Analysis and Interpretation, 2008, pp. 73–76.