Cover Page

Title of the paper

Do video coding impairments disturb the visual attention deployment?

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={ Do video coding impairments disturb the visual attention deployment?},
Year={2010}}

DOI = doi:10.1016/j.image.2010.05.008

Do video coding impairments disturb the visual attention deployment?

O. Le Meur, A. Ninassi, P. Le Callet, D. Barba

Abstract

The visual attention deployment in a visual scene is contingent upon a number of factors. The relationship between the observer's attention and the visual quality of the scene is investigated in this paper: can a video artifact disturb the observer's attention? To answer this question, two experiments have been conducted. First, eye-movements of human observers were recorded, while they watched ten video clips of natural scenes under a free-viewing task. These clips were more or less impaired by a video encoding scheme (H.264/AVC). The second experiment relies on the subjective rating of the quality of the video clips. A quality score was then assigned to each clip, indicating the extent to which the impairments were visible. The standardized method DSIS (Double Stimulus Impairment Scale) was used, meaning that each observer viewed the original clip followed by its impaired version. Both experimental results have conjointly been analyzed. Our results suggest that video artifacts have no influence on the deployment of visual attention, even though these artifacts have been judged by observers as at least annoying.

Index Terms

visual attention, visual quality, saliency, video coding, H.264/AVC

Do video coding impairments disturb the visual attention deployment?

I. Introduction

ONE of the fundamental issues in studying visual perception is to identify the factors having the capability to significantly alter the visual deployment. In natural vision, the observer's attention is influenced both by sensory (also called bottom-up) and task related (called top-down) factors. The former rely on the low-level visual features (e.g. salience) whereas the latter rest on the ability of the visual system to focus on a target related to a given task (e.g., relevance to a task). Previous studies have provided valuable information about the impact of a task on the visual attention. The most famous experience is the one proposed by A. Yarbus in 1967 [1]. He showed that the pattern of eye movement was clearly dependent on the instructions given to the observers. In his experience, seven tasks were given to subjects as they viewed a painting. Seven different patterns of eye movements were obtained indicating that our visual attention is not purely bottom-up. In 1999, Land et al. [2] proposed to record eye movements of several observers, while they performed a familiar task (making tea). They reported that most of the visual fixations were relevant to the task (95%). In other specific domains such as reading [3], [4], the same conclusion was made. Eye movements are closely tied to the task which is being carried out [5], [6]. Conversely to the top-down mechanism, the bottom-up attentional allocation is driven by low-level factors. Indeed, overt visual attention is effortlessly drawn to salient parts present in our visual field. These fixated zones present a local singularity or a local contrast, whether it is luminance, color, texture, motion or even semantic [7]. For instance, Reinagel and Zador [8] showed that fixated areas are more contrasted in term of luminance than non fixated areas. More recently, Parkhurst and Niebur [9] showed that texture contrast contributes to the guidance of attention.

The relationship between bottom-up and top-down mechanisms still remains an open-issue. A part of the answer has been recently given in [10]. Authors reported that the top-down mechanism can override almost immediately the bottom-up one, as soon as a visual search task was given.

Eye movements are thus influenced by many factors. Nevertheless, one aspect has not been taken into consideration. In a free-viewing context, does the visual quality of the scene alter the deployment of the visual attention? In a previous study, L. Itti [11] proposed to use a saliency map in order to blur the non visually interesting salient areas and to keep the original resolution on the salient areas. The filtered video sequence was then encoded. In this approach, the goal was in one hand to reduce the encoding bit rate and in other hand to keep a good visual quality over the regions of interest. As the amount of blur applied on non regions of interest is very high, this approach is likely appropriate for very low bit rate encoding. Indeed, it is important in this context to allocate most of the bit budget on the salient regions. In our study, the level of impairment is not at all the same. The targeted applications are low to medium bit rates, typically those used in a TV broadcast system.

Our main contribution is to examine conjointly the pattern of eye movements and the quality of the viewed natural scene. To explore whether the alteration of the video content has an influence on the deployment of visual attention, two experiments have been conducted. First, eye movements of observers watching in free task either an unimpaired or an impaired natural video sequences are recorded. Besides, subjective tests have been performed in order to assess the subjective quality of the impaired video. Both experiments have been conducted under the same conditions (same room, same average luminance...). This study attempts to determine the influence of video coding artifacts on the visual attention.

In the following sections, the experimental protocols for both the eye tracking and the subjective analysis of the video quality are described. Results are examined with different methods. They indicate that the video artifacts do not have any influence on the deployment of visual attention. This finding has an important implication in the domain of video processing. In particular, this study suggests that the positions of the most interesting parts of a video are the same regardless of the coding impairments. It is however important to outline again the fact that impairments which are under analysis are only H.264/AVC coding artifacts.

TABLE I
NAME AND FEATURES OF THE STIMULI. EBU STANDS FOR EUROPEAN BROADCASTING UNION SVT FOR SWEDISH PUBLIC BROADCASTER SVERIGES TELEVISION AND IRT FOR INSTITUT FUR RUNDFUNKTECHNIK .

Name	Origin	Description	Spatial resolution	Temporal resolution
Dancer	EBU	Colourful dancing during soccer-break	720×480	50Hz
PrincessRun	SVT	Running person (camera pan, trees, grass)	720×480	50Hz
Foot	IRT	Fast-action outdoor sports	720×480	50Hz
Hockey	IRT	Fast-action indoor sports	720×480	50Hz
Crowd Run	SVT	Running crowd (No camera movement, trees, grass)	720×480	50Hz
Ducks	SVT	Take off of several ducks (water)	720×480	50Hz
Trees	SVT	Outdoor scenes	720×480	50Hz
Mobcal	SVT	Interiors, man-made environment (camera pan)	720×480	50Hz
ParkRun	SVT	Tracking shot of running man (water, trees, spring)	720×480	50Hz
ParkJoy	SVT	Tracking shot of running people (water, trees, winter)	720×480	50Hz

II. EYE TRACKING EXPERIMENT

A. Participants

Thirty six paid subjects participated to the experiments. They were all from the University of Nantes. They were all between 19 and 51 years old. Most of these participants were male (28 males and 9 female).
Prior to the test, subjects were screened for visual acuity by using a Monoyer optometric table and for normal color vision by using Ishihara's tables. All observers had normal or corrected to normal visual acuity and normal color perception. All were inexperienced observers (not expert in image or video processing) and naive to the purpose of this study.

B. Apparatus

Experiments had been performed with a dual-Purkinje eye tracker from Cambridge Research Corporation. The eye tracker was mounted on a rigid EyeLock headrest that incorporates an infrared camera, an infrared mirror and two infrared illumination sources. Before each trial, the subject's head was correctly positioned on a headrest so that their chin pressed on the chin-rest and their forehead lean against the head-strap. The heights of the chin-rest and head-strap system were adjusted so that the subject sat comfortable and their eye level was aligned with the center of the presentation display. The eye tracker is able to record the movement of one eye only. The eye tracker is fixed according the subject's guiding eye.
To obtain accurate data regarding the diameter of the subjects' pupil, a calibration procedure is required. The calibration aims at presenting to the subject a number of screen targets from a known distance. Once the calibration procedure is complete and a stimulus has been loaded, 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 close-up image of the eye. This video 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. The guaranteed sampling frequency is 50Hz and the accuracy is 0.5 degree of visual angle.
Stimuli were displayed at a viewing distance of four times the height of the picture (66 cm). The screen was a CRT, Dell Trinition, Ultra scan p991. Video were positioned in a random fashion around the center of the screen. Again, the rationale of this decision relies on the willingness to be less sensitive to the middle of the screen. Generally, when observers watch videos on computer 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 [12]–[14].
Tests were conducted in a standardized environment [15].

C. Stimuli

Twenty video sequences have been used. Ten video sequences of various natural contents are original video sequences without any degradation. Figure 1 shows the first picture of each video clip and Table I gives a short description.

(a) Dance (b) PrincessRun (c) Foot (d) Hockey (e) CrowdRun

(f) Ducks (g) Trees (h) Mobcal (i) ParkRun (j) ParkJoy

Fig. 1. Key frames for each video sequence used in the test.

From these ten original sequences, ten sequences were built which were impaired versions of the original ones. These impaired versions present a number of spatial and temporal artifacts. These impairments have been obtained by using an H.264/AVC video encoder. Each clip lasts 8s.

The impairment caused by the encoding is not spatially or temporally uniform. Some areas are impaired, whereas the quality of others remains unchanged. Figure 2 gives the distortion maps for the VQA (Video Quality Assessment) and SSIM (Structural Similarity Index Measurement) metrics for three pictures extracted from the sequence Dancer. Figure 4 shows two stripes extracted from the original and impaired video sequence Dancer. These illustrations indicate that the impairments introduced by the coding are not spatially uniformly distributed within each single frame.

Concerning the temporal variation of the quality, the quality variations estimated by the VQA (more details are given in the next section), the SSIM and the PSNR metric is given on figure 3. For the SSIM and VQA metrics, the quality scale is in the range of 0 (minimum quality) to 1 (maximum quality).

It indicates that the quality varies from frame to frame manner. These ruptures of quality, or these contrasts of quality, are typically those that could influence the deployment of visual attention. Figure 3 shows that these ruptures of quality occur periodically. This is due to the setting of the video encoder:

- GoP (Group of Picture) size = 23-4;
- Constant Bit Rate (one quantization level per frame without adaptive quantization). The bit rate was different for each video sequence. The goal was to reach a level of distortion at least visible. The degradations caused by the video encoder are subjectively considered as, at least, slightly annoying. For the sequence Foot, Hockey and Ducks, observers were very annoyed by the poor quality of the video. These results are described in the next section (see Table IV).

Table II gives three quality scores stemming from three objective quality metrics. The first objective score is given by the well-know peak signal-to-noise ration, called PSNR. The second and the third quality indicators are the metric defined in [16], [18], noted VQA, and the SSIM metric defined in [17]. Both metrics are in the range 0 (minimum quality) to 1 (maximum quality).

PSNR, VQA and SSIM are full-reference quality metrics, meaning that the original signal is available. The full-reference metrics are the best way to achieve a good prediction of the quality perceived by observers, compared to no-reference and reduced-reference quality metrics. Nevertheless, even with the original signal, it is not easy to correlate well with what is perceived. For instance, PSNR has limited performances, simply because PSNR is based on a pixel-based comparison. It is true that pixel-based metrics can successfully predict subjective ratings for a given context (dedicated for a type of distortion). However, without prior knowledge of the targeted applications,

(a) Original (b) Impaired (c) SSIM (d) VQA

Fig. 2. Examples of distortion maps for the sequence called Dancer. From the left to the right: (a) original picture, (b) impaired picture, (c) SSIM distortion map and (d) VQA distortion map (Brighter areas correspond to higher distortions).

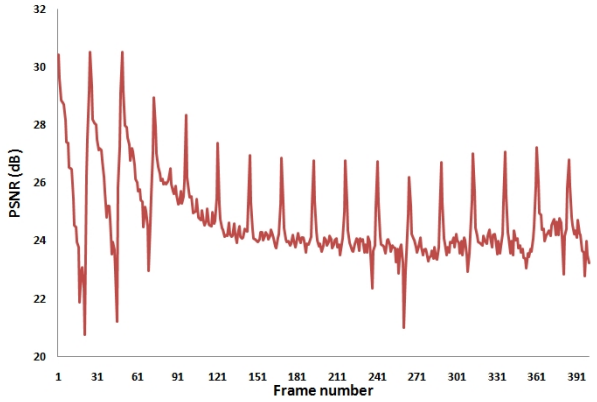
TABLE II
OBJECTIVE ASSESSMENT OF THE VIDEO QUALITY OF THE IMPAIRED SEQUENCES. PSNR=PEAK SIGNAL NOISE RATIO; VQA=VIDEO QUALITY ASSESSMENT (SEE [16]), 1=BEST QUALITY); SSIM=STRUCTURAL SIMILARITY INDEX MEASUREMENT (SEE [17]), 1=BEST QUALITY).

Clip	PSNR(dB)	VQA	SSIM
Dance	27.53	0.49	0.78
PrincessRun	23.94	0.37	0.73
Foot	30.42	0.31	0.80
Hockey	32.24	0.35	0.88
CrowdRun	27.51	0.35	0.81
Ducks	24.72	0.75	0.74
Trees	32.02	0.73	0.78
Mobcal	32.89	0.36	0.90
ParkRun	25.92	0.36	0.79
ParkJoy	25.99	0.43	0.77

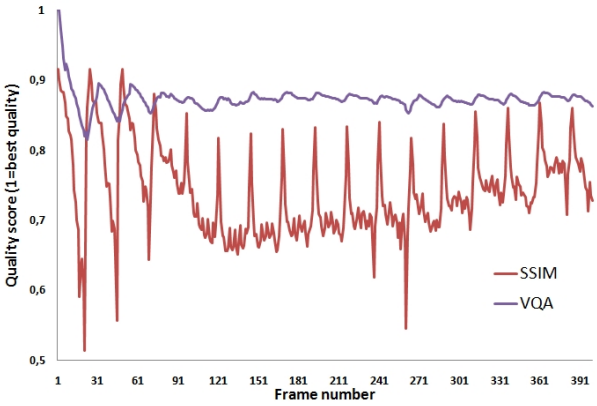
pixel-based metrics are not the most efficient one [19].
Most efficient quality metrics are those that rest on the properties of the human visual system. VQA as well as SSIM belong to this category. For instance, VQA metric relies on the sensitivity of the visual system as well as the interaction between different signals (commonly called visual masking) [16].

D. Stimuli presentation

Figure 5 (a) illustrates the manner stimuli were presented to observers. Each sequence was presented to subjects in a free-viewing task. Subjects were asked to examine freely without any specific objectives the sequence. The objective is to encourage a visual bottom-up behavior and to lessen the top-down effects. However, it is worth



(a)



(b)

Fig. 3. Temporal evolution of the predicted quality for the sequence Ducks. Three objective quality metrics are used: PSNR (a), VQA and SSIM (b).

(a) (b)

Fig. 4. Two stripes extracted from the original (a) and impaired (b) video Dancer (extracted from the impaired picture of the second row of figure 2.

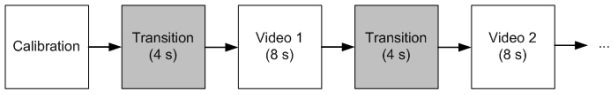
reminding that it is impossible to fully rule out top-down influences.
Prior to the onset of each sequence, two flickered black discs sequentially appeared at two different positions. They appeared for one second each. Then, a gray picture is displayed for two seconds. Note that there is no fixation 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 [14]. Each trial began with the calibration of the eye tracker.
The presentation list is given below. Note that the participant viewed two times each clip during the experiment. However, the original sequence (respectively the impaired one) is never followed by the impaired (respectively the original one). The goal is to reduce a possible memory effect that might be influenced the deployment of the visual attention. In addition, the presentation order is random, meaning that the first sequence viewed is either the original or the impaired one.

1) Trees-src
2) mobcal-src
3) CrowdRun-src
4) foot-src
5) Ducks-deg
6) ParkJoy-deg
7) parkrun-deg
8) mobcal-deg
9) ParkJoy-src
10) hockey-src
11) dance-deg
12) PrincessRun-deg
13) hockey-deg
14) PrincessRun-src
15) foot-deg
16) parkrun-src
17) Trees-deg
18) dance-src
19) Ducks-src
20) CrowdRun-deg

The whole experiment lasts in average 15 minutes.

E. Human priority maps

The visual inspection of the visual field is studied through the eye movements. Analysis of the eye movement record was then carried out off-line after completion of the experiments. The raw eye data is segmented into saccades and fixations. Saccades are very rapid eye movements allowing the viewer to explore his visual field. Fixation is a residual movement of the eye when the eye is locked on a particular area of the visual field. The fixation occurs between two saccade periods. Visual fixation allows the viewer to lock the central part of the retina, the fovea, on a particular target. The fovea plays a critical role in sensing details since most of the visual sensory resources are concentrated on this central part. The start and end time of the fixation were extracted as well as its spatial coordinates. A visual fixation must last at least 100 ms with a maximum velocity of 25 degrees per second [20].
From the spatial coordinates of visual fixation, a human priority map [21] is computed for each observer and for



(a) Stimuli presentation for the eye tracking experiment

(b) Stimuli presentation for the subjective quality experiment

Fig. 5. Stimuli presentation for the eye tracking experiment (a) and for the subjective quality assessment (b).

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 are first parsed in order to separate data into fixation and saccade periods (see [22]). The algorithm used to separate fixation and saccade periods is composed of the following steps. Each sample coming from the eye tracking apparatus is treated as described below:

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

The parsing of the raw eye tracking data leads to the determination of a fixation sequence, called SM^k (for an observer k) given by:

$$SM^k(s,t) = \sum_{i=1}^{M} \delta\left(s - s_i, t - t_i\right) \tag{1}$$

where M is the number of visual fixations, s and t represents a spatial coordiantes and the time, respectively. (s_i, t_i) are the spatial coordinates where $s_i = (x_i, y_i)$ and t_i is the start-time of the visual fixation i. δ is the Kronecker symbol, $\delta(t) = 1$, when $t = 0$, 0 otherwise.

Sequences SM^k are grouped together to form an average fixation sequence SM. SM could be interpreted as a map indicating where an average observer would look at:

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

where N is the number of observers.

This sequence is eventually smoothed with a 2D Gaussian filter (Parzen window method), leading to the human priority map. The rational of the Gaussian filtering is two-fold: Observers do not gaze at a point of the visual field but rather an area having a surface close to the size of the fovea. To simulate this, the standard deviation of the Gaussian filter is set to 0.75 degree of visual angle. The Gaussian filtering is also used to reflect the limited accuracy of the eye tracking apparatus.

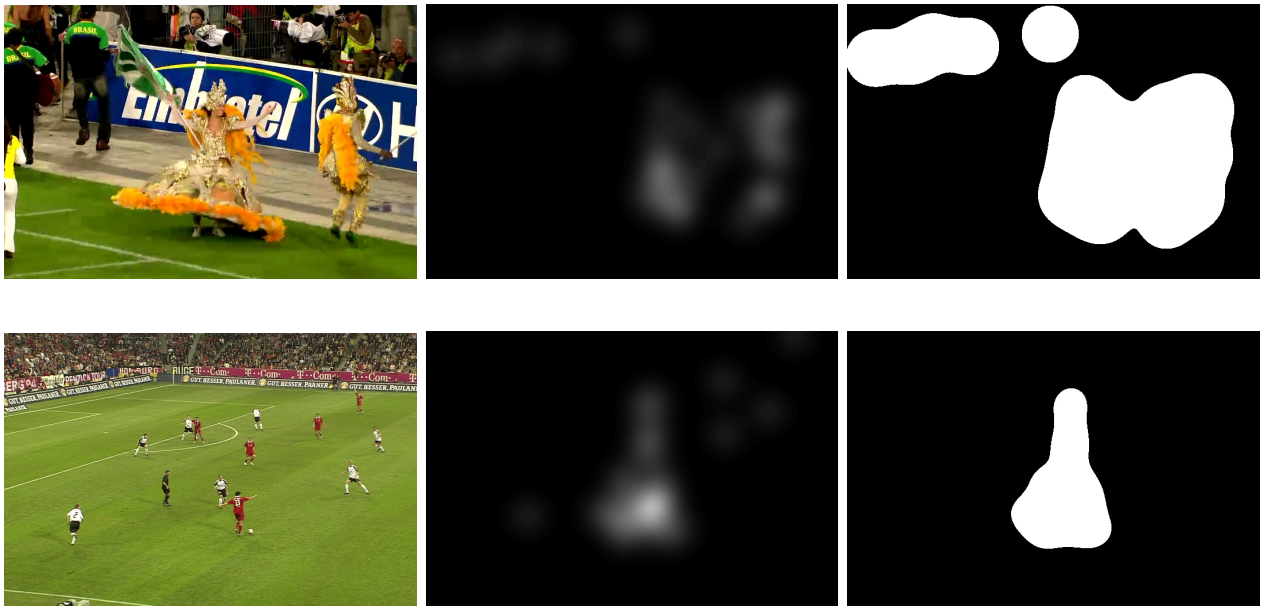
Figure 6 gives two examples of saliency maps. Binary maps are also given. This kind of maps will be used to evaluate the impact of the impairments on the region-of-interest. Two different thresholds are used: one equal to 10 and the other equal to 14 (these values are explained in section IV-B). A small threshold provides an over-detection of the region-of-interest whereas a high level will promote only the most visually interesting areas.

F. Results

1) Fixation durations: Table III gives the average fixation durations per video sequence (original and impaired sequences). The average fixation durations are equal to 434 and 447 ms for the original and impaired video sequences respectively. These values are not statistically different ($p < 0.31$). The video coding impairments do not seem to influence the fixation durations.

2) Distribution of fixation points: Figures 7 and 8 present the visual fixations for 6 clips. The blue dots indicate the spatial position of the fixations. All visual fixations collected during the experiments and for a given video are reported on the first frame. The idea is to subjectively compare the positions of visual fixations obtained when the video sequence is either impaired or not.

Given that the degradations are judged as slightly annoying or even annoying, it was rather logical to presume that the behavior of the visual attention system might be significantly altered by these impairments. It is interesting to notice from figures 7 and 8 that the impairments do not significantly alter the deployment of visual attention, when the observer is freely viewing scenes. The distribution of visual fixations obtained with the original sequence is as focused as the impaired one. Therefore, from this first and global analysis, it appears that the behavior of the visual attention is not modified by a rather coarse video coding.



(a) (b) (c)

Fig. 6. Saliency map examples: (a) original pictures (Dance first row and Foot second row); (b) saliency maps; (c) threshold map (threshold=10 for Dance, and threshold=14 for Foot).

TABLE III
FIXATION DURATION (AVERAGE±CI) AND AVERAGE NUMBER OF FIXATIONS PER SECOND. PAIRED T-TEST, 95% CONFIDENCE INTERVALS (CI).

	Original clips		Impaired clips	
Clip	Average fixation duration (ms)	Number of fixation per second	Average fixation duration (ms)	Number of fixation per second
Dance	425±170	1.84	444±175	1.87
PrincessRun	427±190	1.77	480±227	1.72
Foot	444±163	1.9	418±153	1.85
Hockey	363±109	2.23	455±180	1.9
CrowdRun	413±109	1.97	381±127	2.00
Ducks	473±180	1.8	508±195	1.79
Trees	393±120	2.07	369±106	2.08
Mobcal	330±87	2.4	349±121	2.15
ParkRun	616±397	1.5	600±300	1.52
ParkJoy	459±175	1.67	473±187	1.66
Average	434		447 ($p < 0.31$)	

Figures 7 and 8 suggest that the congruency of fixation locations does not seem to be dependent on the video coding artifacts. However, the congruency between observers differs strongly between the clips. Qualitatively speaking, when clips consist of regions of interest that stand out from the background (the ducks see figure 7 (c), people see figure 8 (b),(c)), the distribution of fixation points is more focused than when there is no region of interest that pops out (see figure 7 (b) to lesser extend (a), figure 8 (b)).

Figures 7 and 8 also indicate that the visual attention is driven by the content.

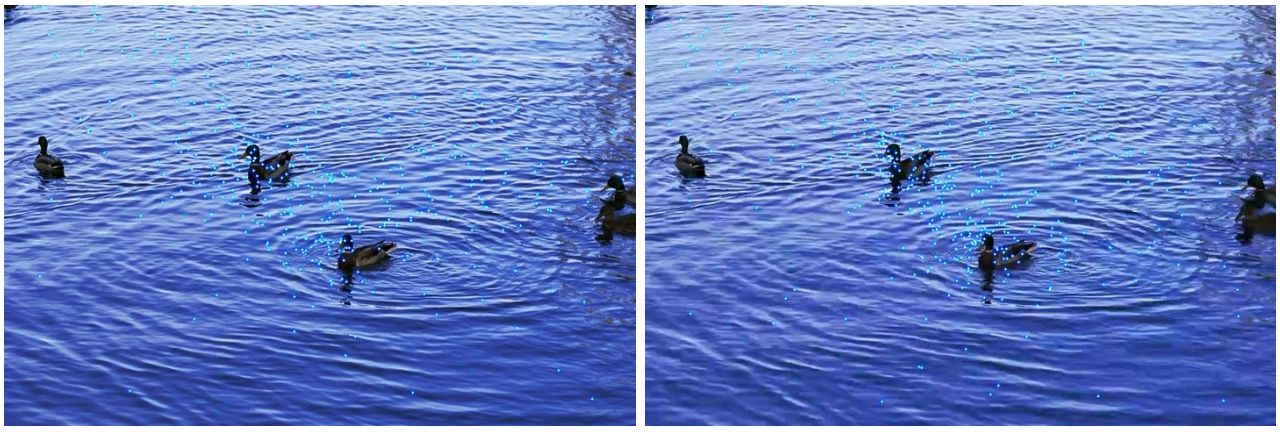
III. SUBJECTIVE QUALITY ASSESSMENT

A. *Participants*

Thirty six paid participants, the same as for the eye tracking experiment, are asked to assign each sequence with a quality score, indicating the extent to which the artifacts were visible.

(a)

(b)



(c)

Fig. 7. First pictures extracted from clips on which fixation points (blue dots) are superimposed. For each clip, two pictures are given. On the left, fixation points are those obtained on the original video sequence. The right picture for each pair, fixation points are those obtained on the impaired video sequence. (a) Trees, (b) CrowdRun, (c) Ducks.

(a)

(b)

(c)

Fig. 8. Same as figure 7 for three clips: (a) Mobcal, (b) Parkrun and (c) ParkJoy

TABLE IV
MOS FOR THE SUBJECTIVE QUALITY ASSESSMENT. MOS STANDS FOR MEAN OPINION SCORE. A VALUE OF 5 INDICATES THE BEST QUALITY. CI = 95 % CONFIDENCE INTERVALS.

Clip	MOS±CI
Dance	2.24±0.26
PrincessRun	2.52±0.33
Foot	1.39±0.22
Hockey	1.30±0.16
CrowdRun	2.94±0.32
Ducks	1.55±0.19
Trees	1.75±0.28
Mobcal	2.84±0.27
ParkRun	2.18±0.29
ParkJoy	2.24±0.3

B. Method

The standardized method DSIS (Double Stimulus Impairment Scale) is used. In DSIS, each observer views an unimpaired reference video sequence followed by its impaired version, each lasting 8s. The subject is told about the presence of the reference as first stimulus in each pair and is asked to rate only the test sequence. Experiments were conducted in standardized conditions (the same as previous [15]). The discrete scale used to score the distortion level is composed of 5 distortion grades:

1) very annoying;
2) annoying;
3) slightly annoying;
4) not annoying;
5) imperceptible.

C. Stimuli

The same stimuli used for the eye-tracking experiment were used.

D. Stimuli presentation

Figure 5 (b) illustrates the manner stimuli were presented to observers. A transition video sequence lasting 2 seconds is used between the reference and the impaired video sequence. This video is composed of average grey level pictures. After the impaired video sequence, the participant had to give a quality score for the whole video sequence. The scale used is given in the section named Method.

E. Results

Mean Opinion Scores (MOS) are given in Table IV. MOS and CI have been computed following the VQEG (Video Quality Expert Group) MM testplan [23], following correponding observers screening method.
The highest quality score (indicating the best quality) is equal to 2.94 (corresponding to slightly annoying) and is obtained by the sequence CrowdRun. The lowest one is equal 1.3 (between very annoying and annoying). It is obtained by the sequence Hockey. The average and the median MOS is equal to 2 (annoying) and 2.21 (between annoying and slightly annoying), respectively.

IV. COMPARISON OF HUMAN PRIORITY MAPS AND INTER-OBSERVER CONGRUENCY

The main question of this article is whether or not video coding impairments influences overt attention. Two indicators are examined. The first is the degree of similarity between the salience deduced from the impaired or the original sequence is assessed. Does the similarity degree become more variable with video coding artifacts? This specific problem is tackled by comparing the saliency values stemming from the two sets of results. The comparison

is done by using three different indicators: the linear correlation coefficient, the KullbackLeiblur divergence (KL-divergence) and the Receiver Operating Characteristic, commonly called ROC.
The second is the congruency between observers. If video coding artifacts influence overt attention, the inter-observer congruency could be higher in impaired video sequences than in original video sequences.

A. Correlation coefficient and KL-divergence

Two global metrics are used to assess the similarity degree between the two saliency maps. The first metric used is the 2D linear correlation coefficient, noted cc. It is a measure of dependence between two data sets. The correlation coefficient detects linear dependencies between two data sets. If the two data sets are independent, the correlation coefficient is 0. There is almost a perfect linear relationship between the two variables when the correlation value is close to -1 or 1:

$$cc(SM_p, SM_h) = \frac{cov(SM_p, SM_h)}{\sigma_p \sigma_h} \tag{3}$$

with SM_h and SM_p are the unimpaired and impaired saliency maps, respectively. $cov(SM_p, SM_h)$ is the covariance value between the two maps.
The second metric is the Kullback-Leibler divergence, noted KL [24]. The KL-divergence estimates the dissimilarity between two probability density functions:

$$KL(p|h) = \sum_{i=0}^{N-1} p(s_i) log \frac{p(s_i)}{h(s_i)} \tag{4}$$

with h and p the 2D probability density functions of the experimental priority maps. s_i represent the spatial coordinates of the pixel i. N is the number of pixel in the picture. h and p are deduced from SM_p and SM_h, respectively:

$$p(x_i) = \frac{SM_p(s_i)}{\sum_{k=0}^{N-1} SM_p(s_k)} \tag{5}$$

$$h(x_i) = \frac{SM_h(x_i)}{\sum_{k=0}^{N-1} SM_h(s_k)} \tag{6}$$

Therefore, p and h are homogeneous to a probability density function: $\sum_{k=0}^{N-1} p(s_k) = 1$ and $\leq p(s_k) \leq 1$, $\forall k$.

When the two probability densities are strictly equal, the KL-divergence value is zero. An upper-bound for the KL-divergence is obtained by computing the degree of similarity between the probability density deduced from the experimental priority map (based on the unimpaired sequence) and a uniform probability density function.

Average coefficients, obtained by averaging the coefficient values across frames, are presented in Table V. The similarity degree between the two maps is strong. Indeed, for most of the sequences, the correlation is above 0.7. The worst results are obtained for the sequence CrowdRun and Mobcal.

Concerning the KL-divergence, the dissimilarity between priority maps is about 4.56 whereas the dissimilarity between a priority and a uniform map is about 22. The dissimilarity between priority maps could be explained by the dispersion that exists between observers over time. This dispersion, or the inter-observer congruency is examined in subsection IV-C.

B. ROC analysis

In this study, the ROC analysis rests on a binary classification of the two data sets. Pixels of the original and impaired saliency sequences are labeled as fixated or not.
The reference, or the ground truth, is the unimpaired priority map. On this reference, the areas having a salience greater than a given threshold are labeled as fixated regions. Two thresholds (TH_{orig})are used: 10 and 14. As the saliency maps are encoded in grey level on 8 bits, the maximum value 255 is obtained in the case where all observers looked at the same areas at the same time. Given that the eye tracking experiment involved 36 observers,

TABLE V
AVERAGE CORRELATION COEFFICIENT (CC) AND KL-DIVERGENCE COMPUTED ON THE WHOLE SEQUENCE.

Clip	cc±CI	KL±CI	KL±CI (uniform)
Dance	0.77±0.02	3.98±0.31	21.07±0.28
PrincessRun	0.83±0.02	5.43±0.5	23.4±0.32
Foot	0.89±0.01	3.75±0.27	24.46±0.18
Hockey	0.86±0.01	4.23±0.28	23.86±0.13
CrowdRun	0.58±0.02	5.74±0.37	20±0.38
Ducks	0.80±0.01	3.23±0.24	22.9±0.24
Trees	0.72±0.01	4.3±0.29	22.11±0.29
Mobcal	0.62±0.02	7.1±0.39	21±0.26
ParkRun	0.92±0.01	3.82±0.3	24.6±0.27
ParkJoy	0.83±0.01	4.05±0.34	23.21±0.25
Average	0.782	4.56	22.67

TABLE VI
CONFUSION MATRIX FOR A BINARY CLASSIFIER. THE VALUES TP, FP, FN AND TN ARE OBTAINED FOR A GIVEN COUPLE OF THRESHOLDS (TH_{orig}, TH_{deg}).

		Original	
		Fixated	Not fixated
Impaired	Fixated	TP	FP
	Not fixated	FN	TN

the contribution of a particular observer is around 7 on the grey level scale. Therefore, a threshold equal to 14 means that an area with the label fixated is an area that has been simultaneously and exactly fixated by at least two observers. Figure 6 gives two examples: on the first row (c), a threshold equal to 14 is used. The resulting map is a binary map, where black areas represent the non fixated areas and the white areas are the fixated areas. A small threshold value conducts to an over detection whereas a higher threshold favors the most salient areas of the map. Concerning the impaired priority map, or outcome in ROC's terminology, 128 thresholds TH_{deg} are used. These thresholds are linearly distributed on 0 to 255. For each threshold, a binary map is deduced. This map is then compared to the binary map coming from the unimpaired video sequence.
For each pair of thresholds $((TH_{orig}, TH_{deg}))$, we can compute four numbers featuring the quality of the classification. These four numbers are grouped into a 2×2 confusion matrix as illustrated by Table VI. The four numbers represent the true positives (TP), the false positives (FN), the false negatives (FN) and the true negatives (TN). The true positive number is the number of fixated areas in the reference (original video sequence) that are also labeled as fixated in the impaired video sequence.

The true positive rate (TPR), also called sensitiviy or recall, is defined as:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

The false positive rate (FPR), also called the false acceptance rate, is defined as:

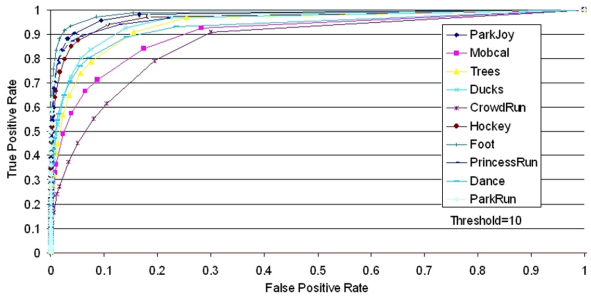
$$FPR = \frac{FP}{TP + FN} \tag{8}$$

A ROC curve is a plot of TPR versus FPR for different thresholds. The properties of the ROC curve are briefly recalled below:

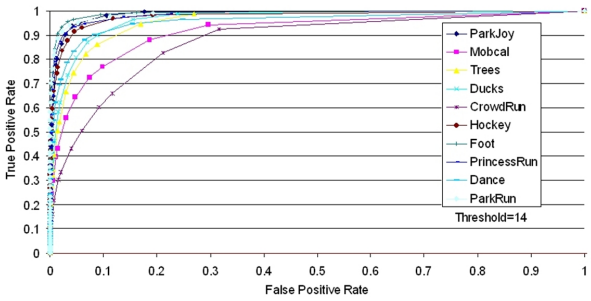
- the best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space;
- the diagonal line divides the ROC space in areas of good or bad classification.

TPR and FPR are first computed for each frame of a given sequence. Then, the average values of TPR and FPR are deduced. The quality of the classification is summarized using the area under curve (AUC). Higher AUC scores are better. The maximum is one. It indicates that the two binary maps are exactly the same.

Figure 9 gives the ROC curves for each sequence considering a threshold TH_{orig} and the set of thresholds TH_{deg} and table VII gives the corresponding AUC. The AUC is close to one, meaning that a threshold value can be easily determined to achieve a good classification.



(a)



(b)

Fig. 9. ROC curves for each sequence: (a) with a threshold equal to 10 for the human priority map coming from the unimpaired video; (b) same as before but with a threshold equal to 14.

Results of figure 9 and table VII suggest there are no significant differences between the attentional allocation on the original and on the impaired sequence. The worst cases are obtained with the sequences CrowdRun, Mobcal, Trees and Ducks. Two explanations can be proposed, one dependent on the amount and the distribution of impairment and the other related to the content of the video. The first explanation is the less probable. Indeed, all sequences have been encoded with the same video scheme and with the same encoding setting. Moreover, the subjective quality scores for these three sequences are not dramatically different from the other quality scores. They vary between 1.55 and 2.94 (we remind that a quality score of 1 means a very annoying quality whereas 3 means slightly annoying). Finally, as the quantization is the same for all the pictures, we cannot say that the salient areas are more impaired than the other ones.
The second explanation is the most plausible one and concerns the inter-observer congruency. The next section examines it.

C. Inter-observer congruency

As in Torralba et al. [25], the inter-observer congruency is measured for each frame of a sequence. A priority map, noted SM^{all}, is computed from the visual fixations generated by all-expect-one observers. This map excluding fixations coming from participant i is then compared to the priority map of participant i. We call this map SM^{i}. Following the proposition of Torralba et al. [25], the priority map SM^{all} is thresholded to select the most important salient areas. The threshold is chosen arbitrary small to favor the selection of the salient regions. The consistency across participants is determined by the percentage of visual fixations of the i^{th} participant that fell within the

TABLE VII
AUC FOR THE CLIPS, FOR THE TWO CONSIDERED THRESHOLDS ($AUC = 0.5$ (NO DISCRIMINATION); $AUC = 0.6 - 0.7$ (POOR); $AUC = 0.7 - 0.8$ (FAIR); $AUC = 0.8 - 0.9$ (GOOD); $AUC > 0.9$ (OUTSTANDING))

Clip	Th=10	Th=14
Dance	0.93	0.95
PrincessRun	0.96	0.98
Foot	0.98	0.99
Hockey	0.96	0.97
CrowdRun	0.86	0.87
Ducks	0.95	0.96
Trees	0.94	0.95
Mobcal	0.9	0.91
ParkRun	0.97	0.98
ParkJoy	0.97	0.98

salient areas of SM^{all} for a given frame. The final result is obtained by averaging the consistency obtained for all participants and all frames of the sequence.

Table VIII gives the congruency between observers for the original and impaired video sequences. Results reveal that there is a significant difference in the inter-observer congruency for most of the sequences. However, results indicate that the differences of the fixation behavior are not due to the impairment level but rather to the order of presentation of stimuli. Indeed we observe that an impaired video sequence does not provoke systematically a decrease of the inter-observer congruency. For instance, for the sequences Ducks, Mobcal and Parkrun, the highest inter-observer congruency is obtained when the sequence is impaired. What it is also interesting is to compare these results to the presentation list given in subsection II-D: the congruency between observers is significantly smaller for video sequence viewed in the second position, whatever the quality. Therefore, although the order of presentation of the stimuli was designed to lessen a possible memory effect, the fixation behavior of individual subjects during the second viewing was influenced by the first viewing of the sequence. This influence is statistically significant for 6 sequences. The three sequences for which the difference is not significant are Hockey, ParkRun and ParkJoy. It means that the content of these sequences leads subjects to fixate more similar locations than for other sequences. Sequences ParkRun and ParkJoy are similar sequences (see figure 1 and table I). There is a strong region of interest (man and people running) and this feature can explain why subjects tend to fixate more similar locations. Concerning the sequence Hockey, there is also a strong region of interest which is the hockey player.

Table VIII also indicates that the inter-observer congruency differs strongly between the different sequences. For the first viewing, the congruence values vary between 1.78% and 60% of similar fixation locations. For three sequences (CrowdRun, Trees and Mobcal), the congruency value is inferior to 5%. The content of these sequences does not present a region of interest that clearly pops out from the background (see figure 1) and therefore it is difficult to predict where a subject is going to fixate. It also explains why we obtained on these sequences a low degree of similarity of priority maps. As mentioned previously, these three sequences had the lowest AUC, KL and CC values. Note that performances of computational models of visual attention are more and more compared to the inter-observer congruency. This value is considered as an upper bound of prediction.

Finally, in spite of the memory effect, these results suggest that the video coding degradations do not change the fixation behavior of individual subjects. A lack of similarity between priority maps is probably due to the intrinsic inter-observer congruency.

V. DISCUSSION

In this study, we investigated whether or not the presence of strong visual coding degradations disturbed the deployment of visual attention.

We found that the saliency sequences for the impaired sequences are not significantly different from the original ones, indicating that the visual attention is almost invariant to video coding artifacts (impairements affect attention but the effect is rather small). At this stage, it is again worth mentioning that the degradations of the video clips are at least estimated as annoying by a panel of observers.

Considering that the deployment of the visual attention is significantly influenced by the low-level visual properties

TABLE VIII
AVERAGE PERCENTAGE OF FIXATIONS OF ONE PARTICIPANT THAT FALL WITHIN THE MOST SALIENT AREAS OF A PRIORITY MAP DEDUCED FROM ALL-EXPECT-ONE OBSERVER. THE CONGRUENCY IS COMPUTED ON BOTH THE ORIGINAL AND THE IMPAIRED VIDEO SEQUENCES. RESULTS ARE GIVEN FOR THE FIRST AND SECOND VIEWING OF THE SEQUENCE (SEE THE PRESENTATION ORDER OF THE STIMULI II-D).

Clip	First viewing	Second Viewing	Paired t-test
Dance	38.21	17.17^{orig}	$p = 10^{-5}$
PrincessRun	36.67	26.11^{orig}	$p = 0.056$
Foot	60.3^{orig}	47.23	$p = 0.03$
Hockey	43.41^{orig}	40.70	$p = 0.31$
CrowdRun	5.25^{orig}	3.18	$p = 0.036$
Ducks	17.82	7.27^{orig}	$p = 0.0004$
Trees	16.88^{orig}	8.39	$p = 0.002$
Mobcal	1.78	4.91^{orig}	$p = 0.001$
ParkRun	38.09	32.51^{orig}	$p = 0.22$
ParkJoy	34.16^{orig}	26.87	$p = 0.07$

(especially under free viewing [26]) and that the quality of the video was significantly reduced (to be at least annoying when a specific task of quality was given), it was not absurd to presume that observers would watch the video clips in a different way than those watching the same unimpaired clips. This is not the case, even though great care was taken on the way the quality of the video sequences was degraded. Indeed, as depicted by figures 3, 2 and 4 that shows the temporal evolution of objective quality scores, some distortion maps and two examples of distortion, respectively, the amount of impairment is not at all uniformly distributed spatially as well as temporally. Therefore, there are numbers of variations of quality that might potentially disturb the attention of the observer. How could we explain that there is no modification of the overt visual attention?

This result would indicate that the oculomotor behavior is also influenced by factors others than the low-level visual features, under free viewing task. It is not surprising since the transformation of visual precepts is the result of a series of complex biological and mental processes. As stated by Lester [27], visual perception is a function of the meaning we associate -through learned behavior or intelligent assumptions- with the object we see.
The fact that there is no explicit task does not mean that top-down influences are ruled out. To catch a total comprehension and understanding of visual images, observers use their own knowledge (memory, shape recognition...) to understand, to recognize and to interpret the scene.
No one can dispute the importance of early vision. An approximation to human fixations can be accomplished with mathematical models purely based on the low-level visual features. Several models, purely based on the low-level visual features, exist in the literature. A short review has been proposed in [28]. However, Torralba et al. [25] demonstrated that the performances of a model predicting where observers would look at on still pictures are much better when a combination of contextual information and low-level visual features is used than when the low-level visual features are used solely.
Finally, the fact that there is no significant modification in the deployment of visual attention in presence of distortion would suggest again that the fixation points are closely linked with the semantic and the context of the scene semantic, as suggested by [7], [29].
However, it could be argued that the amount of degradation on the shape of the object is likely not sufficient both to annoy the recognition of patterns and shapes and to disturb the comprehension of the scene. In a recent study, Rouse and Hemami [30] introduced the concept of similarity metric. Conversely to fidelity metric (PSNR, SSIM and WQA), a similarity metric is used to assess the quality of edges of the shapes. This kind of metric assesses the visual equivalence between two pictures by providing a score indicating the usefulness or utility of the content. A future work will consist in impairing the set of video sequences to dramatically reduce their utility scores. New eye tracking experiments will be conducted with these new materials. In addition, we will examine the influence of the transmission artifacts. Finally, the ability of computation models to predict regions of interest on impaired video sequences will be examined.

REFERENCES

[1] A. Yarbus, "Eye movements and vision," in *New-York, Plenum*, 1967.

[2] M. Land, M. Mennie, and J. Rusted, "Eye movements and vision," *Perception*, vol. 28, pp. 1311–1328, 2009.

[3] J. O'Regan, "Eye movements and reading," *in Eye Movements and Their Role in Visual and Cognitive Processes Ed.E Kowler (Amsterdam: Elsevier)*, pp. 395–453, 1990.

[4] K. Rayner, "Eye movements and cognitive processes in reading, visual search, and scene perception," *in Eye Movement Research: Mechanisms, Processes and Applications Eds J.M. Findlay, R Walker, R W Kentridge (Amsterdam: North-Holland)*, pp. 3–22, 1995.

[5] J. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews, Neuroscience*, vol. 5, pp. 495–501, 2004.

[6] J. Wolfe, "Guided search 4.0: Current progress with a model of visual search," *In W. Gray (Ed.), Integrated Models of Cognitive Systems, New York: Oxford.*, pp. 99–119, 2007.

[7] J. Henderson and A. Hollingworth, "High level scene perception," *Annual Review of Psychology*, vol. 25, pp. 210–228, 1999.

[8] P. Reinagel and A. Zador, "Natural scene statistics at the centre of gaze," *Network Comput. Neural Syst.*, vol. 10, pp. 341–350, 1999.

[9] D. Parkhurst and E. Niebur, "Texture contrast attracts overt visual attention in natural scenes," *European Journal of Neuroscience*, vol. 19, pp. 783–789, 2004.

[10] W. Einhauser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *Journal of Vision*, vol. 8, no. 2, pp. 1–19, 1998.

[11] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, p. 13041318, 2004.

[12] B. Tatler, "The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 1–17, 2007.

[13] D. Parkhurst and E. Niebur, "Scene content selected by active vision," *Spatial Vision*, vol. 16, pp. 125–154, 2003.

[14] 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.

[15] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland," Recommendation BT.500-11, 2002.

[16] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions," *IEEE Journal of selected topics in signal processing*, vol. 3, no. 2, pp. 253–265, 2009.

[17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. On Image Processing*, vol. 13, pp. 600–612, 2004.

[18] 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," in *Proc. SPIE Human Vision and Electronic Imaging XIII (HVEI'08)*, 2008.

[19] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Comm.*, vol. 43, no. 12, pp. 2959–2965, 1995.

[20] A. T. Bahill, A. Brockenbrough, and B. T. Troost, "Variability and development of a normative data based for saccadic eye movements," *Invest. Ophthalmol. Vis. Sci.*, vol. 21, no. 1, pp. 116–125, 1981.

[21] J. Fecteau and D. Munoz, "Salience, relevance and firing: a priority map for target selection," *Trends in Cognitive Sciences*, vol. 10, no. 8, pp. 617–631, 2006.

[22] D. Salvucci and J. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the Eye Tracking Research and Applications Symposium*, 2000, pp. 71–78.

[23] VQEG, "Final report of vqeg's multimedia phase i validation test," Tech. Rep., 2008.

[24] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[25] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychovisual Review*, vol. 113, pp. 766–786, 2006.

[26] L. Elazary and L. Itti, "Interesting objects in natural scenes are more salient," *Journal of Vision*, vol. 8, no. 3, pp. 1–15, 2008.

[27] P. Lester, "Visual communication: Images with messages," in *Belmont, CA:Wadsworth*, 1995.

[28] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters: visual attention and applications," in *ICIP*, 2009.

[29] A. Hollingworth and J. Henderson, "Semantic informativeness mediates the detection of changes in natural scenes," *Visual Cognition, Special Issue on Change Blindness and Visual Memory*, vol. 7, pp. 213–235, 2000.

[30] D. Rouse and S. Hemami, "Natural image utility assessment using image contours," in *International Conference on Image Processing, ICIP*, September 2009.